Mixtures of g-Priors in Generalized Linear Models

Yingbo Li[*]and Merlise A. Clyde[†]

arXiv:1503.06913v3 [stat.ME] 4 May 2018

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.

[*]Yingbo Li, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634 (e-mail: carolli13@gmail.com)

[†]Merlise A. Clyde is Professor and Chair, Department of Statistical Science, Duke University, Durham, NC 27708 (e-mail: clyde@duke.edu)

We illustrate inference using these priors and contrast them to other approaches via simulation and real data examples. The methodology is implemented in the `R package` `BAS` and freely available on `CRAN`.

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

1 Introduction

Subjective elicitation of prior distributions for variable selection, quickly becomes intractable as the number of potential variables p increases, motivating objective or conventional prior distributions for default usage (Berger and Pericchi 2001). In the context of linear models, Zellner's g-prior and 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), leading to the preference of these priors over many other conventional prior distributions (Forte et al. 2016).

Zellner (1983, 1986) proposed the g-prior as a simple partially informative distribution in Gaussian regression models $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathsf{N}(0, \sigma^2 \mathbf{I}_n)$. Through the use of imaginary responses at the observed design matrix $\mathbf{X}$, he obtained a conjugate Gaussian prior distribution $\boldsymbol{\beta} \mid \sigma^2 \sim \mathsf{N}(\boldsymbol{b}_0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$, with an informative mean $\boldsymbol{b}_0$, but having 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 shrinkage towards the prior mean and dispersion of the posterior covariance through the shrinkage factor $g/(1+g)$.

Using Zellner's g-prior for Bayesian variable selection (BVS) and Bayesian model averaging (BMA) requires specification of the hyper parameters $\boldsymbol{b}_0$ and g for each of the 2^p submodels,

---

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

indexed by $\mathcal{M} \in \{0,1\}^p$,

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

where $\mathbf{1}_n$ is a column vector of ones of length n, α is the intercept, $\mathbf{X}_{\mathcal{M}}$ is a model specific design matrix with $p_{\mathcal{M}}$ columns assumed to be 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 for BMA/BVS (Fernández et al. 2001; Liang et al. 2008) assigns an independent Jeffreys prior to α and σ^2

$$p(\alpha) \propto 1, \tag{2}$$
$$p(\sigma^2) \propto 1/\sigma^2, \tag{3}$$

and a g-prior using the centered predictors $(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$ of the form

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \sigma^2, g, \mathcal{M} \sim \mathrm{N}\left(\mathbf{0}_{p_{\mathcal{M}}}, g\sigma^2(\mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}})^{-1}\right), \tag{4}$$

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 $\mathbf{1}_n$. To justify the use of the improper reference priors on the intercept and variance (2)-(3), it is often assumed the columns of the design matrix $\mathbf{X}_{\mathcal{M}}$ must be orthogonal to $\mathbf{1}_n$ so that the expected Fisher Information is block diagonal. In that case $\mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$ in (4) reduces to the more familiar $\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}}$. Bayarri et al. (2012), however, argue that measurement invariance, which leads to (2) and (3), combined with predictive matching (discussed in more detail later) lead to the form of the g-prior above without the explicit centering of the predictors in the sampling model (1). Both model parameterizations, however, lead to equivalent posterior distributions through a change of variables.

It is well known that the choice of g affects shrinkage in estimation/prediction as well as posterior probabilities of models, with various approaches being put forward to determine a g with desirable properties. Independent of Zellner, Copas (1983, 1997) arrived at g-priors

3

in linear and logistic regression by considering shrinkage of maximum likelihood estimators (MLEs) to improve prediction and estimation, as in James-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 perspective. Fernández et al. (2001) studied consistency of BMA under g-priors in linear models, recommending $g = \max(p^2, n)$, 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 (4), include the Cauchy prior of Zellner and Siow (1980), 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. Mixtures of g-priors not only inherit the 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 and other attractive theoretical properties (Liang et al. 2008; Maruyama and George 2011; Bayarri et al. 2012; Feldkircher and Zeugner 2009; Celeux et al. 2012; Ley and Steel 2012; Feldkircher 2012; Fouskakis and Ntzoufras 2013). 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 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

$$p(Y_i) = \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{a(\phi_0)} + c(Y_i, \phi_0)\right\}, \quad i = 1, \ldots, n, \tag{5}$$

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 (5), $Y_1, \ldots, Y_n$ 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

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{1}_n\alpha + \mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} \tag{6}$$

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 parameters are fixed, with $a(\phi_0) = \phi_0/w_i$ with known ϕ_0 and w_i, a 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 ϕ_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.

In BVS/BMA, 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

$$p(\mathbf{Y} \mid \mathcal{M}, g) = \iint p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\alpha) p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g)\ d\alpha\ d\boldsymbol{\beta}_{\mathcal{M}} \qquad (7)$$

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 α and $\boldsymbol{\beta}_{\mathcal{M}}$. When the likelihood is non-Gaussian, normal priors no longer are conjugate, 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 α and $\boldsymbol{\beta} = \mathbf{0}$. Un-

der 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 α 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 $\mathcal{I}(\eta_{\mathcal{M},i}) = -\mathbb{E}\left[\partial^2 \log p(Y_i \mid \eta_i, \mathcal{M})/\partial \eta_i^2\right]$, 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 $1/c = \mathcal{I}(\eta) = -\mathbb{E}\left[\partial^2 \log p(Y \mid \eta, \mathcal{M}_{\varnothing})/\partial \eta^2\right]$ 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:

$$\boldsymbol{\beta}_{\mathcal{M}} \mid g, \mathcal{M} \sim \mathrm{N}\left(\mathbf{0},\ gc(\mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}})^{-1}\right) \tag{8}$$

where imaginary samples are generated from the null model $\mathcal{M}_{\varnothing}$ and the constant c is inverse

of the unit information given above evaluated at the MLE of α under $\mathcal{M}_\emptyset$. For the normal linear regression, $c = \sigma^2$ recovers the usual g-prior.

Under large sample approximations to the likelihood, the g-prior in (8) 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 Hinkley 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 $\boldsymbol{\beta}_{\mathcal{M}}$ and the local geometry under model $\mathcal{M}$. 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, provides analytic expressions that permit deeper understanding of their theoretical properties, and leads to computationally efficient algorithms under large sample approximations to likelihoods.

2.2 Local Information Metric g-Prior

The invariance and predictive matching criteria in Bayarri et al. (2012) lead to adoption of (2)-(3) for location-scale families. Although the Poisson and Bernoulli families are not location-scale 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 (2) and a second order Taylor expansion of the likelihood 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*

$$
\begin{aligned}
p(\mathbf{Y} \mid \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) &= \int p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\alpha) d\alpha \\
&\propto p\left(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right)^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right)\right\},
\end{aligned} \quad (9)
$$

where the approximation (9) *is precise $O(n^{-1})$, and the observed information of $\boldsymbol{\eta}_{\mathcal{M}}$, α, 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*

$$
\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) = \mathsf{diag}(d_i) \text{ where } d_i = -Y_i\,\theta''(\hat{\eta}_{\mathcal{M},i}) + (b \circ \theta)''(\hat{\eta}_{\mathcal{M},i}) \text{ for } i = 1, \ldots, n, \quad (10)
$$

$$
\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{1}_n, \quad (11)
$$

$$
\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^T(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}, \quad (12)
$$

respectively, and

$$
\mathcal{P}_{\mathbf{1}_n} = \mathbf{1}_n \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{1}_n\right)^{-1} \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \quad (13)
$$

is the orthogonal projection onto the span $\mathbf{1}_n$ under the information $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ inner product, $\mathbf{u}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{v}$ for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

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

The approximate marginal likelihood in (9) 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,

$$
\mathbf{X}_{\mathcal{M}}^c \triangleq (\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}, \quad (14)
$$

where the column means for centering are weighted averages $\bar{\mathbf{x}}_{\mathcal{J},j} = \sum_i d_i x_{ij} / \sum_i d_i$, with the weights proportional to d_i in (10). 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 $\mathcal{M}$

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g \sim \mathrm{N}\left(\mathbf{0},\ g \cdot \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}\right). \tag{15}$$

The advantage of (15) 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 (6)

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\mathcal{M}} \boldsymbol{\beta}_{\mathcal{M}} \tag{16}$$

where (with apologies for abuse of notation) α 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 (9).

In hypothesis testing, where parameter $\boldsymbol{\beta}$ is tested against a null value $\boldsymbol{\beta}_0$ with a nuisance parameter α, Jeffreys (1961) argues that when the Fisher information is block diagonal for all values of $\boldsymbol{\beta}$ and α, improper uniform priors on α can 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 α 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 (8) 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 an $O(n^{-1/2})$ 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. For model averaging, however, we may wish to have more precise estimates of Bayes Factors in the posterior probabilities. For estimation, local orthogonality at the MLE, as in the g-prior in (15), captures the large sample geometry of the likelihood parameters $(\alpha, \boldsymbol{\beta}_{\mathcal{M}})$ better than null orthogonality, and as we will see, greatly simplifies posterior derivations and theoretical calculations, and reduces computational complexity.

Bayarri et al. (2012) note that orthogonalization is not required for adopting a uniform prior on α, but instead the criteria of predictive matching and location invariance are used to justify the choice. Integration with respect to an improper uniform measure on α leads to a marginal likelihood involving a "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 $\mathbf{X}$. The uniform prior on the intercept in either parameterization with the g prior (15) leads to equivalent posterior distributions on $\boldsymbol{\eta}_{\mathcal{M}}$. For ease of exposition, however, we will adopt the centered parameterization in (16) for the remainder of the article, and drop the superscript c for simplification of notation when there is no ambiguity.

2.3 Posterior Distributions of Parameters

Under the g-prior (15) on $\boldsymbol{\beta}_{\mathcal{M}}$ and a uniform prior (2) on α for the centered parameterization (16), asymptotic limiting distribution theory (Bernardo and Smith 2000, pp. 287) under a Laplace approximation yields the approximate posterior distributions conditional on $\mathcal{M}$ as

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}, g \xrightarrow{D} \mathrm{N}\left(\frac{g}{1+g}\,\hat{\boldsymbol{\beta}}_{\mathcal{M}},\ \frac{g}{1+g}\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}\right), \tag{17}$$

$$\alpha \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} \mathrm{N}\left(\hat{\alpha}_{\mathcal{M}},\ \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-1}\right), \tag{18}$$

where the symbol $\xrightarrow{D}$ indicates convergence in distribution, and $\hat{\alpha}_{\mathcal{M}}$ and $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ are MLEs of α and $\beta_{\mathcal{M}}$ respectively under model $\mathcal{M}$. Due to local orthogonality, the posterior distributions of $\boldsymbol{\beta}_{\mathcal{M}}$ and α are asymptotically independent. Furthermore, for large n, the asymptotic marginal posterior distribution of α is proper, although its prior distribution is improper. Similar results are obtained by Held et al. (2015) under the assumption 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.

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). As discussed in Copas (1983, 1997), shrinking predicted values toward the center of responses, or equivalently, shrinking regression coefficients towards zero, may alleviate over-fitting, and thus yield optimal predictive performance. In Section 5.2, using the GUSTO-I data and logistic regression, we find that methods that favor 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 (8) and (15) 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 α and g-prior in (15) for any model $\mathcal{M}$, the approximate marginal likelihood for $\mathcal{M}$ and g in (7) has a closed form expression

$$
\begin{aligned}
p(\mathbf{Y} \mid \mathcal{M}, g) &= \int p(\mathbf{Y} \mid \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M}) p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, g) \, d\boldsymbol{\beta}_{\mathcal{M}} \\
&\propto p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} (1 + g)^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{ -\frac{Q_{\mathcal{M}}}{2(1 + g)} \right\},
\end{aligned} \tag{19}
$$

where the approximation (19) is precise to $O(n^{-1})$, $p_{\mathcal{M}}$ is the column rank of $\mathbf{X}_{\mathcal{M}}$, and

$$Q_{\mathcal{M}} = \hat{\boldsymbol{\beta}}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\hat{\boldsymbol{\beta}}_{\mathcal{M}} \tag{20}$$

is the Wald statistic (under observed information). For the null model $\mathcal{M}_{\o}$ where $p_{\mathcal{M}_{\o}} = 0$, $Q_{\mathcal{M}_{\o}} = 0$ so that (19) still holds. The approximate marginal likelihood (19) is a function of MLEs, which is fast to compute using existing algorithms such as the iterative weighted least squares (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 $\mathsf{BF}_{\mathcal{M}_1:\mathcal{M}_2} = p(\mathbf{Y} \mid \mathcal{M}_1, g)/p(\mathbf{Y} \mid \mathcal{M}_2, g)$, is commonly used in Bayesian model selection, assuming the two models are equally likely *a priori*. If $\mathsf{BF}_{\mathcal{M}_1:\mathcal{M}_2}$ 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 (19), the g-prior (15) yields closed form Bayes factors

$$\mathsf{BF}_{\mathcal{M}:\mathcal{M}_{\o}} = \frac{p(\mathbf{Y} \mid \mathcal{M}, g)}{p(\mathbf{Y} \mid \mathcal{M}_{\o})} = \exp\left\{\frac{z_{\mathcal{M}}}{2}\right\} \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_{\o}})}{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})}\right]^{\frac{1}{2}} (1+g)^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}}}{2(1+g)}\right\}, \tag{21}$$

where

$$z_{\mathcal{M}} = 2\log\left\{\frac{p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M})}{p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}_{\o}}, \mathcal{M}_{\o})}\right\} \tag{22}$$

is the change in deviance or two times the likelihood ratio test statistic for comparing model $\mathcal{M}$ to $\mathcal{M}_{\o}$. 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 (21) is closely related to the test-based Bayes factors

(TBF) of Hu and Johnson (2009); Held et al. (2015, 2016)

$$\mathsf{TBF}_{\mathcal{M}:\mathcal{M}_\emptyset} = \frac{\mathrm{G}\left(z_{\mathcal{M}}; \frac{p_{\mathcal{M}}}{2}, \frac{1}{2(1+g)}\right)}{\mathrm{G}\left(z_{\mathcal{M}}; \frac{p_{\mathcal{M}}}{2}, \frac{1}{2}\right)} = (1+g)^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{\frac{g\, z_{\mathcal{M}}}{2(1+g)}\right\}, \qquad (23)$$

which is derived from the asymptotic distributions of $z_{\mathcal{M}}$ under $\mathcal{M}$ and $\mathcal{M}_\emptyset$; $\mathrm{G}(z_{\mathcal{M}}; a, b)$ denotes the density of a Gamma distribution with mean a/b, evaluated at $z_{\mathcal{M}}$. Under the null or a local alternative where $\boldsymbol{\beta}_{\mathcal{M}}$ is in an $O(n^{-1/2})$ neighborhood of the null, the Wald statistic $Q_{\mathcal{M}}$ and deviance statistic $z_{\mathcal{M}}$ are asymptotically equivalent and the ratio $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\emptyset})/\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})$ in (21) converges to one in probability, resulting in the data-based Bayes factor in (21) (or DBF for short) with $Q_{\mathcal{M}}$ replaced by $z_{\mathcal{M}}$ being equivalent asymptotically to 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 both are close to BIC. In Section 5, using simulation and real examples, we find that with $g = n$, TBF and the DBF (21) 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 investigate the possible use of g-priors (15) 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 with known dispersion. We will return to the case of g-priors in linear models with unknown dispersion in the non-full rank case in Section 3.3.

For binary regression models with a finite sample size, data separation problems may lead to MLEs that are not unique nor finite (Albert and Anderson 1984; Heinze and Schemper 2002; Ghosh et al. 2017). 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

$$\gamma_0 + \mathbf{x}_{\mathcal{M},i}^T \boldsymbol{\gamma} \geq 0 \ \text{ if } Y_i = 1, \quad \gamma_0 + \mathbf{x}_{\mathcal{M},i}^T \boldsymbol{\gamma} \leq 0 \ \text{ if } Y_i = 0, \quad \text{for all } i = 1, \ldots, n. \quad (24)$$

In particular, there is complete separation if in (24) strict inequalities hold for all observations. In the absence of complete separation, there is quasi-complete separation if (24) holds with equality for at least one sample.

This implies that the information metric is no longer a valid inner product and that the operator in (13) is not an orthogonal projection. While it is possible to define projections in the case where $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ is not full rank (Christensen 2011, Chapter 10), we will restrict attention to the case where $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ is of full rank and conditions for asymptotic normality hold to avoid additional technicalities.

Design matrices that are not full rank lead to identifiability problems with MLEs of $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ in GLMs. Consider a model $\mathcal{M}$ where $\mathsf{rank}(\mathbf{X}_{\mathcal{M}}) = \rho_{\mathcal{M}} < p_{\mathcal{M}}$, and a full rank design matrix $\mathbf{X}_{\mathcal{M}'}$ that contains $\rho_{\mathcal{M}}$ 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,

$$\hat{\boldsymbol{\eta}}_{\mathcal{M}} = \mathbf{1}_n \hat{\alpha}_{\mathcal{M}} + \mathbf{X}_{\mathcal{M}} \hat{\boldsymbol{\beta}}_{\mathcal{M}} = \mathbf{1}_n \hat{\alpha}_{\mathcal{M}'} + \mathbf{X}_{\mathcal{M}'} \hat{\boldsymbol{\beta}}_{\mathcal{M}'} \quad (25)$$

and $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ is unique and positive definite. The precision matrix of the g-prior (15), $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^{cT} \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}}^c$ is well-defined, however, since $\mathsf{rank}(\mathbf{X}_{\mathcal{M}}^c) = \mathsf{rank}(\mathbf{X}_{\mathcal{M}}) = \rho_{\mathcal{M}} < p_{\mathcal{M}}$, it is not invertible. Note that the null-based g-prior (8) 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 rank deficient.

Proposition 2. *Suppose* $\mathsf{rank}(\mathbf{X}_{\mathcal{M}}) = \rho_{\mathcal{M}} < p_{\mathcal{M}}$, *then*

$$BF_{\mathcal{M}:\mathcal{M}_\varnothing} = \frac{p(\mathbf{Y} \mid \mathcal{M}, g)}{p(\mathbf{Y} \mid \mathcal{M}_\varnothing)} = \exp\left\{\frac{z_{\mathcal{M}}}{2}\right\} \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\varnothing})}{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})}\right]^{\frac{1}{2}} (1+g)^{-\frac{\rho_{\mathcal{M}}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}}}{2(1+g)}\right\}. \quad (26)$$

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.4. 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 exist, 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:

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

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

For example, under the g-prior (15), 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 (8), 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 $\hat{g}_{\mathcal{M}}^{\text{GEB}} = \arg\max_{g \geq 0} \sum_{\mathcal{M}} p(\mathcal{M})p(\mathbf{Y} \mid \mathcal{M}, g)$. 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 recommend 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 (9), it yields marginal likelihoods and (data-based) Bayes factors in closed form, for both GLMs (5), 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 (19) 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, $p(\mathbf{Y} \mid \mathcal{M}, u) \propto u^{p_{\mathcal{M}}/2} \exp(-uQ_{\mathcal{M}}/2)$, a conjugate prior for u (given ϕ_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

$$p(u \mid t, q, r, s, v, \kappa) = \frac{v^t \exp(s/v)}{B(t,q) \, \Phi_1(q, r, t+q, s/v, 1-\kappa)} \frac{u^{t-1}(1-vu)^{q-1}e^{-su}}{[\kappa + (1-\kappa)vu]^r} \, \mathbf{1}_{\{0 < u < \frac{1}{v}\}} \quad (27)$$

where parameters $t > 0, q > 0, r \in \mathbb{R}, s \in \mathbb{R}, v \geq 1$, and $\kappa > 0$. Here, $B(t,q)$ is the Beta function and $\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} (\alpha)_{m+n}(\beta)_n x^m y^n / [(\gamma)_{m+n} m! n!]$ 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 $n \in \mathbb{N}$. 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 3. *Let $u = 1/(1+g)$ have the prior distribution*

$$u \sim tCCH\left(\frac{a}{2}, \frac{b}{2}, r, \frac{s}{2}, v, \kappa\right) \tag{28}$$

where $a, b, \kappa > 0, r, s \in \mathbb{R}$, and $v \geq 1$, then for GLMs with a fixed dispersion ϕ_0, integrating the marginal likelihood in (19) with respect to the prior on u yields the marginal likelihood for $\mathcal{M}$ which is proportional to

$$\begin{aligned}
p(\mathbf{Y} \mid \mathcal{M}) &\propto p\left(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} v^{-\frac{p_{\mathcal{M}}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}}}{2v}\right\} \\
&\quad \cdot \frac{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2v}, 1-\kappa\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)},
\end{aligned} \tag{29}$$

where $p_{\mathcal{M}}$ is the rank of $\mathbf{X}_{\mathcal{M}}$, and $Q_{\mathcal{M}}$ is given in (20). The posterior distribution of u under model $\mathcal{M}$ is also a tCCH distribution asymptotically

$$u \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} tCCH\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}, r, \frac{s+Q_{\mathcal{M}}}{2}, v, \kappa\right) \tag{30}$$

allowing conjugate updating under integrated Laplace approximations.

