Bayesian Hierarchical Models for Model Choice

by

Yingbo Li

Department of Statistical Science
Duke University

Date: __________________________

Approved:

____
Merlise A. Clyde, Supervisor

____
James O. Berger

____
Edwin S. Iversen

____
Elizabeth R. Hauser

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

Abstract

Bayesian Hierarchical Models for Model Choice

by

Yingbo Li

Department of Statistical Science
Duke University

Date: ______________________________

Approved:

____
Merlise A. Clyde, Supervisor

____
James O. Berger

____
Edwin S. Iversen

____
Elizabeth R. Hauser

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2013

Abstract

With the development of modern data collection approaches, researchers may collect hundreds to millions of variables, yet may not need to utilize all explanatory variables available in predictive models. Hence, choosing models that consist of a subset of variables often becomes a crucial step. In linear regression, variable selection not only reduces model complexity, but also prevents over-fitting. From a Bayesian perspective, prior specification of model parameters plays an important role in model selection as well as parameter estimation, and often prevents over-fitting through shrinkage and model averaging.

We develop two novel hierarchical priors for selection and model averaging, for Generalized Linear Models (GLMs) and normal linear regression, respectively. They can be considered as "spike-and-slab" prior distributions or more appropriately "spike-and-bell" distributions. Under these priors we achieve dimension reduction, since their point masses at zero allow predictors to be excluded with positive posterior probability. In addition, these hierarchical priors have heavy tails to provide robustness when MLE's are far from zero.

Zellner's g-prior is widely used in linear models. It preserves correlation structure among predictors in its prior covariance, and yields closed-form marginal likelihoods which leads to huge computational savings by avoiding sampling in the parameter space. Mixtures of g-priors avoid fixing g in advance, and can resolve consistency problems that arise with fixed g. For GLMs, we show that the mixture of g-priors using a Compound Confluent Hypergeometric distribution unifies existing choices in the literature and maintains their good properties such as tractable (approximate) marginal likelihoods and asymptotic consistency for model selection and parameter estimation under specific values of the hyper parameters.

While the g-prior is invariant under rotation within a model, a potential problem with the g-prior is that it inherits the instability of ordinary least squares (OLS) estimates when predictors are highly correlated. We build a hierarchical prior based on scale mixtures of independent normals, which incorporates invariance under rotations within models like ridge regression and the g-prior, but has heavy tails like the Zeller-Siow Cauchy prior. We find this method out-performs the gold standard mixture of g-priors and other methods in the case of highly correlated predictors in Gaussian linear models. We incorporate a non-parametric structure, the Dirichlet Process (DP) as a hyper prior, to allow more flexibility and adaptivity to the data.

To Chunying Lin and Yizhong Li

Contents

List of Tables

List of Figures

Acknowledgements

First, I am so grateful for my advisor Merlise Clyde, who has been extremely supportive to me over the years. She has everything a student could ask for in a mentor. I would like to thank her, for her patience and insights. She generously shares with me her experience and wisdom as a scholar, and encourages me to pursuit academic career. She also teaches me the rigorous attitude towards research. She sets a role model as an intelligent and knowledgeable scholar for me.

I would like to thank my committee members, Ed Iversen for introducing me to a colorful world of genetic research; thank Jim Berger and Beth Hauser for their insightful comments. I especially want to thank David Banks, a great mentor, for his generous help to both my research and job hunting. I would also like to thank Mine Çetinkaya-Rundel and Dalene Stangl, for their advice and help for my summer teaching. I am very thankful to Fan Li, for her valuable comments and ideas for both work and life. My further thanks go to my mentors at Avaya Labs, Lorraine Denby, Jim Landwehr and Pat Tendick. I learned a lot from them during my summer intern. I must also thank my English teacher, Diane Bryson, for her generous help with paper revising. I thank Andrew Cron, Cliburn Chan and Jacob Frelinger, for helping me implementing their HDPGMM model.

This research was supported by the National Science Foundation grant DMS-1106891. I thank the National Science Foundation (NSF) and the Statistical and Applied Mathematical Science Institute (SAMSI) for their financial support to my

graduate study.

I feel so fortunate to have made friends with a lot of kind and smart people, especially, Hongxia, Minhui, Jianyu, Fangpo and Monika. I learn from them and enjoy their company.

Last but not least, I would like to thank my parents and my husband. I am thankful to my parents, Guangsu Li and Juhong Ying, for their unconditional love and support. I am also very thankful to my husband, Qiang Wang, who makes me want to be a better person. I love you, too.

1

Introduction

In linear regressions, variable selection is routinely used to reduce model complexity and prevent over-fitting. From a Bayesian perspective, model selection is driven via prior specifications. In this dissertation, we develop two novel hierarchical priors for variable selection and model averaging. This chapter is organized as follows. Section 1 introduces the background of Bayesian model selection and model averaging in normal linear models. Section 2 describes the "spike-and-slab" prior, the class of prior distributions that contain point masses at zero as mixture components. Sections 3 and 4 review two widely utilized prior distributions, mixtures of g-priors and scale mixtures of independent normals respectively. Overviews of our new methods are included in these two sections.

1.1 Background: Bayesian Model Selection and Model Averaging in Linear Regression

From a model selection prospective, suppose we have $(q + p)$ number of potential predictors, among which the first q ones $\mathbf{X}_0 = (\mathbf{X}_{0,1}, \ldots, \mathbf{X}_{0,q})$ should always be included according to background information sources or the modeling structure (e.g. intercept); while a subset of the remaining p predictors $\mathbf{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_p)$ may be redundant or null predictors and may be excluded. We denote a normal linear regression model with predictors $(\mathbf{X}_0, \mathbf{V}_{\mathcal{M}})$ as model $\mathcal{M}$, which can be written as

$$\text{Model } \mathcal{M}: \quad \mathbf{Y} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{V}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_n) \tag{1.1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^T$ is the vector of n independent responses, and $\mathbf{V}_{\mathcal{M}}$ is the design matrix that consists of certain $p_{\mathcal{M}}$ columns of $\mathbf{V}$.

Bayesian solutions to the model selection problem require prior specifications on the model space, $p(\mathcal{M})$, and also on the parameters $\boldsymbol{\psi}_{\mathcal{M}} = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_{\mathcal{M}}, \sigma)$. After the prior specification, for each model $\mathcal{M}$, its marginal likelihood $f(\mathbf{Y} \mid \mathcal{M})$ and its posterior probability $p(\mathcal{M} \mid \mathbf{Y})$ can be computed as

$$f(\mathbf{Y} \mid \mathcal{M}) = \int f(\mathbf{Y} \mid \boldsymbol{\psi}_{\mathcal{M}}, \mathcal{M})\; p(\boldsymbol{\psi}_{\mathcal{M}} \mid \mathcal{M})\; d\boldsymbol{\psi}_{\mathcal{M}}$$

$$p(\mathcal{M} \mid \mathbf{Y}) = \frac{f(\mathbf{Y} \mid \mathcal{M})\; p(\mathcal{M})}{\sum_{\mathcal{M}'} f(\mathbf{Y} \mid \mathcal{M}')\; p(\mathcal{M}')}$$

A widely used selection criterion is to select the model with the highest posterior probability $p(\mathcal{M} \mid \mathbf{Y})$. In addition, model posterior probabilities also serves as weights in Bayesian model averaging (BMA), which uses the weighted average of posterior mean estimates of coefficients given each model,

$$\tilde{\boldsymbol{\beta}}_j = \sum_{\mathcal{M}} p(\mathcal{M} \mid \mathbf{Y})\; \mathbb{E}(\beta_j \mid \mathbf{Y}, \mathcal{M})\; \mathbf{1}_{\{\mathbf{X}_j \in \mathbf{X}_{\mathcal{M}}\}}$$

Therefore, for both model selection and parameter estimation, calculating marginal likelihoods $f(\mathbf{Y} \mid \mathcal{M})$ is essential.

1.2 Prior Distributions with Point Masses at Zero

For Bayesian model selection and model averaging, prior specification for parameters plays an important role. The "spike-and-slab" type of priors are popular choices for regression coefficients. Originally, the spike-and-slab prior (Mitchell and Beauchamp, 1988) refers to a mixture distribution of a points mass at zero (the spike) and a uniform distribution on a bounded interval centered at zero (the slab). This concept nowadays is usually used to describe a class of prior distributions that are mixtures of point masses at zero and continuous distributions or "spike-and-bell" priors. With the point masses, a subset of predictors can be excluded with positive probability, which can be treated as direct shrinkage to zero. The continuous components in the prior also pull the coefficients included in the model towards their prior centers, which are usually zero, to achieve another layer of shrinkage. When dealing with high-dimensional data, in addition to the spike-and-slab type of priors, another class of prior distributions, continuous shrinkage priors are also widely adopted. The density functions of these priors have high peaks around zero (or even diverge at zero), which can impose heavy shrinkage on the coefficients towards zero but cannot strictly exclude predictors unless posterior mode estimates are used, or some additional decision theoratic approach is adapted.

1.3 The g-prior and Mixture of g-priors

Among the spike-and-slab priors, Zellner's g-prior is a very popular choice. In the regression problem $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, when there is some information about the value of the coefficient $\boldsymbol{\beta}$ but little information about σ and the prior covariance of $\boldsymbol{\beta}$, Zellner (1986) proposes the g-prior on $(\boldsymbol{\beta}, \sigma)$,

$$\begin{aligned} \boldsymbol{\beta} \mid g, \sigma &\sim \mathrm{N}\left(\boldsymbol{\beta}_0, g\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \\ p(\sigma) &\propto 1/\sigma \end{aligned}$$

which incorporates the possible value of the coefficient through the prior mean $\boldsymbol{\beta}_0$. Since for variable selection problems, selecting variables $\mathbf{X}$ is equal to testing hypotheses $H_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $H_a : \boldsymbol{\beta} \neq \mathbf{0}$, hence here the possible value of the coefficient is $\boldsymbol{\beta}_0 = \mathbf{0}$. In the g-prior, the normal standard deviation σ typically has an improper diffuse prior. Improper priors introduce arbitrary constants into the marginal likelihoods generally leading to ill determined Bayes factors, which may invalidate model comparison based on Bayes factors. Hence Bayarri et al. (2012) proposes the Basic Criterion for priors in model selection, which suggests that all model specific parameters should have proper conditional prior distributions. Common orthogonal parameters are exceptions, due to the cancellation of the vague constants in the Bayes factors (Berger et al., 1998).

It is convenient to consider an equivalent parameterization of model (1.1) so that the common predictors $\mathbf{X}_0$ and remaining model specific predictors are orthogonal for all models. To achieve orthogonality, in (1.1) we decompose $\mathbf{V}_{\mathcal{M}}$ by projecting it onto the hyper plane spanned by the columns of $\mathbf{X}_0$,

$$\begin{aligned} \text{Model } \mathcal{M}: \quad \mathbb{E}(\mathbf{Y}) &= \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathcal{P}_{\mathbf{X}_0}\mathbf{V}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} + (\mathbf{I}_n - \mathcal{P}_{\mathbf{X}_0})\mathbf{V}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} \qquad (1.2) \\ &= \mathbf{X}_0\boldsymbol{\alpha} + \mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} \qquad (1.3) \end{aligned}$$

where $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + (\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{V}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}$ is the parameters on common predictors after translation, $\mathcal{P}_{\mathbf{X}_0} = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T$ is the projection matrix, and

$$\mathbf{X}_{\mathcal{M}} = (\mathbf{I}_n - \mathcal{P}_{\mathbf{X}_0})\mathbf{V}_{\mathcal{M}} \qquad (1.4)$$

is the new model specific predictors such that

$$\mathbf{X}_0^T\mathbf{X}_{\mathcal{M}} = \mathbf{0} \qquad (1.5)$$

Formula (1.5) implies that the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ are orthogonal in the sense of the information matrix of $(\boldsymbol{\alpha}, \boldsymbol{\beta}_{\mathcal{M}})$ being block diagonal. Note that the above orthogonality holds under all 2^p models, so $\boldsymbol{\alpha}$ can be considered as a common parameter among different models. In the special case where the only common predictor is the intercept $\mathbf{X}_0 = \mathbf{1}_n$, in normal linear regression transforming $\mathbf{V}_{\mathcal{M}}$ to $\mathbf{X}_{\mathcal{M}}$ is equivalent to centering the columns of $\mathbf{V}_{\mathcal{M}}$. And after this orthogonalization, the intercept α can be considered as the center of $\mathbf{Y}$, which does not change with any specific model $\mathcal{M}$. Therefore, in linear models according to the most widely used version of the Zellner's g-prior, the intercept α has an improper flat prior (e.g., Liang et al. (2008), the null-based approach)

$$\boldsymbol{\beta}_{\mathcal{M}} \mid g, \sigma, \mathcal{M} \sim \mathrm{N}\left(\mathbf{0}, g\sigma^2(\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}})^{-1}\right) \tag{1.6}$$

$$p(\alpha, \sigma \mid \mathcal{M}) \propto 1/\sigma \tag{1.7}$$

This version of the g-prior has several ideal properties. The marginal likelihoods yielded by it are in closed form expression, and can be represented as simple functions of the coefficient of determination or R^2. In addition, it maintains the same correlation structure in the prior distribution as the likelihood and is invariant under orthogonal transformation of designs. However, choosing the value of the hyper parameter g is not straight-forward. Arbitrary values of g in the g-prior usually lead to the information paradox (Liang et al., 2008). In addition, Lindley's paradox occurs when g is large, because the prior density is too flat and hence always favors the smaller model. To resolve these problems, fully Bayes approaches propose prior distributions on g, e.g. Zellner and Siow (1980), Liang et al. (2008), Maruyama and George (2011), Bayarri et al. (2012), Celeux et al. (2012), Ley and Steel (2012).

1.3.1 Overview of Chapter 2

New mixtures of g-priors have been extensively studied in linear models, however choice of prior distributions in Generalized Linear Models (GLMs) remains an open problem. In Chapter 2 of this thesis we extend mixtures of g-priors to Generalized Linear Models (GLMs) by assigning a conjugate prior, the Confluent Hypergeometric distribution, to the shrinkage factor $\frac{g}{1+g}$. Our CH-g prior encompasses common mixtures of g-priors in the literature such as the Hyper-g prior, and naturally extends them to be applicable in GLMs. Under a Laplace approximation, it yields marginal likelihoods in computationally tractable forms. We demonstrate theoretically the asymptotic consistency for model selection and BMA estimation holds under the CH-g prior. With our default choice of hyper parameters, the CH-g prior satisfies the intrinsic consistency of Bayarri et al. (2012) implicitly. In addition, we illustrate its use in simulation and real examples.

1.4 Scale Mixtures of Independent Normals

In addition to mixtures of g-priors, shrinkage methods with continuous priors in the family of scale mixtures of independent normals (West, 1987) are also prevalently used, for example, the relevance vector machine (Tipping, 2001), the Normal-exponential-gamma prior (Griffin and Brown, 2005), the Bayesian lasso (Park and Casella, 2008), (Hans, 2009), the Bayesian elastic net (Li and Lin, 2010) and the horseshoe (Carvalho et al., 2010). Under orthonormal designs, the (conditional) posterior mean of each regression coefficient may can be represented as the MLE multiplied by a shrinkage factor, which takes value between 0 and 1.

The posterior distribution under the g-prior inherits the instability of ordinary least square (OLS) estimate when the design matrix is nearly singular. Ridge regression, lasso estimates or estimates under scale mixtures of independent normals are not as affected by the correlation among the predictors. Carvalho et al. (2010) claim that the horseshoe performs almost as well as the gold standard of Bayesian model averaging (BMA) under the Zellner-Siow prior for prediction. However, continuous priors cannot shrink coefficients to exact zeros, and lack selection procedures that can be validated by optimizing any loss function. On the other hand, "spike-and-slab" priors (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005; Scott and Berger, 2006) allow coefficients to be exactly zero (so that they can be excluded from the model) by adding positive probability masses at zero to the priors. Our results suggest that scale mixtures of independent normals may out-perform the mixtures of

g-priors if the predictors are highly correlated.

1.4.1 Overview of Chapter 3

In normal linear regression, empirical studies suggest that ridge regression outperforms the lasso in parameter estimation and prediction when regression coefficients are small or covariates are highly correlated. Unlike the lasso, which depends on the choice of coordinate system used to represent the model, ridge regression is invariant under the orthogonal rotation of the explanatory variables. Inspired by the rotation invariant property of ridge regression, in Chapter 3 we propose the Local Rotation Invariant prior (LoRI). This Bayesian approach has a local rotation invariant structure, which is induced by the DP prior on variance parameters in normal prior distributions for the regression coefficients. Due to the natural grouping structure induced by the DP, our shrinkage prior acts like a multivariate Cauchy prior within the group. Point masses at zero in the DP base measure can achieve sparse solutions like the lasso or "spike-and-slab" type of Bayesian variable selection priors. Compared with continuous shrinkage methods, it has the advantage of valid built-in variable selection. Meanwhile, the Cauchy tails of the prior lead to bounded prior influence that can preserves large effects. Both simulation and real-world examples show that the LoRI achieves high accuracy in parameter estimation and prediction.

2

The Confluent Hypergeometric g-prior for GLMs

2.1 Introduction

In linear regression, mixtures of g-priors (Zellner and Siow, 1980; Liang et al., 2008; Maruyama and George, 2011; Bayarri et al., 2012; Celeux et al., 2012; Ley and Steel, 2012) are widely used for model selection and model averaging. They yield (exact or approximate) marginal likelihoods in tractable form, which may avoid sampling regression coefficients in MCMC to achieve computational efficiency. They maintain correlation structure among predictors by allowing the correlation in the prior covariance to mimic that induced by the likelihood, which also leads to their invariance under change of measurement. Mixtures of g-priors not only inherit the ideal features of the g-prior, but also resolve the information paradox (Liang et al., 2008) and Lindley's paradox (Lindley, 1968) that occur under fixed g.

In this paper, we build a unified framework of mixture of g-priors for GLMs. Our hyper prior on g based on the Confluent Hypergeometric distribution encompasses most common hyper priors, such as the Hyper-g prior, and naturally extends their corresponding mixtures of g-priors to GLMs. Under a Laplace approximation, our choice of hyper prior is conjugate, and yields computationally tractable forms for marginal likelihoods. We provide conditions for asymptotical consistency of model selection and parameter estimation under our mixture of g-prior for GLMs.

Section 2 reviews the g-prior for GLMs. Section 3 develops the mixture of g-priors for GLMs. Section 4 examines the model selection consistency, information consistency and Bayesian model averaging consistency. Section 5 discusses our default choices of hyper parameters, and shows its performance in both simulation and real examples.