The proof is available in supplementary material Appendix A.5.

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

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

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 function in two variables and the g-prior is derived using the information matrix; the hierarchical prior formed by (2), (15) and (28) will be denoted as the CHIC g-prior.

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	κ	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, $\mathcal{M}_\varnothing$
Uniform	2	2	0	0	1	1	No, $\mathcal{M}_\varnothing$
Jeffreys	0	2	0	0	1	1	No, $\mathcal{M}_\varnothing$
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
TruncGamma	$2a_t$	2	0	$2s_t$	1	1	If $s_t = O(n)$
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

In the conjugate updating scheme (30), 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 κ "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 section.

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

$$u \sim \mathrm{CH}\left(\frac{a}{2}, \frac{b}{2}, \frac{s}{2}\right), \tag{31}$$

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

$$u \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} \mathrm{CH}\left(\frac{a + p_{\mathcal{M}}}{2}, \frac{b}{2}, \frac{s + Q_{\mathcal{M}}}{2}\right), \tag{32}$$

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

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 behavior. 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 behavior like a multivariate Cauchy.

Proposition 4. *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.,*

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

22

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.6. 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 $s = 0$, (31) reduces to a Beta distribution, and when $b = 2$, it reduces to a truncated Gamma distribution. For the CH prior, we let parameter a be fixed, and parameters b and s be either fixed, or on the order of $O(n)$. The CH 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)

$$u \sim \mathrm{TG}_{(0,1)}(a_t, s_t) \Longleftrightarrow p(u) = \frac{s_t^{a_t}}{\gamma(a_t, s_t)} u^{a_t-1} e^{-s_t u} \, \mathbf{1}_{\{0<u<1\}} \tag{33}$$

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: $u \mid \mathbf{Y}, \mathcal{M} \sim \mathrm{TG}_{(0,1)}(a_t + p_{\mathcal{M}}/2, s_t + Q_{\mathcal{M}}/2)$. When $a_t = 1, s_t = 0$, (33) 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 $g \sim \mathrm{IG}(1/2, n/2)$.

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

$$u \sim \mathrm{Beta}\left(\frac{a_h}{2} - 1, 1\right), \text{ where } 2 < a_h \le 4 \tag{34}$$

with default value $a_h = 3$. When $a_h = 4$, (34) 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 (34)

can also be expressed as a Gamma distribution truncated to the interval $[0, 1]$, and hence has conjugate updating in GLMs,

$$u \sim \mathrm{TG}_{(0,1)} \left(\frac{a_h}{2} - 1, 0 \right) \implies u \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} \mathrm{TG}_{(0,1)} \left(\frac{p_\mathcal{M} + a_h}{2} - 1, \frac{Q_\mathcal{M}}{2} \right). \qquad (35)$$

Beta-prime prior (Maruyama and George 2011)

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

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)

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

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

$$p_r(u) = a_r \left[\rho_r(b_r + n) \right]^{a_r} \frac{u^{a_r - 1}}{\left[1 + (b_r - 1)u \right]^{a_r + 1}} \mathbf{1}_{\left\{ 0 < u < \frac{1}{\rho_r(b_r + n) + (1 - b_r)} \right\}} \qquad (36)$$

where $a_r > 0, b_r > 0$ and $\rho_r \geq b_r / (b_r + n)$. The robust prior 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 (31) as a special case, and vice versa.

In normal linear models, the robust prior yields closed form marginal likelihoods involving the Appell F_1 function (Appell 1925; Weisstein 2009). Similarly in GLMs, evaluation of

the special function Φ_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 (36) reduces to a truncated Gamma, which leads to

$$u \sim \text{TG}_{(0, \frac{p_{\mathcal{M}}+1}{n+1})}\left(\frac{1}{2}, \ 0\right) \implies u \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} \text{TG}_{(0, \frac{p_{\mathcal{M}}+1}{n+1})}\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{Q_{\mathcal{M}}}{2}\right), \qquad (37)$$

and with marginal likelihood proportional to

$$p(\mathbf{Y} \mid \mathcal{M}) \propto p\left(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}\right) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \left(\frac{n+1}{p_{\mathcal{M}}+1}\right)^{\frac{1}{2}}$$
$$\cdot \left(\frac{Q_{\mathcal{M}}}{2}\right)^{-\frac{p_{\mathcal{M}}+1}{2}} \cdot \gamma\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+1)}\right). \qquad (38)$$

Comparing (35) and (37) 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 (34), and 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 the well studied case of normal linear regressions with unknown variances, special cases of the CHIC g-prior, such as the hyper-g, hyper-g/n, 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; Bayarri et al. 2012; Sabanés Bové et al. 2015). For normal linear regression, Liang et al. (2008) show that under the g-prior (2)-(4), the marginal likelihood conditional on g (or u) is

$$p(\mathbf{Y} \mid \mathcal{M}, g) = \frac{p(\mathbf{Y} \mid \mathcal{M}_{\o}) \, (1+g)^{\frac{n-p_{\mathcal{M}}-1}{2}}}{[1+g(1-R_{\mathcal{M}}^2)]^{\frac{n-1}{2}}} \iff p(\mathbf{Y} \mid \mathcal{M}, u) = \frac{p(\mathbf{Y} \mid \mathcal{M}_{\o}) \, u^{\frac{p_{\mathcal{M}}}{2}}}{[(1-R_{\mathcal{M}}^2) + R_{\mathcal{M}}^2 u]^{\frac{n-1}{2}}}. \quad (39)$$

Under the general tCCH prior (28), 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 the special cases discussed in Section 3.2. We present results for the general normal linear model relaxing the assumption that $\mathbf{X}$ is full rank as suggested by Liang et al. (2008).

Proposition 5. *Consider a linear model* $\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \sigma^2 \sim N(\mathbf{1}_n\alpha + \mathbf{X}_{\mathcal{M}}^c\boldsymbol{\beta}_{\mathcal{M}}, \sigma^2\mathbf{W}^{-1})$, *with* $\mathbf{W}$ *a fixed* $n \times n$ *positive definite matrix and centered predictors* $\mathbf{X}_{\mathcal{M}}^c = (\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$ *where* $\mathcal{P}_{\mathbf{1}_n}$ *is the orthogonal projection onto the column space spanned by* $\mathbf{1}_n$ *using* $\mathbf{W}$ *in place of the observed information in* (13). *Define the coefficient of determination as*

$$R_{\mathcal{M}}^2 = \frac{\|\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c}\mathbf{Y}\|_{\mathbf{W}}^2}{\|(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{Y}\|_{\mathbf{W}}^2} \quad (40)$$

where $\|\mathbf{u}\|_{\mathbf{W}}^2 = \mathbf{u}^T\mathbf{W}\mathbf{u}$ *for* $\mathbf{u} \in \mathbb{R}^n$ *and* $\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c} = \mathbf{X}_{\mathcal{M}}^c(\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}\mathbf{X}_{\mathcal{M}}^c)^-\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}$ *is the rank* $\rho_{\mathcal{M}}$ *orthogonal projection onto the column space spanned by* $\mathbf{X}_{\mathcal{M}}^c$ *using the information inner product with* $\mathbf{W}$. *Under the prior distributions* $p(\alpha, \sigma^2) \propto 1/\sigma^2$, *g-prior* $\boldsymbol{\beta}_{\mathcal{M}} \mid \sigma^2, g, \mathcal{M} \sim N(\mathbf{0}, g\sigma^2(\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}\mathbf{X}_{\mathcal{M}}^c)^-)$, *and the tCCH prior on* $1/(1+g)$, *analytic expressions for marginal likelihoods are available for the following cases:*

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

$$p(\mathbf{Y} \mid \mathcal{M}, \mathbf{W}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset, \mathbf{W}) \, B\left(\frac{a+\rho_\mathcal{M}}{2}, \frac{b}{2}\right) \Phi_1\left(\frac{b}{2}, \frac{n-1}{2}, \frac{a+b+\rho_\mathcal{M}}{2}, \frac{s}{2v}, \frac{R_\mathcal{M}^2}{v-(v-1)R_\mathcal{M}^2}\right)}{v^{\frac{\rho_\mathcal{M}}{2}} \left[1 - (1 - \frac{1}{v})R_\mathcal{M}^2\right]^{\frac{n-1}{2}} B\left(\frac{a}{2}, \frac{b}{2}\right) {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, \frac{s}{2v}\right)}; \quad (41)$$

(2) if $s = 0$, then

$$p(\mathbf{Y} \mid \mathcal{M}, \mathbf{W}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\emptyset, \mathbf{W}) \, \kappa^{\frac{a+\rho_\mathcal{M}-2r}{2}} B\left(\frac{a+\rho_\mathcal{M}}{2}, \frac{b}{2}\right)}{v^{\frac{\rho_\mathcal{M}}{2}} (1 - R_\mathcal{M}^2)^{\frac{n-1}{2}} B\left(\frac{a}{2}, \frac{b}{2}\right) {}_2F_1\left(r, \frac{b}{2}; \frac{a+b}{2}, 1 - \kappa\right)} \tag{42}$$
$$\cdot F_1\left(\frac{a+\rho_\mathcal{M}}{2}; \frac{a+b+\rho_\mathcal{M}+1-n-2r}{2}, \frac{n-1}{2}; \frac{a+b+\rho_\mathcal{M}}{2}; 1 - \kappa, 1 - \kappa - \frac{R_\mathcal{M}^2\kappa}{(1 - R_\mathcal{M}^2)v}\right).$$

Furthermore, if the rank of $\mathcal{P}_{\mathbf{X}_\mathcal{M}^c}$ is $n - 1$, the Bayes factor $\mathsf{BF}_{\mathcal{M}, \mathcal{M}_\emptyset} = 1$.

A proof of Proposition 5 is provided in supplementary material Appendix A.8, along with a brief summary of relevant special functions in supplementary material Appendix A.7. Note that (1) applies to the CH prior and all its special cases, and (2) applies to robust, hyper-g/n, and intrinsic priors.

Similarly, 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 ϕ:

$$p(Y_i \mid \theta_i, \phi) = \phi^{\frac{1}{2}} p(Y_i \mid \theta_i)^\phi p(Y_i \mid \theta_i = t_i)^{1-\phi}, \quad i = 1, \ldots, n, \tag{43}$$

where $p(Y_i \mid \theta_i)$ follows the GLM density (5), and $t_i = \arg\max_{\theta_i} p(Y_i \mid \theta_i)$ is a constant that depends on the data. In this formulation, the MLEs $\alpha_\mathcal{M}, \boldsymbol{\beta}_\mathcal{M}$ do not depend on ϕ and the observed information of $\alpha_\mathcal{M}, \boldsymbol{\beta}_\mathcal{M}$ is block diagonal $\mathcal{J}_{n,\phi}\left(\hat{\alpha}_\mathcal{M}, \hat{\boldsymbol{\beta}}_\mathcal{M}\right) = \mathsf{diag}\left\{\phi \mathcal{J}_n(\hat{\alpha}_\mathcal{M}), \phi \mathcal{J}_n(\hat{\boldsymbol{\beta}}_\mathcal{M})\right\}$, where $\mathcal{J}_n(\hat{\alpha}_\mathcal{M})$ and $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_\mathcal{M})$ are the observed information matrices for standard GLMs as in (11) and (12). A CHIC g-prior to account for over-dispersion based on the observed informa-

tion

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

provides closed form approximate marginal likelihoods after integrating out ϕ

$$p(\mathbf{Y} \mid \mathcal{M}, u) \propto \ \frac{[\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})]^{-\frac{1}{2}} \, u^{\frac{p_{\mathcal{M}}}{2}}}{\left\{ uQ_{\mathcal{M}} + 2\sum_{i=1}^n \left[Y_i(t_i - \hat{\theta}_i) - b(t_i) + b(\hat{\theta}_i) \right] \right\}^{\frac{n-1}{2}}}. \tag{44}$$

A derivation of (44) is provided in supplementary material Appendix A.9. Since the kernel function of u (44) is of the same form as (39), there exists a similar result to Proposition 5 for tractable marginal likelihoods after integrating out u under the CHIC prior.

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 applied statisticians. 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 α, nor orthogonalization of α (Bayarri et al. 2012). For the g-prior (15), 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 (15), 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 α (2) 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 σ^2 in (3) 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

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

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

29

null based g-prior in (8) 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.10. 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 (4) with fixed g (Liang et al. 2008), in the sense that the Bayes factor $\mathsf{BF}_{\mathcal{M}:\mathcal{M}_\emptyset}$ (39) 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 the prior $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \alpha, \mathcal{M})$ should be independent of n and remain proper, instead of degenerating to a point mass (Bayarri et al. 2012). By Lemma 1 in the supplementary materials, $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = O_P(n)$ if $\mathcal{M} \supset \mathcal{M}_T$, so with any fixed value of g, the g-prior (15) 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 6. *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.11. 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 7. *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 $t \in \mathbb{R}$, $\phi_{z\mid\mathbf{Y},\mathcal{M}_T}(t) \triangleq \mathbb{E}\left(e^{itz}\right) \xrightarrow{\text{P}} \exp(it)$. Therefore, all moments of $p(z \mid \mathbf{Y}, \mathcal{M}_T)$ converge to 1 in probability. In particular, the posterior mean $\mathbb{E}(z \mid \mathbf{Y}, \mathcal{M}_T) \xrightarrow{\text{P}} 1$ and the posterior variance $\mathbb{V}(z \mid \mathbf{Y}, \mathcal{M}_T) \xrightarrow{\text{P}} 0$.*

The proof is given in supplementary materials Appendix A.12.

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

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = p(\mathcal{M}_T \mid \mathbf{Y})\, p(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T) + \sum_{\mathcal{M} \neq \mathcal{M}_T} p(\mathcal{M} \mid \mathbf{Y})\, p(\boldsymbol{\beta}_\mathcal{M} \mid \mathbf{Y}, \mathcal{M}) \tag{45}$$

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., $p(\mathcal{M}_T \mid \mathbf{Y}) \xrightarrow{\text{P}} 1$, the second term in (45) vanishes in the limit, so we just need to study the posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}_T}$. When $\mathcal{M}_T = \mathcal{M}_\emptyset$, even 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, 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.13. 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 2 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. The R package `BAS`, 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, which have pairwise correlation $\mathrm{cor}(\mathbf{X}_i, \mathbf{X}_j) = r^{|i-j|}$ for $1 \leq i < j \leq p$, with $r = 0$ (independent design) or $r = 0.75$ (correlated design). We consider four different levels of 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 $2^{17} \approx 131,000$ 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 $p_{\mathcal{M}_T} \ll p/2$.

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 6 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 $g = O(1)$.

34

Table 2: Values of the 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**	**0**	**0**	**0**
Sparse	5	-0.5	**b**	**0**	**0**	**0**
Medium	10		**b**	**0**	**b**	**0**
Full	20		**b**	**b**	**b**	**b**

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[2] when $\mathcal{M}_T = \mathcal{M}_\emptyset$. 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 $p = 100$. The results under the unit information prior, i.e., the g-prior with $g = n$, 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

---

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

$p = 100$ 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 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 probabilities, 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	11	11	61	6
CH$(a=1,b=n)$	85	82	60	30	37	8	6	0	15	9	61	6
CH$(a=1/2,b=n/2)$	86	84	46	28	30	**12**	8	0	3	2	62	6
CH$(a=1,b=n/2)$	70	73	45	30	30	11	8	0	8	4	**63**	6
Beta-prime	**92**	88	61	29	**38**	8	7	0	11	6	61	6
ZS adapted	85	82	60	30	37	8	6	0	8	11	61	6
Benchmark	91	**93**	28	**31**	19	8	16	0	6	3	62	6
Robust	86	83	41	29	29	10	8	0	4	1	52	5
Intrinsic	76	77	40	29	26	10	8	0	2	3	56	5
Hyper-g/n	77	73	37	**31**	23	7	16	0	0	0	1	0
DBF, $g=n$	73	79	**67**	29	31	2	0	0	**68**	26	55	3
TBF, $g=n$	73	79	**67**	29	31	2	0	0	**68**	27	55	3
Jeffreys	NA	NA	28	28	17	7	16	0	0	0	1	0
Hyper-g	6	9	25	29	15	8	16	**1**	0	0	0	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	0	0
AIC	3	7	5	9	13	5	12	0	1	2	**63**	**15**
BIC	73	79	**67**	29	31	2	0	0	67	**28**	55	3

Estimation and prediction are often more important than identifying the true model,

Table 4: Logistic regression simulation example: 100 times the average SSE $= \sum_{j=0}^{p}(\tilde{\beta}_j - \beta^*_{j,\mathcal{M}_T})^2$ of 100 realizations. Column-wise minimum is in bold type.

p	\multicolumn{8}{c}{20}	\multicolumn{4}{c}{100}					
$p(\mathcal{M})$	\multicolumn{8}{c}{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	51	96	94	184	109	135	**26**	78
CH$(a=1, b=n)$	3	4	21	**43**	51	96	94	183	119	139	**26**	77
CH$(a=1/2, b=n/2)$	4	5	22	**43**	50	92	87	172	158	182	**26**	75
CH$(a=1, b=n/2)$	4	5	22	**43**	50	92	86	172	160	189	27	74
Beta-prime	3	3	21	44	51	96	94	183	123	142	**26**	78
ZS adapted	3	4	21	**43**	51	96	94	183	121	144	**26**	77
Benchmark	4	7	21	44	49	89	73	**158**	169	195	**26**	75
Robust	4	5	23	44	52	91	90	165	252	292	193	139
Intrinsic	4	6	23	44	52	91	90	165	239	284	143	90
Hyper-g/n	3	4	21	**43**	**48**	**88**	**72**	158	197	226	441	326
DBF, $g=n$	3	3	**20**	47	54	117	113	244	**42**	**65**	27	82
TBF, $g=n$	3	3	**20**	47	54	117	113	245	**42**	**65**	27	83
Jeffreys	2	3	22	45	50	89	74	159	212	231	444	387
Hyper-g	2	3	22	45	51	90	76	160	219	233	451	396
Uniform	2	2	22	46	52	91	78	161	230	236	459	411
Local EB	**1**	**1**	22	45	50	89	74	**158**	245	236	608	434
AIC	8	15	29	51	59	93	103	**158**	287	353	39	**71**
BIC	3	3	21	47	55	117	113	245	**42**	**65**	27	82

particularly for large p. To evaluate the performance for parameter estimation, we report $\text{SSE}(\boldsymbol{\beta}) = \sum_{j=0}^{p}(\tilde{\beta}_j - \beta^*_{j,\mathcal{M}_T})^2$ in Table 4 where $\tilde{\beta}_j$ represents the posterior mean estimates under BMA (here β_0 corresponds to the intercept α); 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. One exception is the $g = O(1)$ priors and the local EB, which have small SSE under the null despite their 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.