2.2 The Generalized g-prior for GLMs

2.2.1 Generalized Linear Models

Suppose that the n dimensional response vector $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ follows a distribution in the exponential family, and according to McCullagh and Nelder (1989), the likelihood function can be written as

$$f(\mathbf{Y} \mid \boldsymbol{\theta}, \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} + c(Y_i, \phi) \right\}, \tag{2.1}$$

where $a(\cdot), b(\cdot)$ and $c(\cdot, \cdot)$ are specific functions which determine the GLM density. The mean and variance for an observation Y can be written using these functions:

$$\mathbb{E}(Y) = b'(\theta), \tag{2.2}$$

$$\mathbb{V}(Y) = a(\phi)b''(\theta), \tag{2.3}$$

where $b'(\cdot)$ and $b''(\cdot)$ are the first and second order derivatives. Due to (2.3), it is reasonable to assume that $b''(\cdot) \geqslant 0$ in most cases. The canonical parameter $\theta_i = \theta(\eta_i)$ can be connected with the linear combination of predictors $\mathbf{V}_i$, i.e.,

$$\boldsymbol{\eta} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{V}\boldsymbol{\beta} \tag{2.4}$$

by the link function $\theta(\cdot)$, where $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)$. In particular, the canonical link, $\theta_i(\eta_j) = \eta_i$, is the most widely used form of link. We restrict the scale $a(\phi) = 1$, which includes many common exponential family distributions, such as Bernoulli, Poisson and Normal with known variance (see Table 2.1).

Table 2.1: Three commonly used distributions in the exponential family.

distribution	$a(\phi)$	θ	$b(\theta)$	$b'(\theta)$	$b''(\theta)$
$\mathrm{N}(\mu, \sigma^2)$	σ^2	μ	$\frac{\theta^2}{2}$	θ	1
$\mathrm{Ber}(p)$	1	$\log \frac{p}{1-p}$	$\log(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta}$	$\frac{e^\theta}{(1+e^\theta)^2}$
$\mathrm{Poi}(\lambda)$	1	$\log \lambda$	e^θ	e^θ	e^θ

Rather than using all predictors, we may wish to consider models based on a subset of $\mathbf{V}$. Suppose $\mathbf{X}_0$ is the set of predictors common all models and $\mathbf{V}_{\mathcal{M}}$ is the subset of $\mathbf{V}$ in model $\mathcal{M}$, then we can write model in (2.4) as

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{X}_0 \boldsymbol{\alpha}_{0,\mathcal{M}} + \mathbf{V}_{\mathcal{M}} \boldsymbol{\beta}_{\mathcal{M}}, \tag{2.5}$$

where typically, $\mathbf{X}_0 = \mathbf{1}_n$.

In normal linear models, the most common variant of the g-prior is

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \sigma \sim N\left(\mathbf{0}, g\, \mathcal{I}_n^{-1}(\boldsymbol{\beta}_{\mathcal{M}})\right),$$

where $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}} / \sigma^2$. The precision matrix (i.e., inverse covariance) of this g-prior equals the inverse of the hyper parameter g multiplied by the expected information matrix based on all n observations, which is the same as the observed information. Extensions are more complicated for non-Gaussian distributions in the exponential family, because their information matrices depend on the unknown coefficient parameters. Bové and Held (2011) evaluate the expected information matrix at the prior mode zero, while Hansen and Yu (2003) at the MLE estimates $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$. Wang and George (2007) also evaluate the information at the MLE, but use the observed information matrix instead. Gupta and Ibrahim (2009) avoid this choice by keeping the unknown parameter $\boldsymbol{\beta}_{\mathcal{M}}$ in the prior precision matrix, which leads to intractable marginal likelihoods.

2.2.2 "Centering" the Predictors

Bové and Held (2011) point out that majority of the current variants of g-priors for GLMs do not treat the common parameters across models, usually the intercept, differently from the model specific coefficients, so that $\mathbf{X}_0 = \emptyset$. This means that the intercept or other common parameters are shrunk towards zero along with the coefficients, which may be problematic when the true intercept is large relative to the regression coefficients. In the extreme case, in normal linear models if the true intercept approaches infinity, and g is allowed to adapt to the data, then the null model is selected. Hence it is desirable to assume the common parameters and the model specific parameters are independent *a priori*. Motivated by the projection procedure (1.4) in normal linear models, which ensures orthogonality between the common variables $\mathbf{X}_0$ and the model specific predictors $\mathbf{X}_{\mathcal{M}}$, we propose a "centering" procedure for likelihood densities in the exponential family, to ensure that the expected Fisher information is block diagonal.

Proposition 1. *Under any model $\mathcal{M}$, we propose the following "centering" procedure to transform its model specific predictors $\mathbf{V}_{\mathcal{M}}$ to $\mathbf{X}_{\mathcal{M}}$,*

$$\mathbf{X}_{\mathcal{M}} = \left[\mathbf{I}_n - \hat{\mathcal{P}}_{\mathbf{X}_0}\right] \mathbf{V}_{\mathcal{M}}, \tag{2.6}$$

$$\hat{\mathcal{P}}_{\mathbf{X}_0} = \mathbf{X}_0 \left(\mathbf{X}_0^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_0\right)^{-1} \mathbf{X}_0^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}), \tag{2.7}$$

$$\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{X}_0 \boldsymbol{\alpha}_{\mathcal{M}} + \mathbf{X}_{\mathcal{M}} \boldsymbol{\beta}_{\mathcal{M}}, \tag{2.8}$$

where $\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$ is the expected information matrix of $\boldsymbol{\eta}_{\mathcal{M}} = (\eta_{1,\mathcal{M}}, \ldots, \eta_{n,\mathcal{M}})^T$ evaluated at its MLE based on all n observations. After this reparameterization,

$$\mathbf{X}_0^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}} = \mathbf{0}, \tag{2.9}$$

which leads to the result that the expected information matrix for $(\boldsymbol{\alpha}_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}})$ evaluated *at the MLE*

$$\mathcal{I}_n\left(\hat{\boldsymbol{\alpha}}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) = \begin{bmatrix} \mathcal{I}_n(\hat{\boldsymbol{\alpha}}_{\mathcal{M}}) & \mathbf{0}_n^T \\ \mathbf{0}_n & \mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}}) \end{bmatrix} \tag{2.10}$$

is block diagonal. Note that $\hat{\mathcal{P}}_{\mathbf{X}_0}$ is an orthogonal projection on the column space $\mathbf{X}_0$ with inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{y}$.

Proof. Since the linear combination $\boldsymbol{\eta}_{\mathcal{M}}$ does not change under the translation operator, due to the functional invariance of MLEs, $\hat{\boldsymbol{\eta}}_{\mathcal{M}}$ remains the same after (2.6). We can simply verify that the off-diagonal block of the information matrix equals zero (2.9). □

In most GLM variable selection problems, the only common predictor is the intercept $\mathbf{X}_0 = \mathbf{1}_n$. Then after the "centering" step (2.6), the j-th predictor

$$\mathbf{X}_j = \mathbf{V}_j - \mathbf{1}_n \tilde{v}_{j,\mathcal{M}}, \tag{2.11}$$

where $\tilde{v}_{j,\mathcal{M}}$ is a weighted average of elements of the vector $\mathbf{V}_j$ and the weights depending on the information matrix $\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$. In particularly, under normal linear models, these weights are equal and thus $\tilde{v}_{j,\mathcal{M}}$ becomes the column-wise average. Except for normal distributions, the "centering" procedure (2.6) is model specific due to its dependence on model specific MLEs in the inner product $\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})$. Due to the asymptotic consistency of the MLE, we now treat the parameter $\boldsymbol{\alpha}_{\mathcal{M}}$ as a common parameter across models, and $\boldsymbol{\alpha}_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ are treated differently by having independent prior distributions, i.e.,

$$p(\boldsymbol{\alpha}_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}}) = p(\boldsymbol{\alpha}_{\mathcal{M}})\; p(\boldsymbol{\beta}_{\mathcal{M}}). \tag{2.12}$$

The "centering" step also simplifies the calculation of the marginal likelihood. Under most of the distributions in the GLM family (except for normal distribution), the marginal likelihood does not have a closed form. To calculate the marginal likelihood, we apply a Laplace approximation (Tierney and Kadane, 1986) that utilizes a second order Taylor expansion around the MLE $(\hat{\boldsymbol{\alpha}}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})$.

$$p(\mathbf{Y} \mid \mathcal{M}) = \int f_{\mathcal{M}}(\mathbf{Y} \mid \boldsymbol{\alpha}_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}}) p(\boldsymbol{\alpha}_{\mathcal{M}}) p(\boldsymbol{\beta}_{\mathcal{M}}) \; d(\boldsymbol{\alpha}_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}}) \tag{2.13}$$

$$= f_{\mathcal{M}}(\mathbf{Y} \mid \hat{\boldsymbol{\alpha}}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}) \int e^{-\frac{1}{2}(\boldsymbol{\alpha}_{\mathcal{M}} - \hat{\boldsymbol{\alpha}}_{\mathcal{M}})^T \mathcal{I}_n(\hat{\boldsymbol{\alpha}}_{\mathcal{M}})(\boldsymbol{\alpha}_{\mathcal{M}} - \hat{\boldsymbol{\alpha}}_{\mathcal{M}})} \; p(\boldsymbol{\alpha}_{\mathcal{M}}) \; d\boldsymbol{\alpha}_{\mathcal{M}} \tag{2.14}$$

$$\cdot \int e^{-\frac{1}{2}(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}})^T \mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})(\boldsymbol{\beta}_{\mathcal{M}} - \hat{\boldsymbol{\beta}}_{\mathcal{M}})} \; p(\boldsymbol{\beta}_{\mathcal{M}}) \; d\boldsymbol{\beta}_{\mathcal{M}} + O(n^{-1}). \tag{2.15}$$

According to Kass et al. (1990), this Laplace approximation is precise to the order of $O(n^{-1})$. The "centering" step combined with independent prior distributions for $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ allows us to approximate the marginal likelihood by integrating out $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ separately. Next, we will describe the g-prior for GLMs that we adopt, which leads to closed form marginal likelihood under the Laplace approximation (2.14), as well as extensions to mixtures of g-priors.