Figure 1 illustrates heatmaps 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, as in Burnham and Anderson (2004); Raftery (1996) to approximate 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 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 1000 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

---

[3]This dataset is available on the book website http://www.clinicalpredictionmodels.org
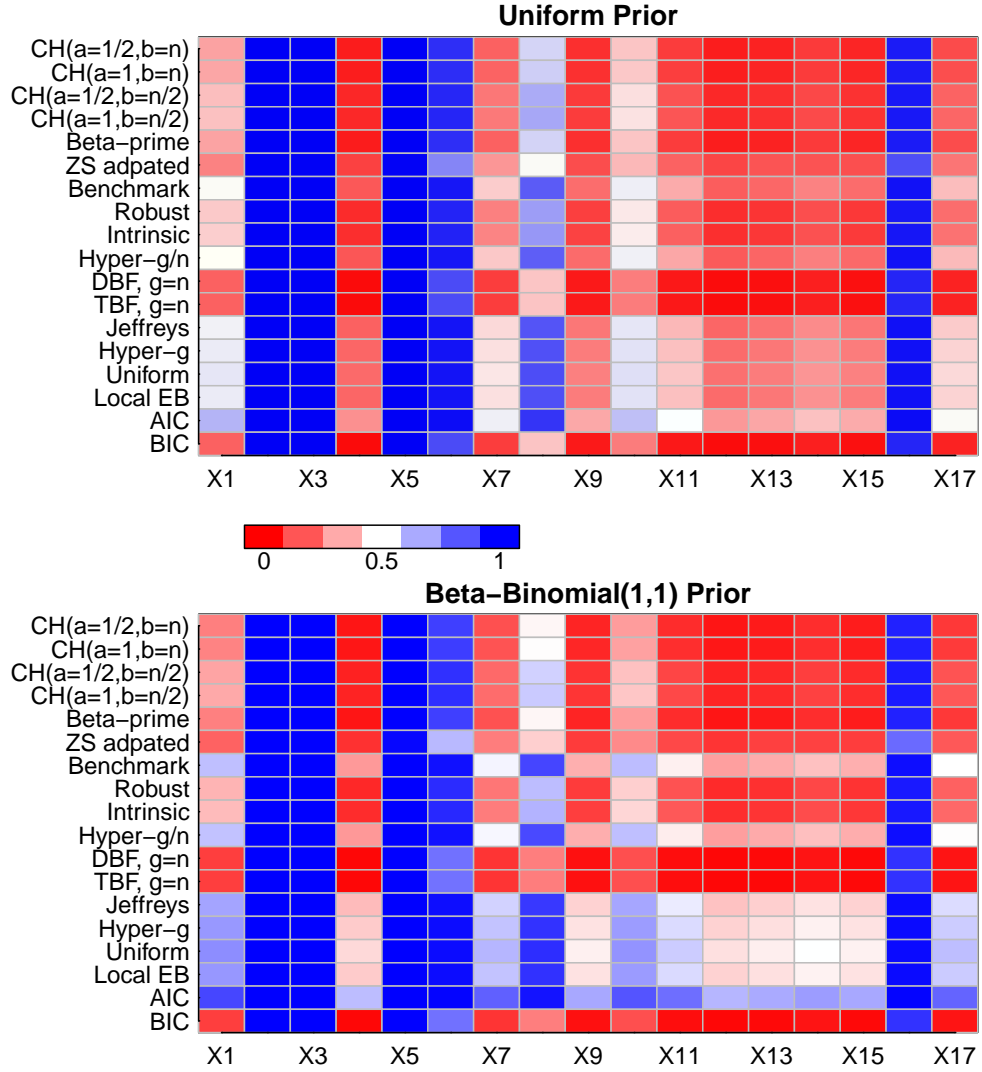
Figure 1: Marginal posterior inclusion probabilities for the GUSTO-I data. The colors are related to the magnitude of the inclusion probability with darkest blue corresponding to one and red to zero, while 0.5 is shown as white.

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, and smaller Brier score indicates 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 particular, the uniform prior on u (a special case of the hyper-g prior) yields the most accurate prediction under all four summaries. Over the model space, the uniform prior slightly outperforms the Beta-Binomial$(1, 1)$, in terms of AUC, CS, and LS.

Table 5: Prediction accuracy for the GUSTO-I data, aggregated from 1000 bootstrap cross validation sets. Bold font marks the largest AUC, the CS closest to one, and the smallest LS and Brier score.

$p(\mathcal{M})$	AUC		CS		LS		Brier	
	Unif	BB(1,1)	Unif	BB(1,1)	Unif	BB(1,1)	Unif	BB(1,1)
CH($a = 1/2, b = n$)	0.8346	0.8338	0.9055	0.9065	0.1848	0.1851	0.0497	0.0497
CH($a = 1, b = n$)	0.8347	0.8339	0.9054	0.9063	0.1848	0.1851	0.0497	0.0497
CH($a = 1/2, b = n/2$)	0.8349	0.8343	0.9054	0.9049	0.1846	0.1849	0.0496	0.0497
CH($a = 1, b = n/2$)	0.8349	0.8343	0.9054	0.9048	0.1846	0.1849	0.0496	0.0497
Beta-prime	0.8346	0.8338	0.9055	0.9065	0.1848	0.1851	0.0497	0.0497
ZS adapted	0.8345	0.8329	0.9338	0.9382	0.1846	0.1854	0.0496	0.0498
Benchmark	**0.8352**	**0.8347**	0.9292	0.9251	0.1841	0.1842	**0.0495**	**0.0495**
Robust	0.8349	0.8344	0.9012	0.8998	0.1847	0.1849	0.0496	0.0497
Intrinsic	0.8350	0.8344	0.9010	0.8993	0.1846	0.1849	0.0496	0.0497
Hyper-g/n	**0.8352**	0.8346	0.9287	0.9265	0.1841	0.1842	**0.0495**	**0.0495**
DBF, $g = n$	0.8338	0.8325	0.9100	0.9126	0.1852	0.1857	0.0498	0.0499
TBF, $g = n$	0.8338	0.8325	0.9101	0.9126	0.1852	0.1857	0.0498	0.0499
Jeffreys	**0.8352**	0.8346	0.9392	0.9373	0.1840	0.1841	**0.0495**	**0.0495**
Hyper-g	**0.8352**	0.8346	0.9446	0.9429	**0.1839**	**0.1840**	**0.0495**	**0.0495**
Uniform	**0.8352**	0.8346	**0.9502**	**0.9485**	**0.1839**	**0.1840**	**0.0495**	**0.0495**
Local EB	**0.8352**	0.8346	0.9391	0.9373	0.1840	0.1841	**0.0495**	**0.0495**
AIC	0.8351	0.8344	0.8813	0.8645	0.1846	0.1850	**0.0495**	0.0496
BIC	0.8338	0.8325	0.9096	0.9122	0.1852	0.1857	0.0498	0.0499

One potential explanation for the better performance of the $g = O(1)$ and the local EB is that shrinkage is better calibrated to the data by avoiding over-fitting (Copas 1983). As
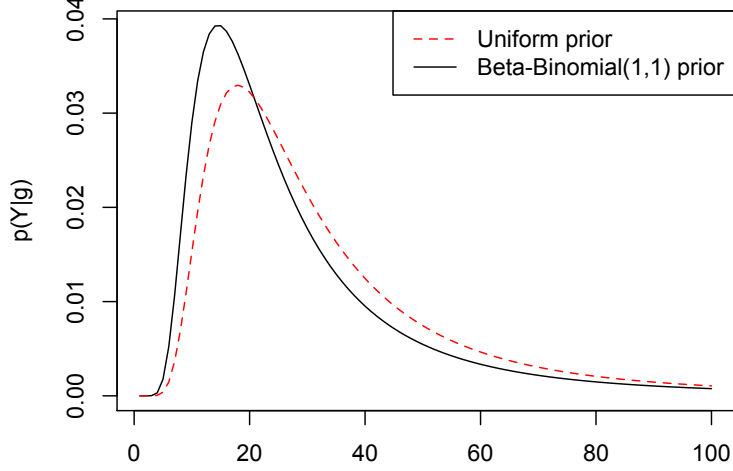
Figure 2: Marginal likelihood of g for the GUSTO-I data ($n = 2188$ and $p = 17$).

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, the BMA posterior estimate of g is 14.7 for the uniform prior on u, 16.5 for hyper-g, 18.4 for local EB, 24.0 for benchmark, 25.4 for hyper-g/n, 50.0 for ZS adapted, 286.5 for intrinsic, 298.1 for CH($a = 1, b = n, s = 0$), 319.6 for Beta-prime, and 321.2 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 $(n - p_{\mathcal{M}})/(p_{\mathcal{M}} + 1) \geq 120.6$ and $n/(p_{\mathcal{M}} + 1) \geq 121.6$, respectively, may not be well supported by the data, when n is large and p is small like the GUSTO-I data.

---

[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 $p(\mathcal{M}) = 1/2^p$.

6 Conclusion

The analytic tractability of asymptotic CHIC posterior distributions allows insight into their theoretical properties and fast computation, serving as a robust conventional prior for most instances. The CHIC family encompasses the majority of mixtures of g-priors used in practice. Under a wide range of hyper parameter choices, CHIC g-priors satisfy various desiderata proposed by Bayarri et al. (2012) for BVS. Based on both theoretical and empirical studies, we recommend priors with the choice $g = O(n)$, such as the CH prior with $b = O(n)$ or $s = O(n)$, hyper-g/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. Because model selection and prediction are two unaligned goals with different objective functions (Copas 1983), it is not surprising that no single prior overwhelmingly outperform others for both goals. Similar to the findings of Ley and Steel (2012) in linear models, we also recommend the benchmark and hyper-g/n priors for general practitioners, due to their balanced performance in selection and prediction.

The approximate marginal likelihoods under the CHIC g-prior require only simple summaries from GLMs, hence the CHIC g-prior has the same computational 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 computed directly without sampling the model specific parameters. All of the methods used in the examples and simulation studies within this article are implemented in the R package `BAS` (Clyde 2018) available on CRAN. Additional refinements for calculating marginal likelihoods for GLMs with canonical links can be obtained

using a correction factor based on a sixth-order Laplace approximation at little increase in computational cost (Raudenbush et al. 2000; Sabanés Bové and Held 2011).

Propositions 2 and 5 establish that g-priors are well defined in the case of non-full rank designs (including the case $p_{\mathcal{M}} > n$). Furthermore under normality, Proposition 5 shows that Bayes factors for models with $p_{\mathcal{M}} > n$ compared to the null are equal to one. In these cases the prior on the model space plays a critical role in model averaging as well as stochastic search algorithms. Prior distributions, such as the sparsity priors of Yang et al. (2016) or truncated Poisson distributions, that place zero prior probability on models with $p_{\mathcal{M}} > n$, are potentially useful in controlling the model size for models that are not identifiable from the data. However, as MCMC algorithms may explore predominantly lower dimensional models that are full rank, g-priors and mixtures may have a computational advantage over the independent "spike and slab" priors, as inversion of the full p dimensional matrices can be avoided under the g-priors.

Proposition 5 can be used to extend the results of Sabanés Bové et al. (2015) who adopted the hyper-g/n prior to generalized additive models using splines in exponential families after marginalizing over the coefficients for the spline basis. This leads to a general linear model with a $\mathbf{W}$ that depends on unknown variance components. Rather than placing a prior directly on the variance components, they exploited the one-to-one correspondence between the degrees of freedom d and the variance components and considered a discrete set of values for d with the objective of learning models indexed by d. Proposition 5 also suggests new directions for BVS/BMA of linear predictors in spatial-temporal models or other covariance structures that depend on a low dimensional number of parameters.

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.

CHIC_examples.zip: R scripts for producing the simulation and real data results.

Acknowledgement

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.

Appell, P. (1925), "Sur les Fonctions Hypergéométriques de Plusieurs Variables," *In Mémoir. Sci. Math. Paris: Gauthier-Villars*.

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.

Burnham, K. P. and Anderson, D. R. (2004), "Multimodel Inference," *Sociological Methods & Research*, 33, 261–304.

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.

Christensen, R. (2011), *Plane Answers to Complex Questions*, Springer-Verlag.

Clyde, M. A. (2018), *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*, R package version 1.4.9. CRAN Comprehensive R Archive Network, DOI: 10.5281/zenodo.1212636 .

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.

Davidson, R. R. and Lever, W. E. (1970), "The Limiting Distribution of the Likelihood Ratio Statistic under a Class of Local Alternatives," *Sankhyā: The Indian Journal of Statistics. Series A*, 32, 209–224.

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

Fahrmeir, L. and Kaufmann, H. (1985), "Consistency and Asymptotic Normality of the Maximum Likelihood Estimator in Generalized Linear Models," *The Annals of Statistics*, 13, 342–368.

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.

Forte, A., Garcia-Donato, G., and Steel, M. (2016), "Methods and Tools for Bayesian Variable Selection and Model Averaging in Univariate Linear Regression," *ArXiv e-prints*.

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., Li, Y., and Mitra, R. (2017), "On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression," *Bayesian Analysis*, advance publication.

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.

Gradshteyn, I. S. and Ryzhik, I. M. (2007), *Table of Integrals, Series, and Products*, Academic Press, 7th ed.

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.

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.

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.

Olver, F. (1997), *Asymptotics and Special Functions*, A K Peters/CRC Press.

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*.

Raftery, A. E. (1996), "Approximate Bayes factors and accounting for model uncertainty in generalised linear models," *Biometrika*, 83, 251–266.

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.

Resnick, S. I. (1999), *A Probability Path*, Birkhäuser.

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.

Sabanés Bové, D., Held, L., and Kauermann, G. (2015), "Objective Bayesian Model Selection in Generalized Additive Models With Penalized Splines," *Journal of Computational and Graphical Statistics*, 24, 394–415.

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.

Self, S. and Mauritsen, R. (1988), "Power/Sample Size Calculations for Generalized Linear Models," *Biometrics*, 44, 79–86.

Self, S., Mauritsen, R., and Ohara, J. (1992), "Power Calculation for Likelihood Ratio Tests in Generalized Linear Models," *Biometrics*, 48, 31–39.

Shieh, G. (2000), "On Power and Sample Size Calculations for Likelihood Ratio Tests in Generalized Linear Models," *Biometrics*, 56, 1192–1196.

Slater, L. (1960), *Confluent Hypergeometric Functions*, Cambridge University Press.

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.

van der Vaart, A. W. (2000), *Asymptotic Statistics*, Cambridge University Press.

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.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016), "On the computational complexity of high-dimensional Bayesian variable selection," *Ann. Statist.*, 44, 2497–2532.

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.

Supplementary Materials: Appendices

A   Assumptions, Theoretical Results, and Proofs

A.1   Assumptions and Regularity Conditions

The following assumptions and standard regularity conditions are used throughout the paper unless specified otherwise.

For functions $b(\cdot)$ **and** $\theta(\cdot)$ in the GLM density (5), their third derivatives exist and are continuous on $\mathbb{R}$. The composite function $b' \circ \theta(\cdot)$, which links $\mathbb{E}(Y)$ and η, is strictly monotonic. The variance function $b'' \circ \theta(\cdot) \geq 0$, and the equality can only occur on the boundary $\pm\infty$.

Finite MLEs $\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ exist and are unique, under all subset models $\mathcal{M}$.

The design matrix X under the full model is known and has a full column rank p. Here, p is fixed. The column space $C(\mathbf{X})$ does not contain $\mathbf{1}_n$. When studying asymptotics, we assume that for $i = 1, \ldots, n$ the norm of the ith row $\|\mathbf{x}_i\|_2$ is bounded by a constant, and for all n, the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}/n$ is bounded from below by a positive constant. These conditions assure weak consistency (convergence in probability) and asymptotic normality for MLEs (Fahrmeir and Kaufmann 1985).

The true model $\mathcal{M}_T$ is among the 2^p subset models to be selected under consideration. In $\mathcal{M}_T$, true values of the intercept and regression coefficients are denoted by $\alpha^*_{\mathcal{M}_T}$ and $\boldsymbol{\beta}^*_{\mathcal{M}_T}$, respectively.

A.2 Proof of Proposition 1

Proof. We first approximate the likelihood by a second order Taylor expansion at the MLE,

$$p(\mathbf{Y} \mid \alpha, \boldsymbol{\beta}_{\mathcal{M}}, \mathcal{M})$$

$$\approx p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M})$$

$$\cdot \exp\left\{ -\frac{1}{2} \begin{bmatrix} \alpha - \hat{\alpha}_{\mathcal{M}} \\ \boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \end{bmatrix}^T \begin{bmatrix} \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n & \mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} \\ \mathbf{X}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n & \mathbf{X}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} \end{bmatrix} \begin{bmatrix} \alpha - \hat{\alpha}_{\mathcal{M}} \\ \boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \end{bmatrix} \right\}$$

$$= p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \exp\left\{ -\frac{1}{2} (\alpha - \hat{\alpha}_{\mathcal{M}} + \mathbf{m})^T \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right) (\alpha - \hat{\alpha}_{\mathcal{M}} + \mathbf{m}) \right.$$

$$\left. -\frac{1}{2} \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right)^T \boldsymbol{\Phi} \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right) \right\},$$

where the above approximation is precise up to a multiplicative term $[1 + O(n^{-1})]$, $\mathbf{m} = \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right)^{-1} \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} \right) \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right)$, and

$$\boldsymbol{\Phi} = \mathbf{X}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} - \left(\mathbf{X}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right) \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right)^{-1} \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} \right).$$

In the above approximate likelihood, the matrix $\boldsymbol{\Phi}$ acts like a precision matrix of $\boldsymbol{\beta}_{\mathcal{M}}$. By using the orthogonal projection $\hat{\mathcal{P}}_{\mathbf{1}_n} = \mathbf{1}_n \left(\mathbf{1}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1} \right)^{-1} \mathbf{1}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$, we can rewrite it as

$$\boldsymbol{\Phi} = \mathbf{X}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} - \mathbf{X}_{\mathcal{M}}^T \hat{\mathcal{P}}_{\mathbf{1}_n}^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \hat{\mathcal{P}}_{\mathbf{1}_n} \mathbf{X}_{\mathcal{M}}$$

$$= \mathbf{X}_{\mathcal{M}}^T (\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n})^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})(\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n}) \mathbf{X}_{\mathcal{M}} = \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}).$$

Under the flat prior $p(\alpha) \propto 1$, an integrated Laplace approximation yields the marginal

likelihood density conditional on $\boldsymbol{\beta}_{\mathcal{M}}$:

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

$$\propto \; p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right)^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right) \right\}$$

$$\cdot \int \exp \left\{ -\frac{1}{2} (\alpha - \hat{\alpha}_{\mathcal{M}} + \mathbf{m})^T \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right) (\alpha - \hat{\alpha}_{\mathcal{M}} + \mathbf{m}) \right\} d\alpha$$

$$\propto \; p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \left[\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n \right]^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right)^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \left(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right) \right\}.$$

$\square$

A.3 Asymptotic Behavior of the Observed Information

Lemma 1. *For any subset model* $\mathcal{M}$,

(1) *if* $\mathcal{M} \supset \mathcal{M}_T$, *then* $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = O_P(n)$ *and* $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = O_P(n)$. *More specifically,* $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})/n -$ $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})/n \xrightarrow{P} 0$, *and* $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n - \mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n \xrightarrow{P} \mathbf{0}$.

(2) *if* $\mathcal{M} \not\supset \mathcal{M}_T$, *then* $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = O_P(n^{\tau_{\mathcal{M}}})$ *and* $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = O_P(n^{\tau_{\mathcal{M}}})$, *where* $0 \leq \tau_{\mathcal{M}} \leq 1$.

Proof. First, we study the asymptotic of MLEs. The assumptions on the design matrix of the full model $\mathbf{X}$ remain to hold for the design matrix $\mathbf{X}_{\mathcal{M}}$ under all subset models, i.e., $\mathbf{x}_{\mathcal{M},i}$ are bounded for all $i = 1, \ldots, n$, and as n tends to infinity, the smallest eigenvalue of $\mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}}/n$ is bounded from below by a positive constant. Since these are stronger than the condition R_c in Fahrmeir and Kaufmann (1985, pp. 355), we have weak consistency and asymptotic normality for MLEs under any $\mathcal{M} \supset \mathcal{M}_T$, i.e., as $n \to \infty$,

$$\left(\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right) \xrightarrow{P} \left(\alpha_{\mathcal{M}}^*, \boldsymbol{\beta}_{\mathcal{M}}^* \right), \quad \mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}^*)^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}_T}^* \right) \xrightarrow{d} \mathrm{N}(0, \mathbf{I}_{p_{\mathcal{M}}}). \tag{46}$$

Here, $\alpha_{\mathcal{M}}^* = \alpha_{\mathcal{M}_T}^*$, and $\boldsymbol{\beta}_{\mathcal{M}}^* = \boldsymbol{\beta}_{\mathcal{M}_T}^*$ in the sense that all entries in $\boldsymbol{\beta}_{\mathcal{M}}^*$ that correspond to predictors not in $\mathcal{M}_T$ are filled with zero. Therefore, if $\mathcal{M} \supset \mathcal{M}_T$, then $\eta_{\mathcal{M},i}^* = \alpha_{\mathcal{M}}^* +$

$\mathbf{x}_{\mathcal{M},i}^T \boldsymbol{\beta}_{\mathcal{M}}^* = \alpha_{\mathcal{M}_T}^* + \mathbf{x}_{\mathcal{M}_T,i}^T \boldsymbol{\beta}_{\mathcal{M}_T}^* = \eta_{\mathcal{M}_T,i}^*$, for all $i = 1, \ldots, n$. On the other hand, if $\mathcal{M} \not\supseteq \mathcal{M}_T$, Self and Mauritsen (1988) and van der Vaart (2000, pp. 45, Theorem 5.7) suggest that the limits of MLEs still exist, i.e., $\left(\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) \xrightarrow{P} (\alpha_{\mathcal{M}}^*, \boldsymbol{\beta}_{\mathcal{M}}^*)$, but the linear predictors in the limit $\eta_{\mathcal{M},i}^* \neq \eta_{\mathcal{M}_T,i}^*$.

Under non-canonical links, observed information matrices are functions of $\mathbf{Y}$, therefore we need a weak law of large numbers for independently but non-identically distributed random variables. In Resnick (1999, pp. 205), by Theorem 7.2.1 and the proof of special case (a), we have that for a sequence of independent random variables $Y_1, \ldots, Y_n$, if their variances are bounded, then as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^n Y_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}(Y_i) \xrightarrow{P} 0. \tag{47}$$

Next we show asymptotic results for $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})$. In (10), for $i = 1, \ldots, n$, the ith diagonal entry of $\mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ can be rewritten as $d_i = b'' \circ \theta(\hat{\eta}_{\mathcal{M},i}) \left[\theta'(\hat{\eta}_{\mathcal{M},i})\right]^2 + \left[b' \circ \theta(\hat{\eta}_{\mathcal{M},i}) - Y_i\right] \theta''(\hat{\eta}_{\mathcal{M},i})$. Hence, for any model $\mathcal{M}$,

$$\frac{1}{n} \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = \frac{1}{n} \mathbf{1}_n \mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n = \frac{1}{n} \sum_{i=1}^n d_i$$

$$= \frac{1}{n} \left\{ \sum_{i=1}^n b'' \circ \theta(\hat{\eta}_{\mathcal{M},i}) \left[\theta'(\hat{\eta}_{\mathcal{M},i})\right]^2 + \left[b' \circ \theta(\hat{\eta}_{\mathcal{M},i}) - Y_i\right] \theta''(\hat{\eta}_{\mathcal{M},i}) \right\}$$

$$\xrightarrow{P} \frac{1}{n} \left\{ \sum_{i=1}^n b'' \circ \theta(\hat{\eta}_{\mathcal{M},i}) \left[\theta'(\hat{\eta}_{\mathcal{M},i})\right]^2 + \left[b' \circ \theta(\hat{\eta}_{\mathcal{M},i}) - b' \circ \theta(\eta_{\mathcal{M}_T,i}^*)\right] \theta''(\hat{\eta}_{\mathcal{M},i}) \right\}$$

$$\xrightarrow{P} \frac{1}{n} \left\{ \sum_{i=1}^n b'' \circ \theta(\eta_{\mathcal{M},i}^*) \left[\theta'(\eta_{\mathcal{M},i}^*)\right]^2 + \left[b' \circ \theta(\eta_{\mathcal{M},i}^*) - b' \circ \theta(\eta_{\mathcal{M}_T,i}^*)\right] \theta''(\eta_{\mathcal{M},i}^*) \right\}, \tag{48}$$

where the second last line is given by (47) and the fact $\mathbb{E}(Y_i) = b' \circ \theta(\eta_{\mathcal{M}_T,i}^*)$, for all $i = 1, \ldots, n$, and the last line is given by the continuous mapping theorem. Since for all $i = 1, \ldots, n$, $\mathbf{x}_i$ is bounded, $\eta_{\mathcal{M},i}^*$ and $\eta_{\mathcal{M}_T,i}^*$ are also bounded. For each term in the summation of (48), it is bounded due to the continuity assumptions on the third derivatives of $b(\cdot)$ and $\theta(\cdot)$. Therefore, $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})/n$ is bounded in probability.

If $\mathcal{M} \supset \mathcal{M}_T$, (48) becomes

$$\frac{1}{n}\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) \xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n} b'' \circ \theta(\eta^*_{\mathcal{M}_T,i}) \left[\theta'(\eta^*_{\mathcal{M}_T,i})\right]^2 = \frac{1}{n}\mathcal{I}_n(\alpha^*_{\mathcal{M}}), \qquad (49)$$

which is also the limit of $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})/n$. Because we assume that $b' \circ \theta(\cdot)$ is strictly monotonic, $\theta(\cdot)$ is also strictly monotonic. For each term in the summation of (49), it is positive because $\theta'(\cdot) \neq 0$ and $b'' \circ \theta(\eta)$ is positive for finite η. Therefore by (49), if $\mathcal{M} \supset \mathcal{M}_T$, then $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})/n$ is positive and bounded in probability, i.e., $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = O_P(n)$. On the other hand, if $\mathcal{M} \not\supset \mathcal{M}_T$, then only (48) holds but not (49). Each term in the summation of (48) can be either positive, zero, or negative. In this case, by (48), $\mathcal{J}_n(\alpha_{\mathcal{M}})/n$ is bounded in probability, and it may equal to zero. Therefore, $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})$ is on the order of $O(n^{\tau_{\mathcal{M}}})$, where $\tau_n \leq 1$, so that it tends to ∞ at a rate no faster than $O_P(n)$.

Last, we show asymptotic results regarding the matrix

$$\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^{cT} \mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{X}_{\mathcal{M}}^c = \mathbf{X}_{\mathcal{M}}^T (\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n})^T \mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}})(\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$$

$$= \mathbf{X}_{\mathcal{M}}^T \left[\mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) - \mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{1}_n \left(\mathbf{1}_n^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{1}_n\right)^{-1} \mathbf{1}_n^T \mathcal{J}(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \right] \mathbf{X}_{\mathcal{M}}.$$

For the (j,k)th entry, $1 \leq j < k \leq p_{\mathcal{M}}$,

$$\frac{1}{n}\left[\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\right]_{j,k} = \frac{1}{n}\sum_{i=1}^{n} d_i x_{i,j} x_{i,k} - \frac{1}{n}\left(\sum_{i=1}^{n} d_i x_{i,j}\right)\left(\sum_{i=1}^{n} d_i\right)^{-1}\left(\sum_{i=1}^{n} d_i x_{i,k}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} d_i x_{i,j} x_{i,k} - \left(\frac{1}{n}\sum_{i=1}^{n} d_i x_{i,j}\right)\left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} d_i x_{i,k}\right)$$

is bounded since all $\mathbf{x}_i$ are bounded. Therefore, $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n$ is bounded in probability.

To show that for any $\mathcal{M} \supset \mathcal{M}_T$, $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n$ does not reduce to zero, we will show that it is a positive definite matrix. For any given non-zero vector $\mathbf{a} \in \mathbb{R}^{p_{\mathcal{M}}}$, we denote $\mathbf{X}_{\mathcal{M}}\mathbf{a} = (t_1, \ldots, t_n)^T$, whose entries are all bounded. When $\mathcal{M} \supset \mathcal{M}_T$, by (49), all d_i's have a positive

lower bound, hence simple calculation gives

$$\frac{1}{n}\mathbf{a}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\mathbf{a} = \frac{1}{n}\sum_{i=1}^{n} d_i t_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} d_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} d_i t_i\right)^2 \geq 0.$$

Here the quality only holds if all t_i's are equal for $i = 1, \ldots, n$, which is impossible here because of the assumption $\mathbf{1}_n \notin C(\mathbf{X}_{\mathcal{M}})$. For large n, the assumption that the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}/n$ being bounded from below by a positive constant suggests that $\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}}/n$ is positive definite, so $\mathbf{a}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\mathbf{a}/n \nrightarrow 0$.

Furthermore, arguing similarly to (49), we also have

$$\frac{1}{n}\sum_{i=1}^{n} d_i t_i^k \xrightarrow{P} \frac{1}{n}\sum_{i=1}^{n} b'' \circ \theta(\eta_{\mathcal{M}_T,i}^*) \left[\theta'(\eta_{\mathcal{M}_T,i}^*)\right]^2 t_i^k,$$

for $k = 0, 1, 2$. Therefore, for any vector $\mathbf{a}$, if $\mathcal{M} \supset \mathcal{M}_T$, then

$$\frac{1}{n}\mathbf{a}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\mathbf{a} - \frac{1}{n}\mathbf{a}^T \mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\mathbf{a} \xrightarrow{P} 0.$$

i.e., $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n$ and $\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n$ are asymptotically the same. $\qquad\square$

A.4   Proof of Proposition 2

Proof. We first use proof by contradiction to show that for $\mathcal{M}$, the MLE of the intercept is unique. If both $(\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1)$ and $(\hat{\alpha}_2, \hat{\boldsymbol{\beta}}_2)$ maximize the likelihood for model $\mathcal{M}$, where $\hat{\alpha}_1 \neq \hat{\alpha}_2$, then

$$\hat{\alpha}_1\mathbf{1}_n + \mathbf{X}_{\mathcal{M}}\hat{\boldsymbol{\beta}}_1 = \hat{\alpha}_2\mathbf{1}_n + \mathbf{X}_{\mathcal{M}}\hat{\boldsymbol{\beta}}_2 \implies (\hat{\alpha}_1 - \hat{\alpha}_2)\mathbf{1}_n = \mathbf{X}_{\mathcal{M}}(\hat{\boldsymbol{\beta}}_2 - \hat{\boldsymbol{\beta}}_1),$$

which is contradicted with $\mathbf{1}_n \notin C(\mathbf{X}_{\mathcal{M}})$. Similarly, we can show this MLE is the same as the one for model $\mathcal{M}'$, i.e., $\hat{\alpha}_{\mathcal{M}} = \hat{\alpha}_{\mathcal{M}'}$.

By (25), between the two models $\mathcal{M}$ and $\mathcal{M}'$,

$$\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}}) = \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}'}), \quad z_{\mathcal{M}} = z_{\mathcal{M}'}.$$

So we just need to show $Q_{\mathcal{M}} = Q_{\mathcal{M}'}$. Since $\hat{\alpha}_{\mathcal{M}} = \hat{\alpha}_{\mathcal{M}'}$, (25) suggests that

$$\mathbf{x}_{\mathcal{M},i}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}} = \mathbf{x}_{\mathcal{M}',i}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}'}, \quad i = 1, \ldots, n.$$

Hence,

$$
\begin{aligned}
\mathbf{X}_{\mathcal{M}}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}} &= \mathbf{X}_{\mathcal{M}} \hat{\boldsymbol{\beta}}_{\mathcal{M}} - \left(\sum_{i=1}^{n} w_i \mathbf{x}_{\mathcal{M},i}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right) \mathbf{1}_n \\
&= \mathbf{X}_{\mathcal{M}'} \hat{\boldsymbol{\beta}}_{\mathcal{M}'} - \left(\sum_{i=1}^{n} w_i \mathbf{x}_{\mathcal{M}',i}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}'} \right) \mathbf{1}_n = \mathbf{X}_{\mathcal{M}'}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}'},
\end{aligned}
$$

where $w_i = d_i / (\sum_{r=1}^{n} d_r)$. Therefore, we have

$$Q_{\mathcal{M}} = \left[\mathbf{X}_{\mathcal{M}}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right]^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \left[\mathbf{X}_{\mathcal{M}}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}} \right] = \left[\mathbf{X}_{\mathcal{M}'}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}'} \right]^T \mathcal{J}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}'}) \left[\mathbf{X}_{\mathcal{M}'}^c \hat{\boldsymbol{\beta}}_{\mathcal{M}'} \right] = Q_{\mathcal{M}'}.$$

$\square$

A.5   Proof of Proposition 3

Proof. The marginal likelihood of the mixture of g-priors is obtained by integrating out g from the marginal likelihood of the g-prior, i.e.,

$$p(\mathbf{Y} \mid \mathcal{M}) = \int_0^{\infty} p(\mathbf{Y} \mid \mathcal{M}, g) p(g) dg$$

Here $p(\mathbf{Y} \mid \mathcal{M}, g)$ is obtained under the integrated Laplace approximation as in (19). Because of the one-to-one mapping between g and u, we rewrite this integral in terms of u.

$$
\begin{aligned}
p(\mathbf{Y} \mid \mathcal{M}) &= \int_0^1 p(\mathbf{Y} \mid \mathcal{M}, u) p(u) du \\
&\propto \int_0^1 p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} u^{\frac{p_{\mathcal{M}}}{2}} e^{-\frac{Q_{\mathcal{M}}}{2} u} \\
&\quad \cdot \frac{v^{\frac{a}{2}} \exp\left(\frac{s}{2v}\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \, \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)} \frac{u^{\frac{a}{2}-1}(1-vu)^{\frac{b}{2}-1} e^{-\frac{s}{2}u}}{[\kappa + (1-\kappa)vu]^r} \mathbf{1}_{\{0<u<\frac{1}{v}\}} \, du \\
&= p(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M}) \mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})^{-\frac{1}{2}} \frac{v^{\frac{a}{2}} \exp\left(\frac{s}{2v}\right)}{B\left(\frac{a}{2}, \frac{b}{2}\right) \, \Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1-\kappa\right)} \\
&\quad \cdot \int_0^1 \frac{u^{\frac{a+p_{\mathcal{M}}}{2}-1}(1-vu)^{\frac{b}{2}-1} e^{-\frac{s+Q_{\mathcal{M}}}{2}u}}{[\kappa + (1-\kappa)vu]^r} \mathbf{1}_{\{0<u<\frac{1}{v}\}} \, du.
\end{aligned}
$$

Since the above integrand is proportional to a tCCH density (27) with updated parameters, the above integral equals $B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \, \Phi_1\left(\frac{b}{2}, r, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2v}, 1-\kappa\right) v^{-\frac{a+p_{\mathcal{M}}}{2}} \exp\left(-\frac{s+Q_{\mathcal{M}}}{2v}\right)$.

$\square$

A.6 Proof of Proposition 4

Proof. The marginal prior on $\boldsymbol{\beta}_{\mathcal{M}}$ after integrating g out is

$$
p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}) \propto \int_0^\infty g^{-\frac{p_{\mathcal{M}}}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n}^2}{2g}\right] g^{\frac{b}{2}-1} \left(\frac{1}{1+g}\right)^{\frac{a+b}{2}} \exp\left[\frac{sg}{2(1+g)}\right] dg \qquad (50)
$$

We will show that as $\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n} \to \infty$, both a lower bound and an upper bound of (50) are proportional to $\left(\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n}^2\right)^{-\frac{a+p_{\mathcal{M}}}{2}}$. Since $s \geq 0$, a lower bound of of the right side of (50) is

$$
\int_0^\infty g^{-\frac{p_{\mathcal{M}}}{2}} e^{-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n}^2}{2g}} g^{\frac{b}{2}-1} \left(\frac{1}{1+g}\right)^{\frac{a+b}{2}} dg = \int_0^\infty \left(\frac{g}{1+g}\right)^{\frac{a+b}{2}} \left(\frac{1}{g}\right)^{\frac{a+p_{\mathcal{M}}-2}{2}} e^{-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n}^2}{2g}} d\left(\frac{1}{g}\right).
$$

Then according to the Watson's Lemma (Olver 1997, pp. 71), as $\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n} \to \infty$, the limit of this lower bound is proportional to $\left(\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{J}_n}^2\right)^{-\frac{a+p_{\mathcal{M}}}{2}}$. Next we find an upper bound of the

right side of (50) as

$$\int_0^\infty g^{-\frac{p_\mathcal{M}}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2(1+g)}\right] g^{\frac{b}{2}-1} \left(\frac{1}{1+g}\right)^{\frac{a+b}{2}} \exp\left[\frac{sg}{2(1+g)}\right] dg$$

$$= e^{-\frac{\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2}} B\left(\frac{b-p_\mathcal{M}}{2}, \frac{a+p_\mathcal{M}}{2}\right) {}_1F_1\left(\frac{b-p_\mathcal{M}}{2}, \frac{a+b}{2}, \frac{s+\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2}\right).$$

According to Abramowitz and Stegun (1970) formula (13.1.4),

$$_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(a)} \exp(s) s^{a-b}[1 + O(|s|^{-1})], \text{ when } \text{Real}(s) > 0, \tag{51}$$

hence as $\|\boldsymbol{\beta}_\mathcal{M}\|_{\mathcal{J}_n} \to \infty$, the limit of the above upper bound converges to

$$\exp\left[-\frac{\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2}\right] \Gamma\left(\frac{a+p_\mathcal{M}}{2}\right) \exp\left[\frac{s+\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2}\right] \cdot \left(\frac{s+\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}}{2}\right)^{-\frac{a+p_\mathcal{M}}{2}} \propto \left(\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}\right)^{-\frac{a+p_\mathcal{M}}{2}}.$$

Therefore, as $\|\boldsymbol{\beta}_\mathcal{M}\|_{\mathcal{J}_n}$ increases, or equivalently, as $\|\boldsymbol{\beta}_\mathcal{M}\|$ increases, both the lower bound and upper bound of $p(\boldsymbol{\beta}_\mathcal{M} \mid \mathcal{M})$ are proportional to $\left(\|\boldsymbol{\beta}_\mathcal{M}\|^2_{\mathcal{J}_n}\right)^{-\frac{a+p_\mathcal{M}}{2}}$. □

A.7 Special Functions: Definition and Useful Properties

We first review a list of special functions, including their definitions and relevant properties, that will be needed in the proof of Proposition 5.