2.2.3 The g-prior for GLMs

In normal linear models, Zellner's g-prior for (1.6) (1.7), assigns the model specific coefficient $\boldsymbol{\beta}_{\mathcal{M}}$ a multivariate normal prior distribution centered at zero, and the inverse of its prior covariance is proportional to the information matrix $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) = \mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}} / \sigma^2$. In GLMs, the expected information matrix becomes

$$\begin{aligned} \mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) &= \mathbf{X}_{\mathcal{M}}^T \; \mathcal{I}_n(\boldsymbol{\eta}_{\mathcal{M}}) \; \mathbf{X}_{\mathcal{M}} \\ &= \mathbf{X}_{\mathcal{M}}^T \; [\Delta(\boldsymbol{\eta}_{\mathcal{M}}) \; \mathcal{I}_n(\boldsymbol{\theta}_{\mathcal{M}}) \; \Delta(\boldsymbol{\eta}_{\mathcal{M}})] \; \mathbf{X}_{\mathcal{M}}, \end{aligned}$$

where $\Delta(\boldsymbol{\eta})$ denotes the diagonal matrix whose i-th element is $\frac{d\theta_i}{d\eta_i}$ evaluated at the i-th linear predictor $\eta_{i,\mathcal{M}} = \alpha_{\mathcal{M}} + \mathbf{x}_{i,\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}}$, and $\mathcal{I}_n(\boldsymbol{\theta}_{\mathcal{M}})$ denotes the expected information matrix of $\boldsymbol{\theta}_{\mathcal{M}}$ based on all n data points. Under canonical links, $\Delta(\boldsymbol{\eta}_{\mathcal{M}})$ becomes the identity matrix, and hence $\mathcal{I}_n(\boldsymbol{\eta}_{\mathcal{M}})$ becomes the diagonal matrix with elements $b''(\eta_{i,\mathcal{M}})$.

In GLMs, after "centering" the design matrix to $\mathbf{X}_{\mathcal{M}}$, we propose the following definition of the g-prior under model $\mathcal{M}$. We let $\boldsymbol{\beta}_{\mathcal{M}}$ have a normal prior with mean

$\mathbf{0}$ and covariance being proportional to inverse of the expected information matrix, and let the intercept $\alpha_{\mathcal{M}}$ have an independent normal prior:

$$\boldsymbol{\beta}_{\mathcal{M}} \mid g, \mathcal{M} \sim \mathrm{N}_{p_{\mathcal{M}}}\left(\mathbf{0},\ g \cdot \mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})^{-1}\right), \tag{2.16}$$

$$p(\alpha_{\mathcal{M}} \mid \mathcal{M}) \sim \mathrm{N}(0, nc), \tag{2.17}$$

where g is a positive parameter, $\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})$ is the expected information matrix based on all the data and evaluated at the MLE $(\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})$ in the form of $\hat{\boldsymbol{\eta}}_{\mathcal{M}}$, and c is a non-negative constant. In the literature, such data dependent priors have been proposed, for example, Kass and Wasserman (1995), Hansen and Yu (2003) and Wang and George (2007). Notice as when $c = \infty$, (2.17) degenerates to the flat prior $p(\alpha_{\mathcal{M}}) \propto 1$. Although in linear model, the flat prior on the intercept is a prevalent choice in almost all existing variants of g-priors that treat the intercept and coefficient separately, we will show in Section 2.4.2 that this may be problematic for other GLM functions.

2.2.4 Laplace Approximate of the Bayes Factor

As discussed in (2.14), we utilize the Laplace approximation to calculate the marginal likelihood for model $\mathcal{M}$. The normal densities of $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ from the likelihood can be combined with the independent normal prior densities on $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ (2.16) (2.17) respectively. Hence we obtain the approximate marginal likelihood in analytical form,

$$p\left(\mathbf{Y} \mid g, \mathcal{M}\right) = f_{\mathcal{M}}\left(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) \left[1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})nc\right]^{-\frac{1}{2}} e^{-\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})\hat{\alpha}^2_{\mathcal{M}}}{2(\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})nc+1)}} \tag{2.18}$$

$$\cdot (1+g)^{-\frac{p_{\mathcal{M}}}{2}} e^{-\frac{Q_{\mathcal{M}}}{2(1+g)}} + O(n^{-1}), \tag{2.19}$$

where

$$Q_{\mathcal{M}} = \left[\hat{\boldsymbol{\beta}}^T_{\mathcal{M}} \mathbf{X}^T_{\mathcal{M}}\right] \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \left[\mathbf{X}_{\mathcal{M}} \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right] \tag{2.20}$$

is the analogue of the regression sum of squares in the linear model, $p_{\mathcal{M}}$ is the number of predictors in $\mathbf{X}_{\mathcal{M}}$, and $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) = \mathbf{1}_n^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{1}_n$ is the expected information of the intercept. Note that for the null model $\mathcal{M}_{\o}$ where $p_{\mathcal{M}_{\o}} = 0$, we can let $Q_{\mathcal{M}_{\o}} = 0$ and then (2.18) remains to hold. When comparing two models $\mathcal{M}_2$ to $\mathcal{M}_1$, the Bayes factor under g-prior can be approximated as the ratio of their marginal likelihoods, which can be rewritten as

$$\mathrm{BF}_{\mathcal{M}_2:\mathcal{M}_1} = \Lambda_{\mathcal{M}_2:\mathcal{M}_1} \cdot \Omega_{\mathcal{M}_2:\mathcal{M}_1} + O(n^{-1}), \tag{2.21}$$

which is decomposed to the product of

$$\Lambda_{\mathcal{M}_2:\mathcal{M}_1} = \frac{f_{\mathcal{M}_2}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_2}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_2})}{f_{\mathcal{M}_1}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_1}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_1})} \left[\frac{1+\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_2})nc}{1+\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_1})nc}\right]^{-\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_2})\hat{\alpha}^2_{\mathcal{M}_2}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_2})nc+1} - \frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_1})\hat{\alpha}^2_{\mathcal{M}_1}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_1})nc+1}\right]} \tag{2.22}$$

and

$$\Omega_{\mathcal{M}_2:\mathcal{M}_1} = \frac{(1+g)^{-\frac{p_{\mathcal{M}_2}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}_2}}{2(1+g)}\right\}}{(1+g)^{-\frac{p_{\mathcal{M}_1}}{2}} \exp\left\{-\frac{Q_{\mathcal{M}_1}}{2(1+g)}\right\}}. \tag{2.23}$$

The first term $\Lambda_{\mathcal{M}_2:\mathcal{M}_1}$ consists of the maximized likelihood ratio and the penalties contributed by the intercept. The second term $\Omega_{\mathcal{M}_2:\mathcal{M}_1}$ comes from the generalized g-prior on the coefficients. In particular, the choice of g effects the Bayes factor only through $\Omega_{\mathcal{M}_2:\mathcal{M}_1}$.

Note that if $c = \infty$, i.e., the prior distribution on $\alpha_{\mathcal{M}}$ is the flat prior, then the approximate Bayes factor and its corresponding $\Lambda_{\mathcal{M}_2:\mathcal{M}_1}$ become

$$p\left(\mathbf{Y}|g, \mathcal{M}\right) = f_{\mathcal{M}}\left(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) (2\pi)^{\frac{1}{2}} \left[\mathcal{I}_n(\hat{\boldsymbol{\alpha}}_{\mathcal{M}})\right]^{-\frac{1}{2}} (1+g)^{-\frac{p_{\mathcal{M}}}{2}} e^{-\frac{Q_{\mathcal{M}}}{2(1+g)}} + O(n^{-1}), \tag{2.24}$$

where

$$\Lambda^{c=\infty}_{\mathcal{M}_2:\mathcal{M}_1} = \frac{f_{\mathcal{M}_2}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_2}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_2}) \left[\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_2})\right]^{-\frac{1}{2}}}{f_{\mathcal{M}_1}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_1}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_1}) \left[\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_1})\right]^{-\frac{1}{2}}}. \tag{2.25}$$

2.2.5 Approximate Conditional Posterior Distributions

For any given model $\mathcal{M}$, here we consider the conditional posterior distributions of $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ under our g-prior (2.16), (2.17) for GLMs. For notation simplification, when there is no ambiguity, we omit the subscript $\mathcal{M}$. Except for the normal distribution, other likelihood densities in GLMs (2.1) are not conjugate with normal prior on $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$. Fortunately, according to the standard Bayesian asymptotic theory (see Bernardo and Smith (2000), p287), as n increases, the conditional posterior densities based on observed data $\{\mathbf{Y}, \mathbf{V}\} = \{(Y_1, \mathbf{v}_1), \dots, (Y_n, \mathbf{v}_n)\}$ converges to normal densities,

$$\alpha_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M} \xrightarrow{\mathrm{d}} \mathrm{N}\left(\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) + \frac{1}{nc}}\, \hat{\alpha}_{\mathcal{M}},\ \frac{1}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) + \frac{1}{nc}}\right), \tag{2.26}$$

$$\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, g, \mathcal{M} \xrightarrow{\mathrm{d}} \mathrm{N}\left(\frac{g}{1+g}\, \hat{\boldsymbol{\beta}}_{\mathcal{M}},\ \frac{g}{1+g}\left[\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\right]^{-1}\right), \tag{2.27}$$

hence we can use these normal distributions as approximates to the conditional posterior distributions. Note that when flat prior is assigned to $\alpha_{\mathcal{M}}$, i.e., $c = \infty$, its approximate conditional posterior is

$$\alpha_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M} \xrightarrow{\mathrm{d}} \mathrm{N}\left(\hat{\alpha}_{\mathcal{M}},\ [\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})]^{-1}\right).$$

Similar to the posterior distribution under Zellner's g-prior in the normal linear models, the approximate conditional posterior mean of $\boldsymbol{\beta}_{\mathcal{M}}$ is shrunk from the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ towards $\mathbf{0}$. We donate the ratio $z = g/(1+g)$ between the posterior mean $\mathbb{E}(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, g, \mathcal{M})$ and the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ as the shrinkage factor. Assume that the expected information $\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})$ based on all data is proportional to n, then the approximate posterior covariance of $\boldsymbol{\beta}_{\mathcal{M}}$ is proportional with $1/n$. Therefor, the conditional posterior $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, g, \mathcal{M})$ becomes more concentrated around its mean as n increases.

2.2.6 Inconsistency of the g-prior

For our model selection problem, suppose among all the 2^p different models, there exists a true model $\mathcal{M}_T$ that generates the data. Under the true model $\mathcal{M}_T$, the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}_T}$ converges to the true parameter $\boldsymbol{\beta}^*_{\mathcal{M}_T}$, while the conditional posterior mean $\mathbb{E}[\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, g, \mathcal{M}]$ becomes more concentrated around $\hat{\boldsymbol{\beta}}_{\mathcal{M}_T}\, g/(1+g)$. Therefore, with any fixed value of g, the posterior mean estimate of $\boldsymbol{\beta}_{\mathcal{M}_T}$ is biased asymptotically.

In addition to the inconsistency in parameter estimation, g-priors with fixed g also exhibits inconsistency in model selection. In normal linear models, Liang et al. (2008) points out the selection inconsistency of g-prior with fixed g. They also suggest that some fully Bayes methods that assigns prior distributions on g, such as the Zellner-Siow prior (Zellner and Siow, 1980), the Hyper-g prior and the Hyper-g/n prior (Liang et al., 2008) can partially or completely resolve this inconsistency. We find that in GLMs, this inconsistency also exists with fixed g. The following counter example shows that in normal linear model with fixed variance, when comparing two nested models, if the smaller model is the true model $\mathcal{M}_T$, the Bayes factor for $\mathcal{M}_T$ compared to $\mathcal{M}$ under g-prior does not go to ∞ asymptotically.

Remark 1. *Under normal linear model with known variance $\sigma^2 = 1$, for any fixed value of g and any model $\mathcal{M} \supset \mathcal{M}_T$, as the sample size n increases, the Bayes factor under the g-prior* (2.16), (2.17)

$$BF_{\mathcal{M}_T:\mathcal{M}} = O(1),$$

which implies the selection consistency does not hold for g-prior with fixed g.

Proof: see Appendix A.2.1.

2.3 The Confluent Hypergeometric Prior on g

We propose a hierarchical prior distribution on g to resolve the inconsistency. Based on the g-prior (2.16), (2.17), we assign a hyper prior distribution,

$$p(g \mid a, b, s) = \frac{g^{\frac{a}{2}-1}(1+g)^{-\frac{a+b}{2}} \exp\left[\frac{s}{2}\left(\frac{g}{1+g}\right)\right]}{B(\frac{a}{2}, \frac{b}{2})\ {}_1F_1(\frac{a}{2}, \frac{a+b}{2}, \frac{s}{2})}, \tag{2.28}$$

where parameters $a > 0, b > 0, s \geqslant 0$, and ${}_1F_1$ is the confluent hypergeometric function (Abramowitz and Stegun, 1970). (See Appendix A.1 for definition of the ${}_1F_1$ function). Gordy (1998a) proposes the Confluent Hypergeometric (CH) distribution, which can be considered as a generalization of Beta distribution and has the following density function

$$p_{\mathrm{CH}}(z \mid a, b, s) = \frac{z^{a-1}(1-z)^{b-1}\exp(-sz)}{B(a, b)\ {}_1F_1(a, a+b, -s)},\ 0 \leqslant z \leqslant 1, \tag{2.29}$$

where parameters $a > 0, b > 0$ and $s \in \mathbb{R}$. When $s = 0$, the $\mathrm{CH}(a, b, s)$ distribution degenerates to $\mathrm{Beta}(a, b)$ distribution. When transforming g to the shrinkage factor z, the prior distribution (2.28) becomes a CH distribution on z, i.e.,

$$z = \frac{g}{1+g} \sim \mathrm{CH}\left(\frac{a}{2}, \frac{b}{2}, -\frac{s}{2}\right), \tag{2.30}$$

which guarantees that the prior distribution (2.28) is well-defined. It is also a conjugate prior, in that the conditional posterior distribution of z also has a CH distribution,

$$z \mid \mathbf{Y}, \mathcal{M} \sim \mathrm{CH}\left(\frac{a}{2}, \frac{b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right), \tag{2.31}$$

where $p_{\mathcal{M}}$ is the model size of $\mathcal{M}$, and $Q_{\mathcal{M}} = \hat{\boldsymbol{\beta}}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}}^T \mathcal{I}_n\left(\hat{\boldsymbol{\eta}}_{\mathcal{M}}\right) \mathbf{X}_{\mathcal{M}} \hat{\boldsymbol{\beta}}_{\mathcal{M}}$ is the analogue of RSS in GLMs. We denote the hierarchical g-prior (2.16), (2.17) and (2.28) as the CH-g prior.

2.3.1 Tail Behavior of the CH-g Prior

Heavy-tailed prior distributions on $\boldsymbol{\beta}_{\mathcal{M}}$ are desirable in model selection since they are robust to large coefficients in terms of not over-shrinking them. Most state of the art prior distributions on g yield the prior densities of $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M})$ with multivariate Student tails, for example, Zellner and Siow (1980), Liang et al. (2008), Maruyama and George (2011) and Bayarri et al. (2012). The following proposition shows that under the CH-g prior, the prior distribution on $\boldsymbol{\beta}_{\mathcal{M}}$ also behaves as a multivariate Student distribution in the tails.

Proposition 2. *Under the CH-g prior* (2.16), (2.17) *and* (2.28), *the marginal prior distribution under model* $\mathcal{M}$

$$p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}) = \int p(\boldsymbol{\beta}_{\mathcal{M}} \mid g, \mathcal{M}) p(g) dg$$

has tails behave as multivariate Student distribution with degrees of freedom b *and scale matrix* $\left[\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\right]^{-1}$, *i.e.,*

$$\lim_{\|\boldsymbol{\beta}_{\mathcal{M}}\| \to \infty} p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}) \propto \left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}\right)^{-\frac{b+p_{\mathcal{M}}}{2}}, \tag{2.32}$$

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

Proof: see Appendix A.2.2.

The choice of the hyper parameter b alone determines the tail behavior of the marginal prior $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M})$. In particular, $b = 1$ corresponds to Cauchy tails.

2.3.2 Approximate Bayes Factor under the CH-g Prior

Similar to the g-prior for GLMs, the CH-g prior also yields closed-form marginal likelihood under Laplace approximations. Denote $u = 1 - z$, then the prior distribution on z (2.30) is equivalent to

$$u = \frac{1}{1+g} \sim \mathrm{CH}\left(\frac{b}{2}, \frac{a}{2}, \frac{s}{2}\right). \tag{2.33}$$

Hence we can integrate g (in the form of u) out from (2.18) and obtain the approximate marginal likelihood under the CH-g prior for model $\mathcal{M}$

$$\begin{aligned} p(\mathbf{Y} \mid \mathcal{M}) &= \int_0^1 p(\mathbf{Y} \mid u, \mathcal{M}) p(u) du \\ &= f_{\mathcal{M}}\left(\mathbf{Y} | \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) \left[1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc\right]^{-\frac{1}{2}} e^{-\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})\hat{\alpha}^2_{\mathcal{M}}}{2(\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})nc+1)}} \\ &\quad \cdot \frac{B\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)}{B\left(\frac{b}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, -\frac{s}{2}\right)} + O(n^{-1}). \end{aligned}$$

Therefore, the Bayes factor comparing $\mathcal{M}_2$ to $\mathcal{M}_1$ under the CH-g prior can be approximated as

$$\mathrm{BF}_{\mathcal{M}_2:\mathcal{M}_1} = \Lambda_{\mathcal{M}_2:\mathcal{M}_1} \cdot \Omega^{\mathrm{CH}}_{\mathcal{M}_2:\mathcal{M}_1} + O(n^{-1}), \tag{2.34}$$

where $\Lambda_{\mathcal{M}_2:\mathcal{M}_1}$ remains the same as in (2.22) and

$$\Omega^{\mathrm{CH}}_{\mathcal{M}_2:\mathcal{M}_1} = \frac{B\left(\frac{b+p_{\mathcal{M}_2}}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b+p_{\mathcal{M}_2}}{2}, \frac{a+b+p_{\mathcal{M}_2}}{2}, -\frac{s+Q_{\mathcal{M}_2}}{2}\right)}{B\left(\frac{b+p_{\mathcal{M}_1}}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b+p_{\mathcal{M}_1}}{2}, \frac{a+b+p_{\mathcal{M}_1}}{2}, -\frac{s+Q_{\mathcal{M}_1}}{2}\right)}. \tag{2.35}$$

We can further let hyper parameters a, b, s to be model specific, then the normalizing constants from the prior in (2.35) can not be canceled, i.e.,

$$\Omega^{\mathrm{CH}}_{\mathcal{M}_2:\mathcal{M}_1} = \frac{B\left(\frac{b_{\mathcal{M}_2}+p_{\mathcal{M}_2}}{2}, \frac{a_{\mathcal{M}_2}}{2}\right) \ {}_1F_1\left(\frac{b_{\mathcal{M}_2}+p_{\mathcal{M}_2}}{2}, \frac{a_{\mathcal{M}_2}+b_{\mathcal{M}_2}+p_{\mathcal{M}_2}}{2}, -\frac{s_{\mathcal{M}_2}+Q_{\mathcal{M}_2}}{2}\right)}{B\left(\frac{b_{\mathcal{M}_1}+p_{\mathcal{M}_1}}{2}, \frac{a_{\mathcal{M}_1}}{2}\right) \ {}_1F_1\left(\frac{b_{\mathcal{M}_1}+p_{\mathcal{M}_1}}{2}, \frac{a_{\mathcal{M}_1}+b_{\mathcal{M}_1}+p_{\mathcal{M}_1}}{2}, -\frac{s_{\mathcal{M}_1}+Q_{\mathcal{M}_1}}{2}\right)} \tag{2.36}$$

$$\cdot \frac{B\left(\frac{b_{\mathcal{M}_1}}{2}, \frac{a_{\mathcal{M}_1}}{2}\right) \ {}_1F_1\left(\frac{b_{\mathcal{M}_1}}{2}, \frac{a_{\mathcal{M}_1}+b_{\mathcal{M}_1}}{2}, -\frac{s_{\mathcal{M}_1}}{2}\right)}{B\left(\frac{b_{\mathcal{M}_2}}{2}, \frac{a_{\mathcal{M}_2}}{2}\right) \ {}_1F_1\left(\frac{b_{\mathcal{M}_2}}{2}, \frac{a_{\mathcal{M}_2}+b_{\mathcal{M}_2}}{2}, -\frac{s_{\mathcal{M}_2}}{2}\right)}. \tag{2.37}$$

2.3.3 Connection with the Literature

Note that the density function of the Confluent Hypergeometric distribution (2.29) is proportional to the densities of both Beta distribution and truncated Gamma distribution, which implies that our $\text{CH}\big(\frac{b}{2}, \frac{a}{2}, \frac{s}{2}\big)$ prior on $u = 1/(1+g)$ encompasses some of the exsiting prior distributions on the hyper parameter g.