- Confluent hypergeometric function (Abramowitz and Stegun 1970, eq 13.2.1): for $\gamma > \alpha > 0$,

$$_1F_1(\alpha, \gamma, x) = \frac{1}{\text{B}(\gamma - \alpha, \alpha)} \int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1} e^{xu} \, du.$$

  - By (Abramowitz and Stegun 1970, eq 13.2.27): ${}_1F_1(\alpha, \gamma, x) = e^x \cdot {}_1F_1(\gamma - \alpha, \gamma, -x)$.

  - By (Abramowitz and Stegun 1970, eq 6.5.12), the incomplete Gamma function:

$$\gamma(a, s) = \int_0^s t^{a-1} e^{-t} dt = {}_1F_1(a, a+1, -s)\frac{s^a}{a}.$$

- $_1F_1(\alpha, \gamma, 0) = 1$.

- Confluent hypergeometric function of two variables (Gordy 1998b)[5]: for $\gamma > \alpha > 0$ and $y < 1$,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \frac{1}{B(\gamma - \alpha, \alpha)} \int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1}(1-yu)^{-\beta}e^{xu} \, du,$$

Special cases:

  - If $x = 0$, then $\Phi_1(\alpha, \beta, \gamma, 0, y) = \, _2F_1(\beta, \alpha; \gamma; y)$.

  - If $\beta = 0$ or $y = 0$, then $\Phi_1(\alpha, 0, \gamma, x, y) = \Phi_1(\alpha, \beta, \gamma, x, 0) = \Phi_1(\alpha, 0, \gamma, x, 0) = \, _1F_1(\alpha, \gamma, x)$.

  - If $x = 0$ and $y = 0$, then $\Phi_1(\alpha, \beta, \gamma, 0, 0) = 1$.

- Hypergeometric function (Abramowitz and Stegun 1970, eq 15.3.1): for $\gamma > \alpha > 0$

$$_2F_1(\beta, \alpha; \gamma; x) = \frac{1}{B(\gamma - \alpha, \alpha)} \int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1}(1-xu)^{-\beta} \, du.$$

  - By (Abramowitz and Stegun 1970, eq 15.3.3): in the definition of $_2F_1$ function above, let $w = \frac{1-u}{1-xu}$, then

$$_2F_1(\beta, \alpha; \gamma; x) = (1-x)^{\gamma-\beta-\alpha} \, _2F_1(\gamma - \beta, \gamma - \alpha; \gamma; x)$$

  - $_2F_1(0, \alpha; \gamma, x) = \, _2F_1(\beta, \alpha; \gamma, 0) = 1$

  - $_2F_1(\beta, 1; \beta, x) = (1-x)^{-1} \, _2F_1(0, \beta-1; \beta, x) = (1-x)^{-1}$

  - By (Abramowitz and Stegun 1970, eq 15.3.4): $_2F_1(\beta, \alpha; \gamma; x) = (1-x)^{-\beta} \, _2F_1\left(\beta, \gamma - \alpha; \gamma, \frac{x}{x-1}\right)$

  - By (Abramowitz and Stegun 1970, eq 15.3.5): $_2F_1(\beta, \alpha; \gamma; x) = (1-x)^{-\alpha} \, _2F_1\left(\alpha, \gamma - \beta; \gamma, \frac{x}{x-1}\right)$

---

[5]Note: the definition in Gordy (1998b) is slightly different from that in Gradshteyn and Ryzhik (2007).

- Hypergeometric function of two variables (Appell function) (Weisstein 2009): for $\gamma > \alpha > 0$,

$$F_1(\alpha; \beta, \beta'; \gamma; x, y) = \frac{1}{\mathrm{B}(\gamma - \alpha, \alpha)} \int_0^1 u^{\alpha-1}(1-u)^{\gamma-\alpha-1}(1-xu)^{-\beta}(1-yu)^{-\beta'} \, du.$$

A.8 Proof of Proposition 5

Proof. To begin we establish that the marginal likelihood conditional on g is well defined under the g-prior when the design matrix is not full rank for a general linear model. We will assume the inner product space defined by the vector space $\mathbb{R}^n$ equipped with inner product $\mathbf{u}^T\mathbf{W}\mathbf{v}$ for two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ where $\mathbf{W}$ is a real, $n \times n$ symmetric positive definite matrix. Similarly, $\|\mathbf{u}\|_{\mathbf{W}}^2 \equiv \mathbf{u}^T\mathbf{W}\mathbf{u}$.

For the model

$$\mathbf{Y} = \mathbf{1}_n\beta_0 + \mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} + \boldsymbol{\epsilon}, \quad \text{with } \boldsymbol{\epsilon} \mid \phi \sim \mathsf{N}(\mathbf{0}_n, \phi^{-1}\mathbf{W}^{-1}),$$

let $\mathcal{P}_{\mathbf{1}} = \mathbf{1}_n(\mathbf{1}_n^T\mathbf{W}\mathbf{1}_n)^{-1}\mathbf{1}_n^T\mathbf{W}$ denote the orthogonal projection onto the column space of $\mathbf{1}_n$ and without loss of generality reparameterize the model

$$\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}_{\mathcal{M}}^c\boldsymbol{\beta}_{\mathcal{M}} + \boldsymbol{\epsilon}$$

where $\mathbf{X}_{\mathcal{M}}^c = (\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{X}_{\mathcal{M}}$ and $\alpha \equiv \beta_0 - (\mathbf{1}_n^T\mathbf{W}\mathbf{1}_n)^{-1}\mathbf{1}_n^T\mathbf{W}\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}$. Adopting the g-prior of the form

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \alpha, \phi, g \sim \mathsf{N}\left(0, \frac{g}{\phi}(\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}\mathbf{X}_{\mathcal{M}}^c)^-\right),$$

where $(\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}\mathbf{X}_{\mathcal{M}}^c)^-$ is any generalized inverse, standard normal theory for the linear combination $\mathbf{X}_{\mathcal{M}}^c\boldsymbol{\beta}_{\mathcal{M}} + \boldsymbol{\epsilon}$ can be used to show that $\mathbf{Y}$ is equal in distribution

$$\mathbf{Y} \mid \alpha, \phi, g, \mathcal{M} \sim \mathsf{N}\left(\mathbf{1}_n\alpha, \phi^{-1}(\mathbf{I}_n + g\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c})\mathbf{W}^{-1}\right) \tag{52}$$

where $\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c} = \mathbf{X}_{\mathcal{M}}^c(\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}\mathbf{X}_{\mathcal{M}}^c)^-\mathbf{X}_{\mathcal{M}}^{cT}\mathbf{W}$ is the $\rho_{\mathcal{M}} \leq p_{\mathcal{M}}$ orthogonal projection onto the column space $\mathbf{X}_{\mathcal{M}}^c$ in the inner product space. As the projection $\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c}$ does not depend on the choice of generalized inverse, this establishes that the marginal likelihood for the model will not depend on the choice of generalized inverse employed in defining the g-prior.

Continuing with integration with respect to α, ϕ under the independent Jeffreys prior $p(\alpha, \phi) \propto \phi^{-1}$,

$$p(\mathbf{Y} \mid g, \mathcal{M}) = \iint (2\pi)^{-\frac{n}{2}} |\mathbf{I}_n + g\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c}|^{-\frac{1}{2}} |\mathbf{W}|^{\frac{1}{2}} \phi^{\frac{n}{2}-1} e^{-\frac{\phi}{2}\left\{(\mathbf{Y}-\mathbf{1}_n\alpha)^T\mathbf{W}(\mathbf{I}_n - \frac{g}{1+g}\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c})(\mathbf{Y}-\mathbf{1}_n\alpha)\right\}} d\alpha \, d\phi \tag{53}$$

rearrangement of terms can be used to show that

$$p(\mathbf{Y} \mid g, \mathcal{M}) = p(\mathbf{Y} \mid \mathcal{M}_\emptyset)(1 + g)^{\frac{n-\rho_{\mathcal{M}}-1}{2}} \left\{1 + g(1 - R_{\mathcal{M}}^2)\right\}^{-\frac{n-1}{2}}$$

where $R_{\mathcal{M}}^2$ is defined in (40) and

$$p(\mathbf{Y} \mid \mathcal{M}_\emptyset) = (2\pi)^{-\frac{n-1}{2}} \Gamma\left(\frac{n-1}{2}\right) |\mathbf{W}|^{\frac{1}{2}} (\mathbf{1}_n^T\mathbf{W}\mathbf{1}_n)^{-\frac{1}{2}} \left[\frac{\|(\mathbf{I}_n - \mathcal{P}_{\mathbf{1}_n})\mathbf{Y}\|_{\mathbf{W}}^2}{2}\right]^{-\frac{n-1}{2}}$$

is the marginal under the null model. Note that in (53), the determinant $|\mathbf{I}_n + g\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c}| = (1+g)^{\rho_{\mathcal{M}}}$, because the eigenvalues of the orthogonal projection $\mathcal{P}_{\mathbf{X}_{\mathcal{M}}^c}$ are one with a multiplicity of $\rho_{\mathcal{M}}$ and zero with a multiplicity of $p_{\mathcal{M}} - \rho_{\mathcal{M}}$. The Bayes Factor for comparing $\mathcal{M}$ to $\mathcal{M}_\emptyset$ is thus

$$\mathsf{BF}[\mathcal{M}, \mathcal{M}_\emptyset] = (1 + g)^{\frac{n-\rho_{\mathcal{M}}-1}{2}} \left\{1 + g(1 - R_{\mathcal{M}}^2)\right\}^{-\frac{n-1}{2}},$$

which will be one for any model $\mathcal{M}$ where $R_{\mathcal{M}}^2 = 1$ and $\rho_{\mathcal{M}} = n - 1$.

For simplicity in the rest of proof, we omit the subscript $\mathcal{M}$ when there is no ambiguity. We now show part (1). In the tCCH distribution, if $r = 0$ or $\kappa = 1$, then

$$\Phi_1\left(\frac{b}{2}, r, \frac{a+b}{2}, \frac{s}{2v}, 1 - \kappa\right) = \Phi_1\left(\frac{b}{2}, 0, \frac{a+b}{2}, \frac{s}{2v}, 0\right) = {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, \frac{s}{2v}\right).$$

Then the marginal likelihood becomes

$$
\begin{aligned}
p(\mathbf{Y} \mid \mathcal{M}) &= \frac{p(\mathbf{Y} \mid \mathcal{M}_\varnothing)\, v^{\frac{a}{2}} \exp\left(\frac{s}{2v}\right)}{B\left(\frac{a}{2},\frac{b}{2}\right)\, {}_1F_1\left(\frac{b}{2},\frac{a+b}{2},\frac{s}{2v}\right)} \int_0^{1/v} \frac{u^{\frac{a+\rho}{2}-1}(1-vu)^{\frac{b}{2}-1}e^{-\frac{su}{2}}}{\left[(1-R^2)+R^2 u\right]^{\frac{n-1}{2}}}\, du \\
&= \frac{p(\mathbf{Y} \mid \mathcal{M}_\varnothing)\, v^{\frac{a}{2}} \exp\left(\frac{s}{2v}\right)}{B\left(\frac{a}{2},\frac{b}{2}\right)\, {}_1F_1\left(\frac{b}{2},\frac{a+b}{2},\frac{s}{2v}\right)} \int_0^{1/v} \frac{u^{\frac{a+\rho}{2}-1}(1-vu)^{\frac{b}{2}-1}e^{-\frac{su}{2}}}{\left\{\left[1-\left(1-\frac{1}{v}\right)R^2\right]\left[\frac{1-R^2}{1-\left(1-\frac{1}{v}\right)R^2}+\frac{R^2/v}{1-\left(1-\frac{1}{v}\right)R^2}\cdot(vu)\right]\right\}^{\frac{n-1}{2}}}\, du \\
&= \frac{p(\mathbf{Y} \mid \mathcal{M}_\varnothing)\, v^{\frac{a}{2}} \exp\left(\frac{s}{2v}\right)}{B\left(\frac{a}{2},\frac{b}{2}\right)\, {}_1F_1\left(\frac{b}{2},\frac{a+b}{2},\frac{s}{2v}\right)} \cdot \frac{B\left(\frac{a+\rho}{2},\frac{b}{2}\right)\,\Phi_1\left(\frac{b}{2},\frac{n-1}{2},\frac{a+b+\rho}{2},\frac{s}{2v},\frac{R^2/v}{1-\left(1-\frac{1}{v}\right)R^2}\right)}{v^{\frac{a+\rho}{2}}\exp\left(\frac{s}{2v}\right)\left[1-\left(1-\frac{1}{v}\right)R^2\right]^{\frac{n-1}{2}}} \\
&= p(\mathbf{Y} \mid \mathcal{M}_\varnothing) \cdot \frac{B\left(\frac{a+\rho}{2},\frac{b}{2}\right)\,\Phi_1\left(\frac{b}{2},\frac{n-1}{2},\frac{a+b+\rho}{2},\frac{s}{2v},\frac{R^2/v}{1-\left(1-\frac{1}{v}\right)R^2}\right)}{v^{\frac{\rho}{2}}\left[1-\left(1-\frac{1}{v}\right)R^2\right]^{\frac{n-1}{2}} B\left(\frac{a}{2},\frac{b}{2}\right)\, {}_1F_1\left(\frac{b}{2},\frac{a+b}{2},\frac{s}{2v}\right)}.
\end{aligned}
$$

Here the second last equality is given by the propriety of the tCCH density function (27).

Then we show part (2). In the tCCH distribution, when $s = 0$, then

$$
\Phi_1\left(\frac{b}{2},r,\frac{a+b}{2},0,1-\kappa\right) = {}_2F_1\left(r,\frac{b}{2},\frac{a+b}{2};1-\kappa\right).
$$

Hence, the marginal likelihood becomes

$$
p(\mathbf{Y} \mid \mathcal{M}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_\varnothing)\, v^{\frac{a}{2}}}{B\left(\frac{a}{2},\frac{b}{2}\right)\, {}_2F_1\left(r,\frac{b}{2};\frac{a+b}{2};1-\kappa\right)} \int_0^{1/v} \frac{u^{\frac{a+\rho}{2}-1}(1-vu)^{\frac{b}{2}-1}}{\left[(1-R^2)+R^2 u\right]^{\frac{n-1}{2}}\left[\kappa+(1-\kappa)vu\right]^r}\, du
\tag{54}
$$

For simplification, we denote $x = 1-1/\kappa$ and $w = 1-(1-vu)/(1-xvu)$. By change of variable,

$$
u = \frac{w}{v(1-x+xw)}, \quad \frac{du}{dw} = \frac{1-x}{v(1-x+xw)^2},
$$

and the integral in (54) is

$$\int_0^{1/v} \frac{u^{\frac{a+\rho}{2}-1}(1-vu)^{\frac{b}{2}-1}}{[(1-R^2)+R^2u]^{\frac{n-1}{2}}[\kappa+(1-\kappa)vu]^r}\,du$$

$$= \int_0^1 \frac{\left[\frac{w}{v(1-x+xw)}\right]^{\frac{a+\rho}{2}-1}\left[\frac{(1-x)(1-w)}{1-x+xw}\right]^{\frac{b}{2}-1}\frac{1-x}{v(1-x+xw)^2}}{\left\{\frac{(1-R^2)v(1-x)+[(1-R^2)vx+R^2]w}{v(1-x+xw)}\right\}^{\frac{n-1}{2}}\left(\frac{1}{1-x+xw}\right)^r}\,dw$$

$$= \frac{(1-x)^{\frac{b}{2}}\,v^{\frac{n-1-a-\rho}{2}}}{[(1-R^2)v(1-x)]^{\frac{n-1}{2}}\,(1-x)^{\frac{a+b+\rho+1-n-2r}{2}}}\int_0^1 \frac{w^{\frac{a+\rho}{2}-1}(1-w)^{\frac{b}{2}-1}}{\left[1-\frac{(1-R^2)vx+R^2}{(1-R^2)v(x-1)}w\right]^{\frac{n-1}{2}}\left(1-\frac{x}{x-1}w\right)^{\frac{a+b+\rho+1-n-2r}{2}}}\,dw$$

$$= \frac{\kappa^{\frac{a+\rho-2r}{2}}\,v^{-\frac{a+\rho}{2}}}{(1-R^2)^{\frac{n-1}{2}}}B\left(\frac{a+\rho}{2},\frac{b}{2}\right)\cdot$$

$$F_1\left(\frac{a+\rho}{2};\frac{a+b+\rho+1-n-2r}{2},\frac{n-1}{2};\frac{a+b+\rho}{2};1-\kappa,\frac{(1-R^2)v(1-\kappa)-R^2\kappa}{(1-R^2)v}\right).$$

$\square$

A.9 Derivation of (44)

Proof. Similar to (19), we apply integrated Laplace approximation to obtain $p(\mathbf{Y}\mid\phi,\mathcal{M},g)$, then marginalize ϕ out as follows.

$$p(\mathbf{Y}\mid\mathcal{M},g) = \int_0^\infty p(\mathbf{Y}\mid\phi,\mathcal{M},g)p(\phi)d\phi$$

$$\propto \int_0^\infty p(\mathbf{Y}\mid\hat{\alpha}_{\mathcal{M}},\hat{\boldsymbol{\beta}}_{\mathcal{M}},\phi,\mathcal{M})\,[\phi\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})]^{-\frac{1}{2}}\,(1+g)^{-\frac{p_{\mathcal{M}}}{2}}e^{-\frac{\phi Q_{\mathcal{M}}}{2(1+g)}}\phi^{-1}d\phi$$

$$\propto [\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})]^{-\frac{1}{2}}\,(1+g)^{-\frac{p_{\mathcal{M}}}{2}}\int_0^\infty \phi^{\frac{n-1}{2}-1}e^{\phi\left\{-\frac{Q_{\mathcal{M}}}{2(1+g)}+\sum_{i=1}^n\left[Y_i(\hat{\theta}_i-t_i)-b(\hat{\theta}_i)+b(t_i)\right]\right\}}d\phi$$

$$\propto [\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}})]^{-\frac{1}{2}}\,(1+g)^{-\frac{p_{\mathcal{M}}}{2}}\left\{\frac{Q_{\mathcal{M}}}{2(1+g)}-\sum_{i=1}^n\left[Y_i(\hat{\theta}_i-t_i)-b(\hat{\theta}_i)+b(t_i)\right]\right\}^{-\frac{n-1}{2}}$$

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

Here, the last step replaces g with $u = 1/(1+g)$. $\square$

A.10 Proof of Model Selection Consistency

We first show a lemma about a non-central χ^2 distribution, which is useful to prove some of the following lemmas and theorems. Here the symbol $\chi_k^2(m)$ denotes a non-central χ^2 distribution with degrees of freedom k and non-centrality parameter m.

Lemma 2. *If a sequence of random variables $\{X_n : n = 1, 2, \ldots\}$ have independent non-central χ^2 distributions: $X_n \sim \chi_k^2(nA_n)$, where random variables $A_n \xrightarrow{D} a_0 \in \mathbb{R}^+ \cup \{0\}$, then as $n \longrightarrow \infty$, $X_n/n \xrightarrow{P} a_0$.*

Proof. For any $n \in \mathbb{N}$, the characteristic function of X_n/n evaluated at $t \in \mathbb{R}$ is

$$
\begin{aligned}
\phi_{X_n/n}(t) &= \mathbb{E}\left(e^{itX_n/n}\right) = \mathbb{E}_{A_n}\left[\mathbb{E}\left(e^{itX_n/n} \mid A_n\right)\right] \\
&= \mathbb{E}_{A_n}\left[\exp\left(\frac{itA_n}{1 - 2it/n}\right)(1 - 2it/n)^{-\frac{k}{2}}\right] = (1 - 2it/n)^{-\frac{k}{2}} \cdot \mathbb{E}_{A_n}\left[\exp\left(\frac{itA_n}{1 - 2it/n}\right)\right].
\end{aligned}
$$

Denote a complex valued random variable $B_n = A_n/(1 - 2it/n)$. Since the limit of A_n is a constant, for the series $\{A_n : n \in \mathbb{N}\}$, convergence in distribution is equivalent to convergence in probability. Because of the continuous mapping theorem, $B_n \xrightarrow{P} a_0$, or equivalently, convergence in distribution. Denote the bounded and continuous function $h(B_n) = \exp(itB_n)$, then according to Portmanteau lemma, $\mathbb{E}[h(B_n)] \longrightarrow \mathbb{E}[h(a_0)] = h(a_0)$. So for any $t \in \mathbb{R}$,

$$
\lim_{n\to\infty} \phi_{X_n/n}(t) = \lim_{n\to\infty}(1 - 2it/n)^{-k/2} \cdot \lim_{n\to\infty} \mathbb{E}[h(B_n)] = h(a_0) = \exp(ita_0),
$$

where the limit is the characteristic function of a degenerated distribution at a_0. Therefore, X_n/n converge in distribution to a constant a_0, which implies convergence in probability.