In normal linear models, to achieve marginal likelihoods in closed forms, prior distributions on u originating from the Beta distribution are conventional. For example, Liang et al. (2008) introduces the Hyper-g prior,

$$\frac{1}{1+g} \sim \text{Beta}\left(\frac{a_h}{2} - 1, 1\right), \tag{2.38}$$

where $2 < a_h \leqslant 4$. When $a_h = 4$, the Hyper-g prior is equal to the uniform prior. The recommended value of the hyper parameter $a_h = 3$ corresponds to a proper prior which puts more mass of $1/(1+g)$ near 0. The choice $a_h = 2$ corresponds to both the reference prior and the Jeffrey's prior, which is improper. While it yields proper posterior distributions, because g does not appear in the model with just $\mathbf{X}_0$, Bayes factors are ill-determined due to the arbitrariness of the constants of proportionality. The Hyper-g prior (2.38) can be viewed as a special case of our CH-g prior, with $a = 2, b = a_h - 2$ and $s = 0$.

The marginal likelihoods under the Hyper-g prior in normal linear models have closed forms that contain the Hypergeometric ${}_2F_1$ function. To further simplify the marginal likelihood, Maruyama and George (2011) proposes the Beta prior distribution on g,

$$\frac{1}{1+g} \sim \text{Beta}\left(\frac{1}{4}, \frac{n - p_{\mathcal{M}}}{2} - \frac{3}{4}\right), \tag{2.39}$$

which eliminates the need to evaluate the ${}_2F_1$ function in the marginal likelihood. An additional benefit is the fact that the second parameter being proportional with n

yields an implicit $O(n)$ choice on g. According to the authors, $g = O(n)$ in the prior is desirable since it prevents the prior variance on $\boldsymbol{\beta}_{\mathcal{M}}$ from decreasing to zero and prevents the likelihood from being dominated by g asymptotically. The Beta prior (2.39) is also a special case of the CH-g prior, with $a = n - p_{\mathcal{M}} - 1.5$ and $b = 0.5$.

In GLMs, when the precision $a(\phi)$ is fixed, the likelihood under Laplace approximate usually contains an exponential term of u, for example, (2.18). Hence conjugate prior densities of u should contain some form of Gamma distribution density. For example, Wang and George (2007) proposes the truncated Gamma prior on u,

$$\frac{1}{1+g} \sim \text{Gamma}_{(0,1)}\left(a_t, b_t\right), \tag{2.40}$$

where the domain is restricted to the interval $(0, 1)$, and $a_t > 0, b_t > 0$. The authors recommend to use a uniform prior on g, which can be achieved by setting $a_t = 1, b_t = 0$. The CH-g prior also encompasses (2.40), with $a = 2, b = 2a_t$ and $s = 2b_t$.

Although our CH-g prior encompasses the above prior distributions on g, it does not include the Hyper-g/n prior (Liang et al., 2008) and the Robust prior (Bayarri et al., 2012). Since the Hyper-g prior cannot yield consistency for model selection when the null model is true, Liang et al. (2008) modify it to the Hyper-g/n prior,

$$p(g) = \frac{a_h - 2}{2n}\left(\frac{1}{1+g/n}\right)^{a_h/2}, \tag{2.41}$$

where $2 < a_h \leqslant 4$.

Under the Robust prior (Bayarri et al., 2012), after transforming the parameter g to $u = 1/(1+g)$, we find that its prior density becomes

$$p_r(u) = a_r\left[\rho_r(b_r+n)\right]^{a_r}\ \frac{u^{a_r-1}}{\left[1+(b_r-1)u\right]^{a_r+1}}\ \mathbf{1}_{\left\{0<u<\frac{1}{\rho_r(b_r+n)+(1-b_r)}\right\}}, \tag{2.42}$$

where $a_r > 0, b_r > 0$ and $\rho_r \geqslant \frac{b_r}{b_r+n}$. In normal linear models, the Robust prior yields closed-form marginal likelihoods in the form of the Appell F_1 function. Based on the various criteria for model selection priors proposed in Bayarri et al. (2012), the recommended values of the hyper parameters in the Robust prior are $a_r = 0.5, b_r = 1$ and $\rho_r = 1/(1 + p_{\mathcal{M}})$. Both the Hyper-$g$ prior and the Hyper-g/n prior are special cases of the Robust prior. More specifically, (2.42) with $a_r = a_h/2 - 1, b_r = 1, \rho_r = 1/(1+n)$ becomes the Hyper-g prior and (2.42) with $a_r = a_h/2 - 1, b_r = n, \rho_r = 0.5$ corresponds to the Hyper-g/n prior. The CH-g prior cannot be obtained as a special case of the Robust prior.

2.3.4 A More General Class of Prior Distributions on g

Both the Robust prior and the CH-g prior are special cases of a more general class of distributions, the Compound Confluent Hypergeometric (CCH) distribution (Gordy, 1998b). The CCH distribution has 6 parameters and can be considered as a generalized version of the Confluent Hypergeometric distribution. Suppose variable u has CCH distribution, then its density function is

$$p_{CCH}(u \mid t, q, r, s, v, \theta) \tag{2.43}$$

$$= \frac{v^t \exp(s/v)}{B(p,q)\ \Phi_1(q, r, t+q, s/v, 1-\theta)}\ \frac{u^{t-1}(1-vu)^{q-1}e^{-su}}{[\theta + (1-\theta)vu]^r}\ \mathbf{1}_{\{0<u<\frac{1}{v}\}}, \tag{2.44}$$

where $t > 0, q > 0, r \in \mathbb{R}, s \in \mathbb{R}, 0 \leqslant v \leqslant 1, \theta > 0$, and

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{(\alpha)_{m+n}(\beta)_n}{(\gamma)_{m+n} m! n!} x^m y^n$$

is the confluent hypergeometric function of two variables (Gordy, 1998b). We can extend the possible domain of the CCH distribution to $(0, 1/v)$ with $v > 1$, so that the upper bound of u can be strictly below 1. The extended CCH distribution as a prior distribution on $u = 1/(1+g)$ unifies a broader variety of prior distributions including both the Robust prior and the CH-g prior. The Robust prior is equal to

$$u \sim \text{CCH}\left(a_r, 1, a_r + 1, 0, \rho_r(b_r + n) + (1 - b_r), 1 + \frac{1 - b_r}{\rho_r(b_r + n)}\right)$$

and the CH-g prior is equal to

$$u \sim \text{CCH}\left(\frac{b}{2}, \frac{a}{2}, 0, -\frac{s}{2}, 1, 1\right).$$

In GLMs, under the extended CCH prior on u, the approximate marginal likelihood also has a closed form.

$$\begin{aligned}
p(\mathbf{Y} \mid \mathcal{M}) &= \int_0^1 p(\mathbf{Y} \mid u, \mathcal{M}) p(u) du \\
&= f_{\mathcal{M}}\left(\mathbf{Y} | \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) \left[1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc\right]^{-\frac{1}{2}} e^{-\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) \hat{\alpha}_{\mathcal{M}}^2}{2(\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc + 1)}} \\
&\quad \cdot \frac{B\left(\frac{2t+p_{\mathcal{M}}}{2}, q\right) \; v^{-\frac{p_{\mathcal{M}}}{2}} e^{-\frac{Q_{\mathcal{M}}}{2v}} \Phi_1\left(q, r, \frac{2t+2q+p_{\mathcal{M}}}{2}, \frac{2s+Q_{\mathcal{M}}}{2v}, 1-\theta\right)}{B(t, q) \; \Phi_1\left(q, r, t+q, \frac{s}{v}, 1-\theta\right)} + O(n^{-1}).
\end{aligned}$$

Under the Robust prior with $b = 1$, the Φ_1 function in $p(\mathbf{Y} \mid \mathcal{M})$ degenerates to a truncated Gamma function, which is easier to compute; that is

$$\begin{aligned}
&p(\mathbf{Y} \mid \mathcal{M}) \\
&= f_{\mathcal{M}}\left(\mathbf{Y} | \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}\right) \left[1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc\right]^{-\frac{1}{2}} e^{-\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) \hat{\alpha}_{\mathcal{M}}^2}{2(\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc + 1)}} a_r \left[\rho_r(1+n)\right]^{a_r} \\
&\quad \cdot \left(\frac{Q_{\mathcal{M}}}{2}\right)^{-\frac{p_{\mathcal{M}}}{2} - a_r} \left\{\Gamma\left(\frac{p_{\mathcal{M}}}{2} + a_r\right) - \Gamma\left(\frac{p_{\mathcal{M}}}{2} + a_r, \frac{Q_{\mathcal{M}}}{2\rho_r(1+n)}\right)\right\} + O(n^{-1}),
\end{aligned}$$

where $\Gamma(a)$ is the Gamma function and $\Gamma(a, s) \equiv \int_s^\infty t^{a-1} e^{-t} dt$ is the incomplete Gamma function.