$\square$

In order to show the asymptotic performance of the Bayes factor $\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}}$, we first study asymptotic behaviors of the terms in the Bayes factors in the following lemmas. When testing nested models, the log likelihood ratio between $\mathcal{M}_T$ and $\mathcal{M}$ converges in distribution to a

central (non-central) χ^2 distribution, when the smaller (larger) model is true. The following lemma studies asymptotic behaviors of the likelihood ratio, which does not require models $\mathcal{M}$ and $\mathcal{M}_T$ to be nested.

Lemma 3. *Denote the the likelihood ratio by*

$$\Lambda_{\mathcal{M}_T:\mathcal{M}} \overset{\triangle}{=} \frac{p(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_T}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_T}, \mathcal{M}_T)}{p(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}, \mathcal{M})} = \exp\left(\frac{z_{\mathcal{M}_T} - z_{\mathcal{M}}}{2}\right) \tag{55}$$

As the sample size n increases,

1) *if $\mathcal{M}_T \subset \mathcal{M}$, then $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O_P(1)$.*

2) *if $\mathcal{M}_T \not\subset \mathcal{M}$, then $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O_P(e^{c_{\mathcal{M}} n})$, where $c_{\mathcal{M}}$ is a positive constant.*

Proof. In the first case where $\mathcal{M} \supset \mathcal{M}_T$, from the well-known results of likelihood ratio test, $z_{\mathcal{M}} - z_{\mathcal{M}_T}$ has a central chi-square distribution $\chi^2_{p_{\mathcal{M}} - p_{\mathcal{M}_T}}$. Therefore, the limiting distribution of the log-likelihood ratio does not depend on n, i.e., $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O_P(1)$.

In the second case where $\mathcal{M} \not\supset \mathcal{M}_T$, we first examine the sub-case where $\mathcal{M} \subset \mathcal{M}_T$. According to the power calculation results for GLM in Self et al. (1992) and Shieh (2000), when testing nested models, if the larger model is true, then we have that $z_{\mathcal{M}_T} - z_{\mathcal{M}}$ converges in distribution to a non-central χ^2 of degrees of freedom $p_{\mathcal{M}_T} - p_{\mathcal{M}}$. The non-centrality parameter Ψ is approximately

$$\Psi \approx \sum_{i=1}^{n} b'(\theta^*_{\mathcal{M}_T,i}) \left(\theta^*_{i,\mathcal{M}_T} - \theta^*_{i,\mathcal{M}}\right) - \left[b(\theta^*_{i,\mathcal{M}_T}) - b(\theta^*_{i,\mathcal{M}})\right],$$

where $\theta^*_{i,\mathcal{M}} = \theta(\eta^*_{i,\mathcal{M}})$, for $i = 1, \ldots, n$. By a Taylor expansion, there exist a $\tilde{\theta}_i$ between $\theta^*_{\mathcal{M}_T,i}$ and $\theta^*_{\mathcal{M},i}$, such that $b(\theta^*_{i,\mathcal{M}}) = b(\theta^*_{i,\mathcal{M}_T}) + b'(\theta^*_{\mathcal{M}_T,i})\left(\theta^*_{i,\mathcal{M}_T} - \theta^*_{i,\mathcal{M}}\right) + b''(\tilde{\theta}_i)\left(\theta^*_{i,\mathcal{M}_T} - \theta^*_{i,\mathcal{M}}\right)^2/2$. This combined with the assumption $b''(\cdot) > 0$ gives that $\lim_{n\to\infty} \Psi/n$ converges to a positive constant $c_{\mathcal{M}}$. Then by Lemma 2, $(z_{\mathcal{M}_T} - z_{\mathcal{M}})/n \overset{P}{\longrightarrow} c_{\mathcal{M}}$, and hence $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O_P(e^{c_{\mathcal{M}} n})$.

In the case where $\mathcal{M}$ and $\mathcal{M}_T$ are not nested, we introduce a third model $\mathcal{M}'$ which

includes all the predictors in both $\mathcal{M}$ and $\mathcal{M}_T$. Using a similar method as in Self et al. (1992), we can treat $\mathcal{M}'$ also as the true model (although with some redundant predictors) when comparing with $\mathcal{M}$ and easily show that $\Lambda_{\mathcal{M}':\mathcal{M}}$ also has a non-central χ^2 distribution. Hence we decompose $\Lambda_{\mathcal{M}_T:\mathcal{M}} = \Lambda_{\mathcal{M}_T:\mathcal{M}'} \cdot \Lambda_{\mathcal{M}':\mathcal{M}}$. Since both pairs $(\mathcal{M}_T, \mathcal{M}')$ and $(\mathcal{M}' : \mathcal{M})$ are nested models, we can apply the previous results twice: $\Lambda_{\mathcal{M}_T:\mathcal{M}'} = O_P(1)$ and $\Lambda_{\mathcal{M}':\mathcal{M}} = O_P(e^{c_{\mathcal{M}} n})$. Therefore, we can conclude that $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O_P(1) \cdot O_P(e^{c_{\mathcal{M}} n}) = O_P(e^{c_{\mathcal{M}} n})$. □

The Bayes factors contain the Wald statistics $Q_{\mathcal{M}_T}$ and $Q_{\mathcal{M}}$. We next study their asymptotic behaviors.

Lemma 4. *The Wald statistic $Q_{\mathcal{M}} = O_P(n^{\xi_{\mathcal{M}}})$, where $0 \le \xi_{\mathcal{M}} \le 1$. In particular,*

1) *If $\mathcal{M}_T \ne \mathcal{M}_{\emptyset}$, then for any $\mathcal{M} \supset \mathcal{M}_T$, $\xi_{\mathcal{M}} = 1$.*

2) *if $\mathcal{M}_T = \mathcal{M}_{\emptyset}$, then for any model $\mathcal{M}$, $\xi_{\mathcal{M}} = 0$.*

Proof. For any $\mathcal{M} \supset \mathcal{M}_T$, we have shown in the proof of Lemma 1 that the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ converges in probability to the true value $\boldsymbol{\beta}_{\mathcal{M}}^*$, and $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})/n$ is a finite positive definite matrix and converges to $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}^*)/n$ in probability. By Lemma 1 and Slutsky's theorem, we can rewrite the asymptotic normality (46) as

$$\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{\frac{1}{2}} \left(\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}}^* \right) \xrightarrow{d} \mathrm{N}(0, \mathbf{I}_{p_{\mathcal{M}}}).$$

Therefore, $Q_{\mathcal{M}} = \hat{\boldsymbol{\beta}}_{\mathcal{M}}^T \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ converges in distribution to a non-central χ^2 random variable with degrees of freedom $p_{\mathcal{M}}$ and non-centrality parameter $\boldsymbol{\beta}_{\mathcal{M}}^{*T} \mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}^*) \boldsymbol{\beta}_{\mathcal{M}}^*$, which is $O(n)$ if $\boldsymbol{\beta}_{\mathcal{M}}^* \ne \mathbf{0}$, and zero otherwise. Since $\boldsymbol{\beta}_{\mathcal{M}}^* = \boldsymbol{\beta}_{\mathcal{M}_T}^*$ in the sense that all entries in $\boldsymbol{\beta}_{\mathcal{M}}^*$ that correspond to predictors not in $\mathcal{M}_T$ are filled with zero, $\boldsymbol{\beta}_{\mathcal{M}}^* = \mathbf{0}$ is equivalent to $\mathcal{M}_T = \mathcal{M}_{\emptyset}$. Therefore, by Lemma 2, if $\mathcal{M}_T \ne \mathcal{M}_{\emptyset}$, then $Q_{\mathcal{M}} = O_P(n)$; if $\mathcal{M}_T = \mathcal{M}_{\emptyset}$, then $Q_{\mathcal{M}} = O_P(1)$.

For any $\mathcal{M} \not\supset \mathcal{M}_T$, since convergence in probability is preserved under addition and multiplication (Resnick 1999, pp. 175), we have $Q_{\mathcal{M}} - \boldsymbol{\beta}_{\mathcal{M}}^{*T} \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \boldsymbol{\beta}_{\mathcal{M}}^* \xrightarrow{P} 0$, i.e., $Q_{\mathcal{M}}$ is at

most on the same order of $\mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})$. By Lemma 1, we have $\xi_{\mathcal{M}} = \tau_{\mathcal{M}}$ if $\boldsymbol{\beta}_{\mathcal{M}}^* \neq \mathbf{0}$, and $\xi_{\mathcal{M}} = 0$ if $\boldsymbol{\beta}_{\mathcal{M}}^* = \mathbf{0}$. $\qquad\square$

Based on the results of Lemma 4, the next lemma discusses the asymptotic properties of $\Omega_{\mathcal{M}_T:\mathcal{M}}^{\text{CH}}$, a term that appears in the Bayes factor under the CH prior.

Lemma 5. *Under the CH prior, denote the term in $\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}}$:*

$$\Omega_{\mathcal{M}_T:\mathcal{M}}^{CH} \triangleq \frac{B\left(\frac{a+p_{\mathcal{M}_T}}{2}, \frac{b}{2}\right) \,{}_1F_1\left(\frac{a+p_{\mathcal{M}_T}}{2}, \frac{a+b+p_{\mathcal{M}_T}}{2}, -\frac{s+Q_{\mathcal{M}_T}}{2}\right)}{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) \,{}_1F_1\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)}. \tag{56}$$

1) If $\mathcal{M}_T \neq \mathcal{M}_\varnothing$, then as n increases,

$$\Omega_{\mathcal{M}_T:\mathcal{M}}^{CH} = \begin{cases} O_P\left(n^{\frac{\xi_{\mathcal{M}} p_{\mathcal{M}} - p_{\mathcal{M}_T} - a(1-\xi_{\mathcal{M}})}{2}}\right) & \text{if } b \text{ is fixed, and } s \text{ is fixed} \\[2mm] O_P\left(n^{\frac{p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}}\right) & \text{if } b = O(n), \text{ or } s = O(n) \end{cases}$$

In particular, if $\mathcal{M} \supset \mathcal{M}_T$, then $\Omega_{\mathcal{M}_T:\mathcal{M}}^{CH} = O_P\left(n^{\frac{p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}}\right)$ for all b and s.

2) If $\mathcal{M}_T = \mathcal{M}_\varnothing$, then as n increases,

$$\Omega_{\mathcal{M}_T:\mathcal{M}}^{CH} = \begin{cases} O_P(1) & \text{if } b \text{ is fixed, and } s \text{ is fixed} \\[2mm] O_P\left(n^{\frac{p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}}\right) & \text{if } b = O(n), \text{ or } s = O(n) \end{cases}$$

Proof. We first show Case 1) where $\mathcal{M}_T \neq \mathcal{M}_\varnothing$, by Lemma 4, $\xi_{\mathcal{M}_T} = 1$. We consider the following three scenarios about parameters b and s being fixed or $O(n)$.

Scenario 1: Both b, s are fixed. By Abramowitz and Stegun (1970) formula (13.1.5),

$${}_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(b-a)}(-s)^{-a}[1 + O(|s|^{-1})], \text{ when Real}(s) < 0. \tag{57}$$

Continuous mapping theorem suggests that for any model $\mathcal{M}$ whose $Q_{\mathcal{M}} = O_P(n^{\xi_{\mathcal{M}}})$,

$$\Omega^{CH}_{\mathcal{M}_T:\mathcal{M}} \approx \frac{\Gamma\left(\frac{a+p_{\mathcal{M}_T}}{2}\right)\left(\frac{s+Q_{\mathcal{M}_T}}{2}\right)^{-\frac{a+p_{\mathcal{M}_T}}{2}}}{\Gamma\left(\frac{a+p_{\mathcal{M}}}{2}\right)\left(\frac{s+Q_{\mathcal{M}}}{2}\right)^{-\frac{a+p_{\mathcal{M}}}{2}}} \propto \frac{(s+Q_{\mathcal{M}_T})^{-\frac{a+p_{\mathcal{M}_T}}{2}}}{(s+Q_{\mathcal{M}})^{-\frac{a+p_{\mathcal{M}}}{2}}} = O_P\left(n^{\frac{\xi_{\mathcal{M}}p_{\mathcal{M}}-p_{\mathcal{M}_T}-a(1-\xi_{\mathcal{M}})}{2}}\right).$$

(58)

Scenario 2: b is fixed, and $s = O(n)$. Since $s + Q_{\mathcal{M}_T} = O(n)$ and $s + Q_{\mathcal{M}} = O(n)$, then by (58), $\Omega^{CH}_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right).$

Scenario 3: $b = O(n)$. Lemma 4 indicates that $Q_{\mathcal{M}}$ is between $O_P(1)$ and $O_P(n)$. By Slater (1960) formula (4.3.3): if b is large, and a, s are bounded, then

$$_1F_1(a, b, s) = 1 + O(|b|^{-1}) \text{ is bounded;}$$

(59)

and by Slater (1960) formulas (4.3.7): if b is large, $s = by$, and a, y are bounded, then

$$_1F_1(a, b, s) = (1-y)^{-a}\left[1 - \frac{a(a+1)}{2b}\left(\frac{y}{1-y}\right)^2 + O(|b|^{-2})\right] \text{ is also bounded.}$$

(60)

Therefore, under the CH prior when parameter $b = O(n)$,

$$\Omega^{CH}_{\mathcal{M}_T:\mathcal{M}} = \frac{B\left(\frac{a+p_{\mathcal{M}_T}}{2}, \frac{b}{2}\right) {}_1F_1\left(\frac{a+p_{\mathcal{M}_T}}{2}, \frac{a+b+p_{\mathcal{M}_T}}{2}, -\frac{s+Q_{\mathcal{M}_T}}{2}\right)}{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right) {}_1F_1\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)} \xrightarrow{P} C \cdot \frac{B\left(\frac{a+p_{\mathcal{M}_T}}{2}, \frac{b}{2}\right)}{B\left(\frac{a+p_{\mathcal{M}}}{2}, \frac{b}{2}\right)}.$$

According to the Stirling's Formula $\Gamma(n) = e^{-n}n^{n-\frac{1}{2}}(2\pi)^{\frac{1}{2}}(1 + O(n^{-1}))$, the above ratio becomes $O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right).$

Next we examine Case 2) where $\mathcal{M}_T = \mathcal{M}_\emptyset$. In this case, Lemma 4 suggests that both $Q_{\mathcal{M}_T}$ and $Q_{\mathcal{M}}$ are on the same order $O_P(1)$. Hence in Scenario 1, where both b and s are fixed, $\Omega^{CH}_{\mathcal{M}_T:\mathcal{M}} = O_P(1)$; In Scenario 2, since both $s + Q_{\mathcal{M}_T}$ and $s + Q_{\mathcal{M}}$ are on the order of $O_P(n)$, the same deviation and result as in Case 1) Scenario 2 apply. In Scenario 3, both $s + Q_{\mathcal{M}_T}$ and $s + Q_{\mathcal{M}}$ are $O_P(1)$ if s is fixed, and $O_P(n)$ if $s = O(n)$, so the same derivation and result as in Case 1) Scenario 3 apply. $\square$

Lemma 6. *Under the robust prior, denote the term in* $\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}}$:

$$\Omega^R_{\mathcal{M}_T:\mathcal{M}} \triangleq \left(\frac{p_{\mathcal{M}}+1}{p_{\mathcal{M}_T}+1}\right)^{\frac{1}{2}} \cdot \frac{Q_{\mathcal{M}_T}^{-\frac{p_{\mathcal{M}_T}+1}{2}}}{Q_{\mathcal{M}}^{-\frac{p_{\mathcal{M}}+1}{2}}} \cdot \frac{\gamma\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+1)}\right)}{\gamma\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+1)}\right)}. \tag{61}$$

As the sample size n *increases,* $\Omega^R_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right)$.

Proof. By Abramowitz and Stegun (1970) formula (6.5.12), the incomplete Gamma function $\gamma(a,s) = \int_0^s t^{a-1}e^{-t}dt$ can be expressed using the $_1F_1$ function

$$\gamma(a,s) = {_1F_1}(a, a+1, -s)\frac{s^a}{a}. \tag{62}$$

Therefore, (61) becomes

$$\left(\frac{p_{\mathcal{M}}+1}{p_{\mathcal{M}_T}+1}\right)^{\frac{1}{2}} \cdot \frac{Q_{\mathcal{M}_T}^{-\frac{p_{\mathcal{M}_T}+1}{2}}}{Q_{\mathcal{M}}^{-\frac{p_{\mathcal{M}}+1}{2}}} \cdot \frac{\left(\frac{p_{\mathcal{M}_T}+1}{2}\right)^{-1}\left(\frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+1)}\right)^{\frac{p_{\mathcal{M}_T}+1}{2}} {_1F_1}\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{p_{\mathcal{M}_T}+3}{2}, -\frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+1)}\right)}{\left(\frac{p_{\mathcal{M}}+1}{2}\right)^{-1}\left(\frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+1)}\right)^{\frac{p_{\mathcal{M}}+1}{2}} {_1F_1}\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{p_{\mathcal{M}}+3}{2}, -\frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+1)}\right)}$$

Since $_1F_1(a,b,0) = 1$, and both $Q_{\mathcal{M}_T}/n, Q_{\mathcal{M}}/n$ are bounded, the ratio between the $_1F_1$ functions is bounded as n increases. Therefore we further simplify $\Omega^R_{\mathcal{M}_T:\mathcal{M}} \propto (n+1)^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right)$. This result holds no matter whether $\mathcal{M}_T = \mathcal{M}_\emptyset$ or not. $\square$

Lemma 7. *Under the intrinsic prior, denote the term in* $\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}}$:

$$\Omega^I_{\mathcal{M}_T:\mathcal{M}} \triangleq \frac{\left(\frac{n+p_{\mathcal{M}}+1}{p_{\mathcal{M}}+1}\right)^{\frac{p_{\mathcal{M}}}{2}} e^{\frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+p_{\mathcal{M}}+1)}} B\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{1}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{p_{\mathcal{M}_T}+2}{2}, \frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)}, -\frac{p_{\mathcal{M}_T}+1}{n}\right)}{\left(\frac{n+p_{\mathcal{M}_T}+1}{p_{\mathcal{M}_T}+1}\right)^{\frac{p_{\mathcal{M}_T}}{2}} e^{\frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)}} B\left(\frac{p_{\mathcal{M}}+1}{2}, \frac{1}{2}\right) \Phi_1\left(\frac{1}{2}, 1, \frac{p_{\mathcal{M}}+2}{2}, \frac{Q_{\mathcal{M}}(p_{\mathcal{M}}+1)}{2(n+p_{\mathcal{M}}+1)}, -\frac{p_{\mathcal{M}}+1}{n}\right)}$$

As the sample size n *increases,* $\Omega^I_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right)$.

Proof. Since $p_{\mathcal{M}_T}, p_{\mathcal{M}}$ are bounded, and $Q_{\mathcal{M}_T}/n, Q_{\mathcal{M}}/n$ are bounded in probability, as $n \to \infty$,

$$\Omega^{\mathrm{I}}_{\mathcal{M}_T:\mathcal{M}} \xrightarrow{\mathrm{P}} C \cdot \frac{\left(\frac{n+p_{\mathcal{M}}+1}{p_{\mathcal{M}}+1}\right)^{\frac{p_{\mathcal{M}}}{2}}}{\left(\frac{n+p_{\mathcal{M}_T}+1}{p_{\mathcal{M}_T}+1}\right)^{\frac{p_{\mathcal{M}_T}}{2}}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right).$$

$\square$

Lemma 8. *Under the local EB, denote the term in* $\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}}$*:*

$$\Omega^{LEB}_{\mathcal{M}_T:\mathcal{M}} \overset{\triangle}{=} \frac{\max\left\{\exp\left(-\frac{Q_{\mathcal{M}_T}}{2}\right), \left(\frac{Q_{\mathcal{M}_T}}{p_{\mathcal{M}_T}}\right)^{-\frac{p_{\mathcal{M}_T}}{2}} \exp\left(-\frac{p_{\mathcal{M}_T}}{2}\right)\right\}}{\max\left\{\exp\left(-\frac{Q_{\mathcal{M}}}{2}\right), \left(\frac{Q_{\mathcal{M}}}{p_{\mathcal{M}}}\right)^{-\frac{p_{\mathcal{M}}}{2}} \exp\left(-\frac{p_{\mathcal{M}}}{2}\right)\right\}}. \tag{63}$$

1) If $\mathcal{M}_T \neq \mathcal{M}_\emptyset$*, then as* n *increases,* $\Omega^{LEB}_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{\frac{\xi_{\mathcal{M}} p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}}\right)$*. In particular, if* $\mathcal{M} \supset \mathcal{M}_T$*, then* $\Omega^{LEB}_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right)$*.*

2) If $\mathcal{M}_T = \mathcal{M}_\emptyset$*, then as* n *increases,* $\Omega^{LEB}_{\mathcal{M}_T:\mathcal{M}} = O_P(1)$*.*

Proof. Case 2) is straightforward, because when $\mathcal{M}_T = \mathcal{M}_\emptyset$, $Q_{\mathcal{M}_T} = O_P(1)$ and $Q_{\mathcal{M}} = O_P(1)$. Now let us focus on Case 1). In (63), the numerator equals $\exp(-Q_{\mathcal{M}_T}/2)$ if and only if $Q_{\mathcal{M}_T} \leq p_{\mathcal{M}_T}$, and the denominator follows the same rule when we replacing $\mathcal{M}_T$ with $\mathcal{M}$. Since $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, $Q_{\mathcal{M}} = O_P(n)$ is greater than $p_{\mathcal{M}}$ for large n. Hence the numerator of (63) is proportional to $(Q_{\mathcal{M}_T}/p_{\mathcal{M}_T})^{-\frac{p_{\mathcal{M}_T}}{2}} \exp(-p_{\mathcal{M}_T}/2) = O_P(n^{-\frac{p_{\mathcal{M}_T}}{2}})$. For model $\mathcal{M}$ whose $Q_{\mathcal{M}} = O_P(n^{\xi_{\mathcal{M}}})$, if $\xi_{\mathcal{M}} > 0$, then when n is large enough, $Q_{\mathcal{M}} > p_{\mathcal{M}}$, so the denominator is $O_P(n^{-\frac{\xi_{\mathcal{M}} p_{\mathcal{M}}}{2}})$. If $\xi_{\mathcal{M}} = 0$, then the denominator is $O_P(1)$, which can also be written as $O_P(n^{-\frac{\xi_{\mathcal{M}} p_{\mathcal{M}}}{2}})$. $\square$

We now examine the model selection consistency.

Proof of Theorem 1

Proof. By Lemma 1, $\mathcal{J}_n(\hat{\alpha}_\mathcal{M}) = O_P(n^{\tau_\mathcal{M}})$, where $0 \leq \tau_\mathcal{M} \leq 1$, and $\tau_\mathcal{M} = 1$ if $\mathcal{M} \supset \mathcal{M}_T$. Hence,

$$\left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_T})}{\mathcal{J}_n(\hat{\alpha}_\mathcal{M})} \right]^{-\frac{1}{2}} = O_P\left(n^{-\frac{1-\tau_\mathcal{M}}{2}} \right).$$

For the CH prior,

$$\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}} = \left[\frac{\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_T})}{\mathcal{J}_n(\hat{\alpha}_\mathcal{M})} \right]^{-\frac{1}{2}} \cdot \Lambda_{\mathcal{M}_T:\mathcal{M}} \cdot \Omega_{\mathcal{M}_T:\mathcal{M}}^{\text{CH}} \cdot [1 + O_P(1/n)]. \tag{64}$$

We first consider the case where both b and s are fixed, by using the results in Lemma 3 and 5. In the case where $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, for any non-true model $\mathcal{M} \supset \mathcal{M}_T$, then $p_\mathcal{M} > p_{\mathcal{M}_T}$, $\tau_\mathcal{M} = 1$, and $\xi_\mathcal{M} = 1$, hence

$$\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}} = O_P(1) \cdot O_P(1) \cdot O_P\left(n^{\frac{p_\mathcal{M} - p_{\mathcal{M}_T}}{2}} \right) \cdot [1 + O_P(1/n)] \xrightarrow{\text{P}} \infty.$$

On the other hand, if $\mathcal{M} \not\supset \mathcal{M}_T$, then

$$\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}} = O_P\left(n^{-\frac{1-\tau_\mathcal{M}}{2}} \right) \cdot O_P\left(e^{c_\mathcal{M} n} \right) \cdot O_P\left(n^{\frac{\xi_\mathcal{M} p_\mathcal{M} - p_{\mathcal{M}_T} - a(1-\xi_\mathcal{M})}{2}} \right) \cdot [1 + O_P(1/n)] \xrightarrow{\text{P}} \infty.$$

In contrast, if $\mathcal{M}_T = \mathcal{M}_\emptyset$, then for any model $\mathcal{M}$, since $\mathcal{M} \supset \mathcal{M}_T$, $\tau_\mathcal{M} = 1$. So the Bayes factor

$$\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}} = O_P(1) \cdot O_P(1) \cdot O_P(1) \cdot [1 + O_P(1/n)]$$

is bounded, which suggests the selection consistency does not hold when $\mathcal{M}_T = \mathcal{M}_\emptyset$.

Next consider the case where $b = O(n)$ or $s = O(n)$. For any model $\mathcal{M} \not\supset \mathcal{M}_T$, the proof is similar as above. If $\mathcal{M} \supset \mathcal{M}_T$, then $\tau_\mathcal{M} = 1$ and $p_\mathcal{M} > p_{\mathcal{M}_T}$, so

$$\mathsf{BF}_{\mathcal{M}_T:\mathcal{M}} = O_P(1) \cdot O_P(1) \cdot O_P\left(n^{\frac{p_\mathcal{M} - p_{\mathcal{M}_T}}{2}} \right) \cdot [1 + O_P(1/n)] \xrightarrow{\text{P}} \infty,$$

which holds even when $\mathcal{M}_T = \mathcal{M}_\emptyset$.

For the robust prior, the intrinsic prior, and local EB, their Bayes factor are given by (64), with $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}}$ replaced by $\Omega^{\text{R}}_{\mathcal{M}_T:\mathcal{M}}$, $\Omega^{\text{I}}_{\mathcal{M}_T:\mathcal{M}}$, and $\Omega^{\text{LEB}}_{\mathcal{M}_T:\mathcal{M}}$, respectively. By Lemma 6, 7, and 8, the proofs are similar to the CH prior, hence omitted. $\qquad\square$

A.11 Proof to Proposition 6

Proof. If $b = O(n)$ then by (59) or (60),

$$\mathbb{E}(1/g) = \frac{B\left(\frac{a}{2}+1,\frac{b}{2}-1\right) {}_1F_1\left(\frac{a}{2}+1,\frac{a+b}{2},-\frac{s}{2}\right)}{B\left(\frac{a}{2},\frac{b}{2}\right) {}_1F_1\left(\frac{a}{2},\frac{a+b}{2},-\frac{s}{2}\right)} \tag{65}$$

$$\propto \frac{B\left(\frac{a}{2}+1,\frac{b}{2}-1\right)}{B\left(\frac{a}{2},\frac{b}{2}\right)} \longrightarrow \frac{a}{b-2} = O(1/n).$$

If b is fixed and $s = O(n)$, then by (57) and (65),

$$\mathbb{E}(1/g) \approx \frac{B\left(\frac{a}{2}+1,\frac{b}{2}-1\right)\Gamma\left(\frac{b}{2}\right)\left(\frac{s}{2}\right)^{\frac{a}{2}}}{B\left(\frac{a}{2},\frac{b}{2}\right)\Gamma\left(\frac{b}{2}-1\right)\left(\frac{s}{2}\right)^{\frac{a}{2}+1}} \propto \frac{1}{s} = O(1/n).$$

$\qquad\square$

A.12 Proof of Proposition 7

Proof. For the CH prior, according to (32), the conditional posterior of $z = 1 - u$ is

$$z \mid \mathbf{Y}, \mathcal{M} \xrightarrow{D} \text{CH}\left(\frac{b}{2}, \frac{a+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right), \tag{66}$$

and its characteristic function is

$$\phi_z(t) = \mathbb{E}\left(e^{itz}\right) = \int \frac{z^{\frac{b}{2}-1}(1-z)^{\frac{a+p_{\mathcal{M}_T}}{2}-1}e^{\left(\frac{s+Q_{\mathcal{M}_T}}{2}+it\right)z}}{B(\frac{b}{2},\frac{a+p_{\mathcal{M}_T}}{2}) {}_1F_1(\frac{b}{2},\frac{a+b+p_{\mathcal{M}_T}}{2},\frac{s+Q_{\mathcal{M}_T}}{2})}dz = \frac{{}_1F_1(\frac{b}{2},\frac{a+b+p_{\mathcal{M}_T}}{2},\frac{s+Q_{\mathcal{M}_T}}{2}+it)}{{}_1F_1(\frac{b}{2},\frac{a+b+p_{\mathcal{M}_T}}{2},\frac{s+Q_{\mathcal{M}_T}}{2})}$$

Lemma 4 shows that if $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, then $s + Q_{\mathcal{M}_T} = O_P(n)$. If $b = O(1)$, then by (51) and the continuous mapping theorem, for any $t \in \mathbb{R}$, as n goes in to infinity,

$$\phi_z(t) \longrightarrow \frac{\exp(\frac{s+Q_{\mathcal{M}_T}}{2} + it) \cdot (\frac{s+Q_{\mathcal{M}_T}}{2} + it)^{-\frac{a+p_{\mathcal{M}_T}}{2}}}{\exp(\frac{s+Q_{\mathcal{M}_T}}{2}) \cdot (\frac{s+Q_{\mathcal{M}_T}}{2})^{-\frac{a+p_{\mathcal{M}_T}}{2}}} \xrightarrow{P} \exp(it).$$

If $b = O(n)$, then using formula (60), we can obtain the same limit.

For the robust prior, we examine the characteristic function of $u = 1 - z$. Based on (37),

$$\phi_u(t) = \mathbb{E}\left(e^{itu}\right) = \frac{\displaystyle\int_0^{\frac{p_{\mathcal{M}_T}+1}{n+1}} u^{\frac{p_{\mathcal{M}_T}+1}{2}-1} e^{\left(it - \frac{Q_{\mathcal{M}_T}}{2}\right)u} \, du}{\displaystyle\int_0^{\frac{p_{\mathcal{M}_T}+1}{n+1}} u^{\frac{p_{\mathcal{M}_T}+1}{2}-1} e^{-\frac{Q_{\mathcal{M}_T}u}{2}} \, du}$$

$$= \frac{\gamma\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{(Q_{\mathcal{M}_T}-2it)(p_{\mathcal{M}_T}+1)}{2(n+1)}\right)}{\gamma\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+1)}\right)} \cdot \left(\frac{Q_{\mathcal{M}_T} - 2it}{Q_{\mathcal{M}_T}}\right)^{-\frac{p_{\mathcal{M}_T}+1}{2}}.$$

Since $Q_{\mathcal{M}_T} = O_P(n)$, for any fixed $t \in \mathbb{R}$, the ratio of the incomplete Gamma functions goes to 1, and so does the second fraction. Therefore, $\phi_u(t) \xrightarrow{P} 1$, which is the characteristic function of the degenerate distribution at 0.

For the intrinsic prior, by (30) and Table 1, the conditional posterior of u is

$$u \mid \mathbf{Y}, \mathcal{M}_T \sim \text{tCCH}\left(\frac{p_{\mathcal{M}_T}+1}{2}, \frac{1}{2}, 1, \frac{Q_{\mathcal{M}_T}}{2}, \frac{n+p_{\mathcal{M}_T}+1}{p_{\mathcal{M}_T}+1}, \frac{n+p_{\mathcal{M}_T}+1}{n}\right), \tag{67}$$

and hence its characteristic function for any $t \in \mathbb{R}$ is

$$\phi_u(t) = \exp\left\{\frac{it(p_{\mathcal{M}_T}+1)}{n+p_{\mathcal{M}_T}+1}\right\} \frac{\Phi_1\left(\frac{1}{2}, 1, \frac{p_{\mathcal{M}_T}+2}{2}, \frac{(Q_{\mathcal{M}_T}-2it)(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)}, -\frac{p_{\mathcal{M}_T}+1}{n}\right)}{\Phi_1\left(\frac{1}{2}, 1, \frac{p_{\mathcal{M}_T}+2}{2}, \frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)}, -\frac{p_{\mathcal{M}_T}+1}{n}\right)}. \tag{68}$$

Since $Q_{\mathcal{M}_T} = O_P(n)$ and

$$\frac{(Q_{\mathcal{M}_T} - 2it)(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)} - \frac{Q_{\mathcal{M}_T}(p_{\mathcal{M}_T}+1)}{2(n+p_{\mathcal{M}_T}+1)} \xrightarrow{P} 0,$$

73

by continuous mapping theorem, the ratio of the two Φ_1 functions in (68) converges to one in probability. Therefore, under the intrinsic prior, $\phi_u(t) \xrightarrow{P} 1$. $\qquad\square$

A.13 Proof of Theorem 2

Proof. For the CH prior, we will prove the BMA estimation consistency in two steps: 1) $\mathcal{M}_T \neq \mathcal{M}_\varnothing$ and 2) $\mathcal{M}_T = \mathcal{M}_\varnothing$. When $\mathcal{M}_T \neq \mathcal{M}_\varnothing$, the model selection consistency always holds, so we just need to show the estimation consistency under the true model $\mathcal{M}_T$. For notation simplicity, we denote $\boldsymbol{\Sigma}_{n,\mathcal{M}} = \mathcal{J}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}$. According to (17) and (66), the characteristic function of the posterior distribution $p(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathcal{M}_T, \mathbf{Y})$ is

$$
\begin{aligned}
\phi_{\boldsymbol{\beta}_{\mathcal{M}_T}}(\mathbf{t}) &= \int e^{i\mathbf{t}^T \boldsymbol{\beta}_{\mathcal{M}_T}} \, p(\boldsymbol{\beta}_{\mathcal{M}_T} | \mathcal{M}_T, \mathbf{Y}) \, d\boldsymbol{\beta}_{\mathcal{M}_T} \\
&= \int e^{i\mathbf{t}^T \boldsymbol{\beta}_{\mathcal{M}_T}} \left\{ \int p(\boldsymbol{\beta}_{\mathcal{M}_T} | z, \mathcal{M}_T, \mathbf{Y}) \, p(z | \mathcal{M}_T, \mathbf{Y}) dz \right\} d\boldsymbol{\beta}_{\mathcal{M}_T} \\
&= \int \left\{ \int e^{i\mathbf{t}^T \boldsymbol{\beta}_{\mathcal{M}_T}} \, p(\boldsymbol{\beta}_{\mathcal{M}_T} | z, \mathcal{M}_T, \mathbf{Y}) \, d\boldsymbol{\beta}_{\mathcal{M}_T} \right\} p(z | \mathcal{M}_T, \mathbf{Y}) dz \\
&= \int e^{z(i\mathbf{t}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}_T} - \frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma}_{n,\mathcal{M}_T} \mathbf{t})} \, p(z | \mathcal{M}_T, \mathbf{Y}) dz
\end{aligned}
$$

In the above calculation, the integrand $e^{i\mathbf{t}^T \boldsymbol{\beta}_{\mathcal{M}_T}}$ has a bounded modulus, so according to Fubini's Theorem, the two integrals (with respect to z and $\boldsymbol{\beta}_{\mathcal{M}_T}$) can be interchanged. Since $Q_{\mathcal{M}_T} = O_P(n)$ and $\boldsymbol{\Sigma}_{n,\mathcal{M}_T} = O_P(n^{-1})$, using methods similar to the proof of Proposition 7 and asymptotic normality of MLE, we can show that for any vector $\mathbf{t}$,

$$
\phi_{\boldsymbol{\beta}_{\mathcal{M}_T}}(\mathbf{t}) \longrightarrow e^{i\mathbf{t}^T \hat{\boldsymbol{\beta}}_{\mathcal{M}_T} - \frac{1}{2}\mathbf{t}^T \boldsymbol{\Sigma}_{n,\mathcal{M}_T} \mathbf{t}} \xrightarrow{P} e^{i\mathbf{t}^T \boldsymbol{\beta}^*_{\mathcal{M}_T}}.
$$

On the other hand, when $\mathcal{M}_T = \mathcal{M}_\varnothing$, under the CH prior model selection consistency does not hold if both b and s are fixed. Hence we need to examine the limit of posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}}$ under all models. Under any model $\mathcal{M}$, the true model is nested in it, so the MLE of the coefficient $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ converges to the true parameters $\mathbf{0}$ in probability as n goes to infinity.

Since the modulus of $e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}}}$ is bounded by a constant 1, which is integrable if regarded as a function of z, so according to the dominated convergence theorem, the characteristic function of the posterior distribution $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M})$ evaluated at any vector $\mathbf{t} \in \mathbb{R}^p$ is

$$\phi_{\boldsymbol{\beta}_{\mathcal{M}}}(\mathbf{t}) = \int e^{z(i\mathbf{t}^T\hat{\beta}_{\mathcal{M}} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_{n,\mathcal{M}}\mathbf{t})} \, p(z \mid \mathcal{M}, \mathbf{Y})dz$$
$$\xrightarrow{P} \int \left[e^{z(i\mathbf{t}^T\mathbf{0} - \frac{1}{2}\mathbf{t}^T\mathbf{0}\mathbf{t})} \right] p(z \mid \mathcal{M}, \mathbf{Y})dz = 1.$$

For the robust and intrinsic priors, model selection consistency always holds. So we just need to consider under $\mathcal{M}_T$. Based on (37) and (67), proofs similar to the above proof of the CH prior can show that either $\mathcal{M}_T \neq \mathcal{M}_\emptyset$ or $\mathcal{M}_T = \mathcal{M}_\emptyset$, the characteristic function of $p(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathcal{M}_T, \mathbf{Y})$ converges to $e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}_T}^*}$ or 1 in probability, respectively. $\qquad\square$

B   Test-Based Bayes Factors

B.1   Test-Based Bayes Factor under the g-Prior

In Bayesian hypothesis testing, while the traditional Bayes factor computes the ratio between marginal likelihoods of data (referred to as data-based BF, or DBF in short), another type of Bayes factor, defined as the ratio between marginal likelihoods of a test statistic, has also been introduced (Johnson 2005, 2008). In particular, based on the likelihood ratio statistic, the test-based Bayes factor (TBF) has been applied in model selection under the g-prior (Hu and Johnson 2009; Held et al. 2015, 2016), where models with high TBFs are preferable.