2.4 Model Selection Consistency

In this section, we will focus on the asymptotic model selection performance of the CH-g prior for GLMs. In addition, we also study its behavior in a special but not rare case, where the sample size is small and there exists a model that fits the data perfectly.

2.4.1 Asymptotic Consistency for Model Selection

When studying the asymptotic properties, we believe it is reasonable to assume that the unit expected information matrices are non-singular.

Assumption 1. *We here assume a mild condition on the predictors* $[\mathbf{1}_n, \mathbf{X}]$ *in the full model. For any* n*-dim vector* $\boldsymbol{\eta}$ *in the space spanning by the predictors* $\mathcal{C}(\mathbf{1}_n, \mathbf{X})$*, i.e.,* $\boldsymbol{\eta} = [\mathbf{1}_n, \mathbf{X}]\, \mathbf{w}$*, where* $\mathbf{w}$ *is a* $(1+p)$*-dim vector of weights, there exists a positive definite matrix* $\boldsymbol{\Sigma}_{\mathbf{w}}$ *such that*

$$\lim_{n\to\infty} \frac{[\mathbf{1}_n, \mathbf{X}]^T \mathcal{I}_n(\boldsymbol{\eta})\, [\mathbf{1}_n, \mathbf{X}]}{n} \longrightarrow \boldsymbol{\Sigma}_{\mathbf{w}} \tag{2.45}$$

In normal linear model, this assumption implies that $\mathbf{X}^T\mathbf{X}/n$ converges to a positive definite matrix $\boldsymbol{\Sigma}_0$, which is a conventional assumption in the model selection literature. Furthermore, if we treat the rows of the full design matrix $\mathbf{X}$ as independent random samples from p-dimensional multivariate distributions which have the same mean and bounded covariance, then (2.45) holds according to the Law of Large Numbers.

Remark 2. *Before studying the asymptotic consistency of the CH-g prior, we want to point out that most asymptotic results which require i.i.d. samples as their conditions also hold under GLMs. Although in GLMs, observations* $Y_1, \ldots, Y_n$ *are conditionally independently but not identically distributed, we can assume that jointly* $(Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)$ *are i.i.d random samples. Thus as long as the marginal distribution of* $\mathbf{x}$ *does not depend on the GLM parameters, the log-likelihood and the score functions do not depend on the marginal distribution of* $\mathbf{x}$*. Hence the asymptotic results related to the MLE and likelihood ratio test also hold here. This underlying assumption is adopted by van der Vaart (2000) (Ch5) when applying MLE consistency in regression examples.*

According to Self and Mauritsen (1988), within the framework of GLM, for every model $\mathcal{M}$, the MLE of its parameters $(\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})$ converges to $(\alpha^*_{\mathcal{M}}, \boldsymbol{\beta}^*_{\mathcal{M}})$ in probability as n increases, where limit $(\alpha^*_{\mathcal{M}}, \boldsymbol{\beta}^*_{\mathcal{M}})$ are the maximizers of the limit of the log-likelihood,

$$(\alpha^*_{\mathcal{M}}, \boldsymbol{\beta}^*_{\mathcal{M}}) = \text{argmax}_{(\alpha,\boldsymbol{\beta})} \lim_{n\to\infty} \frac{1}{n} \log f_{\mathcal{M}}(\mathbf{Y}_n \mid \alpha, \boldsymbol{\beta}).$$

In particular, in the true model $\mathcal{M}_T$, $(\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T})$ are true parameters which generate the data.

We consider the same model selection consistency criteria discussed by Fernandez et al. (2001), Liang et al. (2008) and Bayarri et al. (2012).

Definition 1 (consistency for model selection)**.** *Suppose the true model that generates the data is among the 2^p potential models, and we denote it as $\mathcal{M}_T$. We say that the Bayes rule under the 0-1 loss is consistent for model selection if*

$$plim_{n\to\infty}\ p(\mathcal{M}_T \mid \mathbf{Y}) = 1 \tag{2.46}$$

This means that for any model $\mathcal{M} \neq \mathcal{M}_T$, $\text{plim}_n\ p(\mathcal{M} \mid \mathbf{Y}) = 0$. Hence a sufficient condition for model selection consistency is that

$$\text{plim}_{n\to\infty}\ \text{BF}_{\mathcal{M}_T:\mathcal{M}} = \infty \tag{2.47}$$

for any $\mathcal{M} \neq \mathcal{M}_T$, assuming fixed prior odds. The counter example in Remark 1 shows that for any fixed g, the consistency for model selection do not hold under the fixed g-prior for GLMs, thus we focus on results under the CH-g prior. According to our previous decompositions of the Bayes factors (2.34), it is sufficient to examine the asymptotic properties of $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ and $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}}$.

We will show in the following lemma that the first term $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ (2.22) of the Bayes factor is dominated by the maximized likelihood ratio asymptotically. According to Self et al. (1992), under the alternative model, the log likelihood ratio between $\mathcal{M}_T$ and $\mathcal{M}$ converges in distribution to a non-central χ^2 distribution.

Lemma 1. *As the sample size n increases, the asymptotic property of*

$$\Lambda_{\mathcal{M}_T:\mathcal{M}} = \frac{f_{\mathcal{M}_T}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_T}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{f_{\mathcal{M}}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})} \left[\frac{1+\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T})nc}{1+\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})nc}\right]^{-\frac{1}{2}} e^{-\frac{1}{2}\left[\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T})\hat{\alpha}^2_{\mathcal{M}_T}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T})nc+1} - \frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})\hat{\alpha}^2_{\mathcal{M}}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})nc+1}\right]}$$

is that

1) if $\mathcal{M}_T \subset \mathcal{M}$, then $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O(1)$;

2) if $\mathcal{M}_T \not\subset \mathcal{M}$, then $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O\left(e^{cn}\right)$, where c is a positive constant.

In addition, under the flat prior $p(\alpha_{\mathcal{M}}) \propto 1$, these properties also hold for

$$\Lambda^{c=\infty}_{\mathcal{M}_T:\mathcal{M}} = \frac{f_{\mathcal{M}_T}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_T}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{f_{\mathcal{M}}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})} \left[\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T})}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})}\right]^{-\frac{1}{2}}$$

Proof: see Appendix A.2.3.

The first term $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ in the Bayes factors can be considered as a measure of goodness of fit. If the space spanned by the predictors of $\mathcal{M}$ does not contain all predictors in the true model $\mathcal{M}_T$, $\mathcal{M}$ cannot predict as well as $\mathcal{M}_T$. Therefore, the term $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ overwhelmingly favors $\mathcal{M}_T$ by increasing at an exponential rate of n. On the other hand, when the design space of $\mathcal{M}$ contains all predictors in $\mathcal{M}_T$, $\mathcal{M}$ has the same ability in explaining the response as $\mathcal{M}_T$. Therefore, $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ alone does not favor selecting $\mathcal{M}_T$ against a redundant model $\mathcal{M}$. In this case, the second term in the Bayes factors, (2.23) or (2.35), plays a more important role of placing more penalty on the redundant model. In the case of fixed g, the term $(1+g)^{(p_{\mathcal{M}}-p_{\mathcal{M}_T})/2}$ in $\Omega_{\mathcal{M}_T:\mathcal{M}}$ penalizes $\mathcal{M}$ for the extra dimensions. However, the counter example in Remark 1 illustrates that with fixed g, the penalty being imposed on the redundant model is not strong enough asymptotically. Next, we will focus on the CH-g prior by exploring the asymptotic properties of $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}}$, which yields a stronger penalty on the model size. We first study the asymptotic behavior of the analogue of regression sum of squares (RSS) for GLMs, $Q_{\mathcal{M}_T}$ and $Q_{\mathcal{M}}$ in the following lemma.

Lemma 2. *Let $\boldsymbol{\beta}^*_{\mathcal{M}}$ denote the limit of the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$. The asymptotic properties of*

$$Q_{\mathcal{M}} = \left[\hat{\boldsymbol{\beta}}^T_{\mathcal{M}}\mathbf{X}^T_{\mathcal{M}}\right]\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\left[\mathbf{X}_{\mathcal{M}}\hat{\boldsymbol{\beta}}_{\mathcal{M}}\right]$$

under the true model $\mathcal{M}_T$ and under any other model $\mathcal{M}$ are

1) If $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$, then $Q_{\mathcal{M}_T} = O(n)$; for any other model $\mathcal{M}$, if $\boldsymbol{\beta}^*_{\mathcal{M}} \neq \mathbf{0}$, then $Q_{\mathcal{M}} = O(n)$, and otherwise, $Q_{\mathcal{M}} = O(1)$.

2) If $\mathcal{M}_T = \mathcal{M}_{\emptyset}$, then for any model $\mathcal{M}$, $Q_{\mathcal{M}} = O(1)$; and by the definition of Q under the null model, $Q_{\mathcal{M}_T} = O(1)$.

Proof: see Appendix A.2.4

Theorem 1. *With fixed hyper parameters $a, b > 0$ and $s \geqslant 0$, the CH-g prior* (2.28) *is consistent for model selection* (2.47)*, except for $\mathcal{M}_T = \mathcal{M}_{\emptyset}$. In addition, this result also holds with model specific hyper parameters $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$ that are independent of n.*

Proof: see Appendix A.2.5

Theorem 1 implies that the CH-g prior is desirable as a model selection prior in most cases. However, it fails to impose a strong enough penalty in the case where the null model is true. To resolve this inconsistency, we allow the hyper parameter a to increase with n, such that

$$\lim_{n\to\infty} \frac{a}{n} = a^*, \text{ where } a^* \geqslant 0. \tag{2.48}$$

The following theorem shows that when $a^* > 0$, the selection consistency holds under any $\mathcal{M}_T$, including when $\mathcal{M}_T = \mathcal{M}_{\emptyset}$.