To compute the TBF based on the likelihood ratio deviance $z_{\mathcal{M}}$ (22), first, asymptotic theory (Davidson and Lever 1970) suggests that the limit distribution of $z_{\mathcal{M}}$ under the null model $\mathcal{M}_\emptyset$ and under a local alternative model $\mathcal{M}$ are central and non-central Chi-squares, respectively,

$$z_{\mathcal{M}} \mid \mathcal{M}_\emptyset \sim \chi^2_{p_{\mathcal{M}}}, \quad z_{\mathcal{M}} \mid \mathcal{M} \sim \chi^2_{p_{\mathcal{M}}}(\lambda_{\mathcal{M}}), \text{ where } \lambda_{\mathcal{M}} = \boldsymbol{\beta}_{\mathcal{M}}^T \mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}} = \mathbf{0})\boldsymbol{\beta}_{\mathcal{M}}.$$

Then, as $p(z_\mathcal{M} \mid \mathcal{M}, \boldsymbol{\beta}_\mathcal{M})$ depends on $\boldsymbol{\beta}_\mathcal{M}$ through the non-centrality parameter $\lambda_\mathcal{M}$, integrating $\boldsymbol{\beta}_\mathcal{M}$ out under its prior density yields the marginal likelihood $p(z_\mathcal{M} \mid \mathcal{M})$. Last, the TBF is defined as the ratio

$$\text{TBF}_{\mathcal{M}:\mathcal{M}_\emptyset} = \frac{p(z_\mathcal{M} \mid \mathcal{M})}{p(z_\mathcal{M} \mid \mathcal{M}_\emptyset)} = \frac{\int p(z_\mathcal{M} \mid \boldsymbol{\beta}_\mathcal{M}, \mathcal{M}) p(\boldsymbol{\beta}_\mathcal{M} \mid \mathcal{M}) d\boldsymbol{\beta}_\mathcal{M}}{p(z_\mathcal{M} \mid \mathcal{M}_\emptyset)}. \tag{69}$$

To conduct model selection in GLMs, Held et al. (2015) derive the TBF under the g-prior (8), in whose density, $\boldsymbol{\beta}_\mathcal{M}$ appears in the format of $\lambda_\mathcal{M}$. Thus the conjugacy permits a tractable marginal likelihood $p(z_\mathcal{M} \mid \mathcal{M})$ as a Gamma distribution. Therefore, the resulting TBF has a closed form expression as in (23).

B.2 Comparing Data-Based and Test-Based Bayes Factors

The TBF (23) has a similar expression to the DBF (21). In fact, the two Bayes factors would be the same if $z_\mathcal{M} = Q_\mathcal{M}$ and $\mathcal{J}_n(\hat{\alpha}_{\mathcal{M}_\emptyset}) = \mathcal{J}_n(\hat{\alpha}_\mathcal{M})$. Naturally, it is interesting to examine how different the two Bayes factors are.

We compare DBF (21) and TBF (23) empirically through a logistic regression toy example, with $g = n$ and a single covariate generated from independent standard normal distributions. With the intercept set to $\alpha = 0.5$, three scenarios are studied with different coefficients $\beta = 0, 20/\sqrt{n}, 2$, which correspond to the null, local alternative, and alternative, respectively. To study asymptotics, various sample sizes $n = 100, 500, 1000, 5000$ are taken. For each combination of β and n, 100 independent datasets are generated. To obtain an accurate approximation to the DBF, in addition to the integrated Laplace approximation (ILA) formula (21), we also implement importance sampling (IS), which can be viewed as a gold standard if the number of samples drawn is large. Here we draw $m = 10000$ samples $\alpha^{(t)}, \beta^{(t)}$, independently from Student-t distributions with degrees of freedom 4, with location and scale parameters matching those in the corresponding conditional posteriors (17), (18).

Figure 3 shows that when the null or the local alternative is true, TBF (23) is asymp-
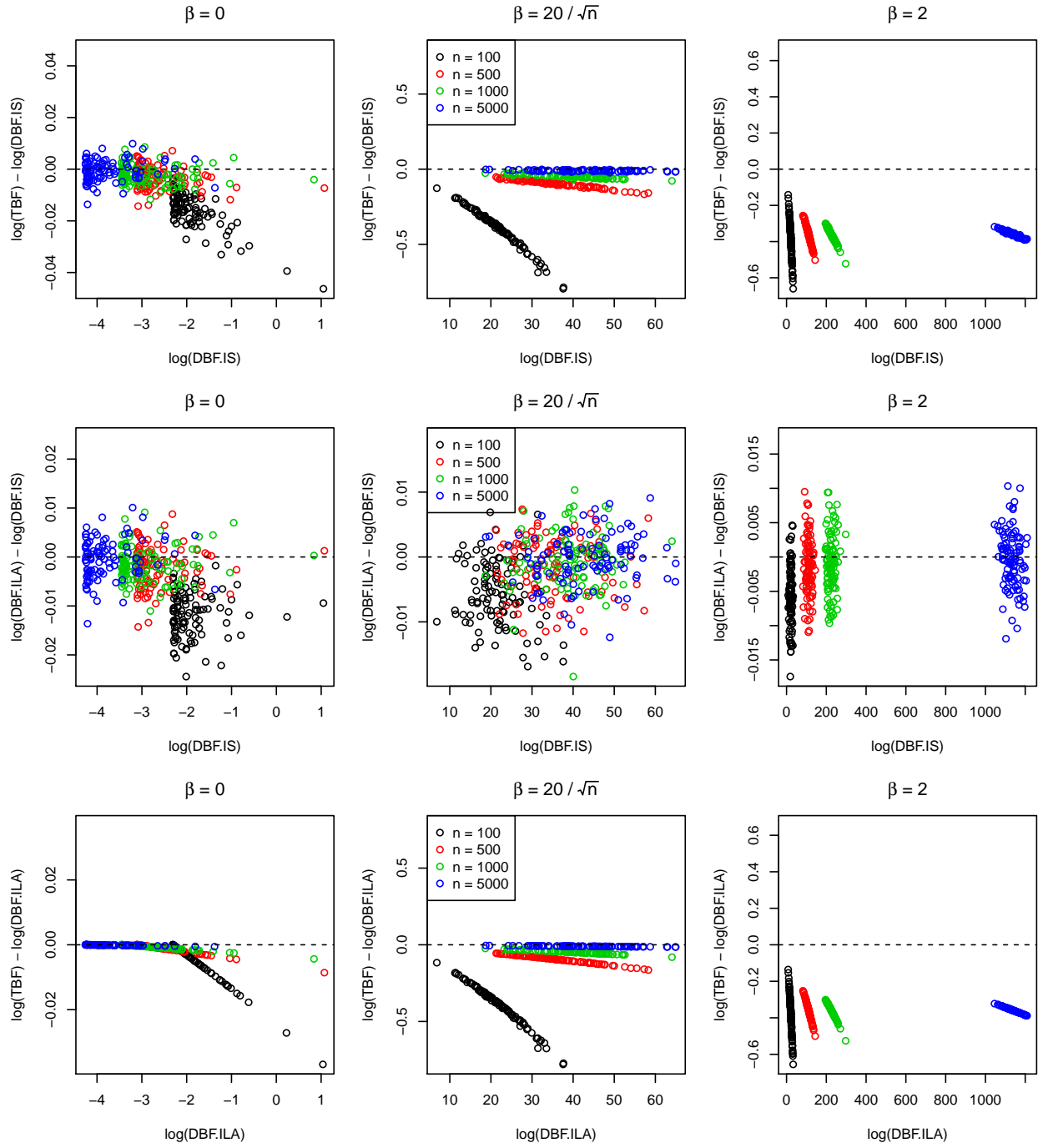
Figure 3: From top to bottom: TBF versus DBF approximated by IS, DBF approximated by ILA vs DBF approximated by IS, and TBF versus DBF approximated by ILA. From left to right: the null, local alternative, and alternative hypotheses.
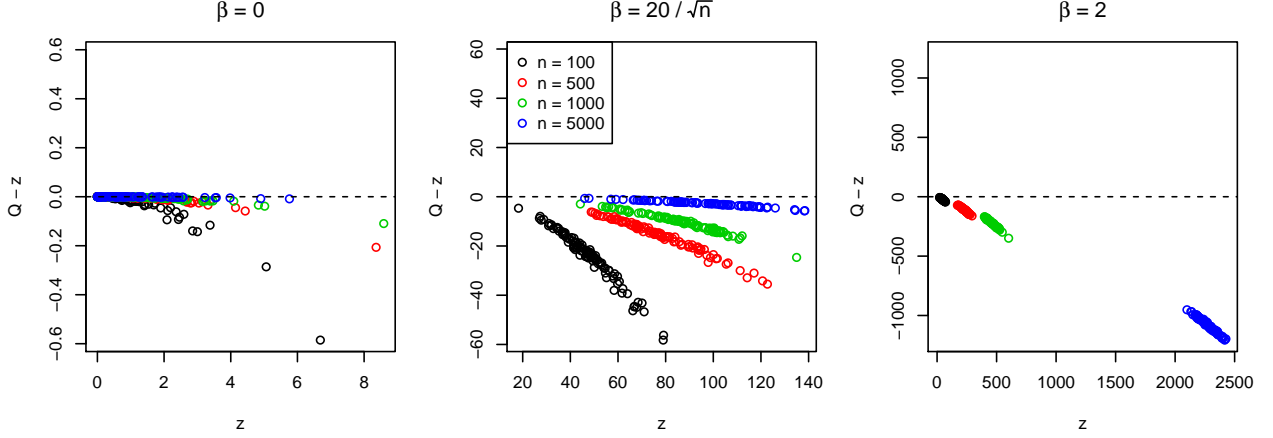
Figure 4: Wald statistic $Q_{\mathcal{M}}$ versus the deviance $z_{\mathcal{M}}$.

totically the same as the DBF computed under either IS or ILA (21). In contrast, when the alternative is true, TBF differs from DBF by a relatively small but systematic amount. Comparison between the Wald statistic $Q_{\mathcal{M}}$ (20) and the deviance $z_{\mathcal{M}}$ (22) suggests a similar phenomenon (Figure 4). They are asymptotically the same under the null or local alternative, but different under the alternative.

In addition to the similarity between the two Bayes factors under g-priors, we notice that as a function of g, the test-based marginal likelihood would have the same kernel $p(z_{\mathcal{M}} \mid \mathcal{M}) \propto (1 + g)^{-p_{\mathcal{M}}/2} \exp\left(-z_{\mathcal{M}}/[2(1 + g)]\right)$ as its data-based counterpart (19) if $z_{\mathcal{M}} = Q_{\mathcal{M}}$. Therefore, all empirical Bayes and fully Bayes approaches on g, discussed in Section 2.6 and Section 3, can be readily applied to test-based methods with minimal changes. Held et al. (2015) apply local empirical Bayes, $p(z_{\mathcal{M}} \mid \mathcal{M}) = \max_{g \geq 0} p(z_{\mathcal{M}} \mid g, \mathcal{M})$, and fully Bayes, $p(z_{\mathcal{M}} \mid \mathcal{M}) = \int p(z_{\mathcal{M}} \mid g, \mathcal{M})p(g)dg$ to compute marginal likelihoods for TBFs. However, we find that these optimized and integrated versions of TBF may no longer be coherent, in the sense that results change with the choice of the baseline model. Elaborating, when testing nested models $\mathcal{M}_1 \subset \mathcal{M}_2$,

$$\text{TBF}_{\mathcal{M}_2:\mathcal{M}_1} \neq \frac{\text{TBF}_{\mathcal{M}_2:\mathcal{M}_\varnothing}}{\text{TBF}_{\mathcal{M}_1:\mathcal{M}_\varnothing}},$$

78

if one computes the left hand side TBF under baseline $\mathcal{M}_1$, but computes the right hand side TBFs under baseline $\mathcal{M}_\emptyset$. The main reason for this incoherence is that for model $\mathcal{M}$, unlike the data-based marginal likelihood which only depends on $\mathcal{M}$ itself, the test statistic $z_{\mathcal{M}}$ also depends on the baseline model. On the other hand, coherence exists for the TBF (23) under fixed g, since $z_{\mathcal{M}_2:\mathcal{M}_1} = z_{\mathcal{M}_2:\mathcal{M}_\emptyset} - z_{\mathcal{M}_1:\mathcal{M}_\emptyset}$ (Johnson 2008). Hence, change of baseline models does not affect the results of the TBF under fixed g, which is also the case with the DBF.

C   Additional Simulation Examples

We first include some additional results from the logistic regression simulation example that are examined in Section 5.1 (see Table 6) and then introduce a different simulation study on Poisson regressions.

The simulation setup of the Poisson regression example is similar to that of the logistic regression in Section (5.1). True values of coefficients (including the intercept) are set to one-fifth of those in the logistic regression, to avoid occasional extremely large values in $\mathbf{Y}$. Tables 7-9 display model selection and parameter estimation performance. Comparison among priors on $\boldsymbol{\beta}_{\mathcal{M}}$ leads to similar conclusions to the logistic regression example. For the Poisson regression, overall model selection accuracy is not as high as the logistic regression when $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, which is likely due to the smaller magnitude of coefficients.

Table 6: Logistic regression simulation example: average size of selected models, out of 100 realizations.

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$(a=1/2, b=n)$	0	0	5	4	10	8	17	13	17	15	5	3
CH$(a=1, b=n)$	0	0	5	5	10	8	17	13	18	15	5	3
CH$(a=1/2, b=n/2)$	0	0	6	5	10	9	17	14	25	20	5	3
CH$(a=1, b=n/2)$	0	0	6	5	10	9	17	14	26	22	5	3
Beta-prime	0	0	5	4	10	8	17	13	19	15	5	3
ZS adapted	0	0	5	5	10	8	17	13	18	15	5	3
Benchmark	0	0	6	6	11	10	18	15	27	24	5	3
Robust	0	0	6	5	11	9	18	14	34	30	21	10
Intrinsic	0	0	6	5	11	9	18	14	32	30	14	5
Hyper-g/n	0	1	6	5	11	10	18	15	69	56	99	80
DBF, $g=n$	0	0	5	4	9	7	15	11	7	5	5	3
TBF, $g=n$	0	0	5	4	9	7	15	11	7	5	5	3
Jeffreys	3	3	6	6	11	10	18	15	70	60	99	91
Hyper-g	4	4	6	6	11	10	18	15	70	61	100	93
Uniform	4	4	7	6	12	10	18	15	70	61	100	97
Local EB	19	19	6	6	11	10	18	15	71	60	100	96
AIC	3	3	8	7	12	11	18	15	34	34	6	4
BIC	0	0	5	4	9	7	15	11	7	5	5	3

Table 7: Poisson regression simulation example: number of times the true model are 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$(a=1/2, b=n)$	94	92	10	**2**	10	0	0	0	2	0	1	0
CH$(a=1, b=n)$	87	89	10	**2**	10	0	0	0	11	1	1	0
CH$(a=1/2, b=n/2)$	91	89	**11**	**2**	10	0	0	0	3	0	1	0
CH$(a=1, b=n/2)$	82	85	**11**	**2**	9	0	0	0	5	**2**	2	0
Beta-prime	94	92	10	**2**	10	0	0	0	7	0	1	0
ZS adapted	87	89	10	**2**	11	0	0	0	6	0	1	0
Benchmark	**97**	**93**	7	0	12	**1**	0	0	4	0	1	0
Robust	91	89	9	**2**	11	0	0	0	1	0	3	0
Intrinsic	85	88	8	**2**	12	**1**	0	0	1	0	3	0
Hyper-g/n	84	87	9	0	12	**1**	0	0	1	0	3	0
DBF, $g=n$	84	88	7	0	8	0	0	0	11	0	1	0
TBF, $g=n$	84	88	7	0	8	0	0	0	**14**	0	1	0
Jeffreys	0	0	7	1	12	**1**	0	0	0	0	3	0
Hyper-g	6	7	7	0	**13**	**1**	0	0	0	0	3	0
Uniform	4	2	7	0	**13**	**1**	0	0	1	1	3	0
Local EB	0	0	7	0	**13**	**1**	0	0	0	0	3	0
AIC	4	4	3	0	6	**1**	**1**	0	0	0	**8**	0
BIC	84	88	7	0	8	0	0	0	13	1	1	0

Table 8: Poisson regression simulation example: average size of selected models, out of 100 realizations.

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$(a=1/2, b=n)$	0	0	4	3	9	5	13	7	12	7	3	2
CH$(a=1, b=n)$	0	0	4	3	9	5	13	7	13	8	3	2
CH$(a=1/2, b=n/2)$	0	0	5	3	9	6	13	8	16	10	3	2
CH$(a=1, b=n/2)$	0	0	5	3	9	6	13	8	17	10	3	2
Beta-prime	0	0	4	3	9	5	13	7	13	7	3	2
ZS adapted	0	0	4	3	9	5	13	7	13	7	3	2
Benchmark	0	0	5	4	10	7	14	9	17	7	3	1
Robust	0	0	5	3	9	6	14	8	20	14	3	2
Intrinsic	0	0	5	3	10	6	14	8	22	13	3	2
Hyper-g/n	0	0	5	4	9	7	14	9	24	31	3	4
DBF, $g=n$	0	0	4	2	8	5	12	6	5	3	3	2
TBF, $g=n$	0	0	4	2	8	5	12	6	6	4	3	2
Jeffreys	2	3	5	4	10	7	14	10	29	36	3	18
Hyper-g	3	4	5	5	10	7	15	10	30	37	3	24
Uniform	4	4	6	5	10	7	15	10	30	38	3	34
Local EB	19	19	5	5	10	7	15	10	32	74	3	76
AIC	3	3	7	6	11	8	16	11	30	28	4	2
BIC	0	0	4	2	8	5	12	6	5	3	3	2

Table 9: Poisson regression simulation example: 1000 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$(a=1/2, b=n)$	5	8	**24**	**61**	34	120	58	198	66	132	37	103
CH$(a=1, b=n)$	6	9	**24**	**61**	34	120	58	197	66	134	37	98
CH$(a=1/2, b=n/2)$	7	11	**24**	**61**	**33**	116	56	188	75	148	36	97
CH$(a=1, b=n/2)$	7	13	**24**	**61**	**33**	115	55	187	77	135	36	94
Beta-prime	5	8	**24**	**61**	34	120	58	197	66	132	37	103
ZS adapted	6	9	**24**	**61**	34	119	55	197	66	125	37	99
Benchmark	8	18	26	65	**33**	**108**	51	170	74	150	36	133
Robust	7	13	25	63	**33**	115	51	183	88	182	36	97
Intrinsic	8	14	25	63	**33**	115	51	182	90	183	35	94
Hyper-g/n	5	12	25	65	**33**	109	52	172	84	162	36	97
DBF, $g=n$	5	6	25	63	37	132	68	231	**40**	**83**	39	101
TBF, $g=n$	5	6	25	63	37	132	68	231	**40**	84	39	101
Jeffreys	4	9	26	67	**33**	**108**	51	169	87	165	35	97
Hyper-g	4	7	26	68	**33**	**108**	51	**168**	87	164	**34**	112
Uniform	3	7	26	70	**33**	**108**	51	**168**	87	164	**34**	121
Local EB	**2**	**4**	26	71	**33**	**108**	51	**168**	99	256	**34**	222
AIC	17	40	28	74	34	115	**46**	171	120	284	37	**79**
BIC	5	6	25	63	37	132	68	231	**40**	84	39	100