Theorem 2. *With hyper parameters $b > 0, s \geqslant 0$, and $\lim_{n\to\infty} a/n = a^* > 0$, the CH-g prior* (2.28) *is consistent under model selection universally, including the case where $\mathcal{M}_T = \mathcal{M}_{\emptyset}$. In addition, this result also holds with model specific hyper parameters $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$, where $\lim_{n\to\infty} a_{\mathcal{M}}/n = a^*_{\mathcal{M}} > 0$.*

Proof: see Appendix A.2.6

2.4.2 Perfect Fitting with Small Sample

We have demonstrated that the CH-g prior for GLMs is consistent for model selection with large n. Now we will explore its selection performance with small samples, in the case of perfect fitting.

In linear regression, Liang et al. (2008) points out that under the Zellner's g-prior with any fixed value of g, there exists the following information paradox. In principle, for $n \gg p_{\mathcal{M}} + 1$, if all the observations fall on a hyperplane ($R^2 = 1$), the Bayes factor should support model $\mathcal{M}$ overwhelmingly over the null model $\mathcal{M}_{\emptyset}$. However, with any fixed g, the $\text{BF}_{\mathcal{M}:\mathcal{M}_{\emptyset}}$ under the g-prior is bounded. To resolve this information paradox, the parameter g should be assigned certain hyper prior distributions in fully Bayes approach, or be estimated by empirical Bayes approach.

Bayarri et al. (2012) provide a formal definition of the information consistency for priors in model selection. If there exists a sequence of datasets with the same size n such that the ratio of maximized likelihoods between $\mathcal{M}$ and $\mathcal{M}_{\emptyset}$ go to infinity, then their Bayes factors should also go to infinity. The condition of this criteria describes the perfect fitting phenomenon under model $\mathcal{M}$ of a diverging likelihood ratio, which is precise in linear regression since the estimate for the normal variance in $\mathcal{M}$ is equal to zero, i.e., $\hat{\sigma}^2 = 0$. However, this form of perfect fitting is not necessarily true with most GLMs, including logistic regression and Poisson regression. Because the response variable is discrete, the maximum likelihood under $\mathcal{M}$ has an upper bound being 1, and no matter how minimal the amount of information the null model $\mathcal{M}_{\emptyset}$ can reveal, its maximum likelihood is always greater than zero. Hence for discrete distributions in the exponential family, even the GLMs can fit the data perfectly, their likelihood ratios are bounded. For example, in logistic regression, suppose model $\mathcal{M}$ fits each binary response perfectly, i.e., $\hat{\mu}_i = 1$ if $Y_i = 1$, and $\hat{\mu}_i = 0$ if $Y_i = 0$; while the estimates of success probability in $\mathcal{M}_{\emptyset}$ are $\hat{\mu}_i = 0.5$. The likelihood ratio

$$\frac{f_{\mathcal{M}}(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})}{f_{\mathcal{M}_{\emptyset}}(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}_{\emptyset}})} = 4^n < \infty.$$

We propose to use the fitted variance of all responses being zero to quantify the perfect fitting phenomenon, that is, under model $\mathcal{M}$,

$$\widehat{\mathbb{V}(Y_i)} = 0, \quad i = 1, \ldots, n, \tag{2.49}$$

which is equivalent to $f_{\mathcal{M}}(\mathbf{Y} \mid \hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}}) = \infty$ in normal distribution. While in the Bernoulli distribution, although the likelihood function is bounded, our criterion (2.49) precisely describes the perfect fitting phenomenon. When the fitted values of the expectation of every binary response $\widehat{\mathbb{E}(Y_i)}$ equals to 0 or 1, the fitted values of the variances are zero.

Another interesting difference we find between normal linear regression and GLMs is whether to favor $\mathcal{M}$ over the null if perfect fitting occurs, but the sample size is relatively small. In normal linear regression, as Liang et al. (2008) and Bayarri et al. (2012) suggest, perfect fitting with any $n \geqslant p_{\mathcal{M}} + 2$ is strong evidence to favor $\mathcal{M}$. However with discrete responses, especially binary ones, perfect fitting is likely to occur by chance when n is just slightly larger than $p_{\mathcal{M}} + 1$. In this case, the Bayes factor should not overwhelmingly support $\mathcal{M}$ over $\mathcal{M}_{\emptyset}$, unless n is large enough. This problem is worth noticing because it is not rare in real world applications where p is close to n, such as genetic studies. In logistic regression or Probit regression, the expected information of the i-th linear combination

$$\mathcal{I}_n(\hat{\eta}_i) = \frac{e^{\hat{\eta}_i}}{(1+e^{\hat{\eta}_i})^2} \text{ or } \frac{\phi(\hat{\eta}_i)^2}{(1-\Phi(\hat{\eta}_i))\Phi(\hat{\eta}_i)}$$

converges to zero as $\hat{\eta}_i = \pm\infty$, where $\phi(\cdot), \Phi(\cdot)$ are pdf and cdf of the standard normal distribution. If perfect fitting occurs under model $\mathcal{M}$, then $\mathcal{I}_n(\alpha_{\mathcal{M}}) = 0$ and $Q_{\mathcal{M}} = 0$. According to the Laplace approximation of marginal likelihoods, $\Lambda^{c=\infty}_{\mathcal{M}:\mathcal{M}_\varnothing} = \infty$ (2.25). Since both $\Omega_{\mathcal{M}:\mathcal{M}_\varnothing}$ (2.23) and $\Omega^{\text{CH}}_{\mathcal{M}:\mathcal{M}_\varnothing}$ (2.35) are bounded, $\text{BF}_{\mathcal{M}:\mathcal{M}_\varnothing}$ diverges under both g-prior and CH-g prior. In contrast, under the normal prior on the intercept, because $\Lambda_{\mathcal{M}:\mathcal{M}_\varnothing}$ is bounded, this problem is resolved. In Section 2.2.3, we recommend using a proper prior (2.17) on the intercept instead of the commonly used improper flat prior, to avoid inconsistency with perfect fitting with small n, where the Bayes factor overwhelmingly supports the larger model.

On the other hand, if perfect fitting occurs under model $\mathcal{M}$ with sufficiently large samples, it is reasonable to let the $\text{BF}_{\mathcal{M}:\mathcal{M}_\varnothing}$ go to infinity. We set the prior variance of $\alpha_{\mathcal{M}}$ proportional to n, so that the normal prior converges to flat prior as n increase which indicates model $\mathcal{M}$ is overwhelmingly favored if it can fit a large sample of responses perfectly and the estimate of $\alpha_{\mathcal{M}}$ is consistent.

2.5 BMA Estimation Consistency

2.5.1 Asymptotic Posterior Estimates

In each model $\mathcal{M} \neq \mathcal{M}_\varnothing$, the conditional posterior mean

$$\mathbb{E}[\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, g, \mathcal{M}] = \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\mathcal{M}}$$

does not converge to the limit of the MLE asymptotically with any fixed g. In a fully Bayes approach, convergence of the posterior distribution of the shrinkage factor $z = g/(1+g)$ to 1 is a necessary condition for the approximate conditional posterior

$z\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ being consistent. Before examining the parameter estimation consistency of the coefficients, we first study the asymptotic behavior of the conditional posterior $p(z \mid \mathbf{Y}, \mathcal{M})$. in the following propositions. Since these results are studied under each model $\mathcal{M}$, they remain true if we allow the hyper parameters $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$ to be model specific. For notation simplicity, we omit their subscript $\mathcal{M}$ here.

Proposition 3. *For the CH-g prior with hyper parameters $a > 0, b > 0, s \geqslant 0$, if the MLE of its coefficient converges to a non-zero vector $\hat{\boldsymbol{\beta}}_{\mathcal{M}} \to \boldsymbol{\beta}^*_{\mathcal{M}} \neq \mathbf{0}$, then the conditional posterior distribution of $z = g/(1+g)$ under any model $\mathcal{M} \neq \mathcal{M}_{\o}$, converges to* 1 *in probability*

$$plim_{n\to\infty}\, p\left(z \mid \mathbf{Y}, \mathcal{M}\right) = \delta_1(z) \tag{2.50}$$

In particular, if the true model is not null $\mathcal{M}_T \neq \mathcal{M}_{\o}$, (2.50) *holds under $\mathcal{M}_T$.*

Proof: see Appendix A.2.7

Proposition 4. *For the CH-g prior with hyper parameters $b > 0, s \geqslant 0$, and $\lim_{n\to\infty} a/n = a^* > 0$, for any true model $\mathcal{M}_T$ including $\mathcal{M}_T = \mathcal{M}_{\o}$, the conditional posterior distribution of the shrinkage factor $z = g/(1+g)$ under any model $\mathcal{M} \neq \mathcal{M}_{\o}$ converges to* 1 *in probability, i.e.,* (2.50).

Proof: see Appendix A.2.8

2.5.2 Parameter Estimation under BMA

Bayesian model averaging (BMA) estimates are widely used to incorporate model uncertainty. We denote the variable $\boldsymbol{\beta}$ as the p dimensional vector of coefficients corresponding to all the potential predictors. In this section, we slightly abuse the notations $\boldsymbol{\beta}_{\mathcal{M}}$ by redefining that as a p-dimensional vectors filled with zeros for variables not included in the model such that $\boldsymbol{\eta}_{\mathcal{M}} = \mathbf{X}_0\boldsymbol{\alpha}_0 + \mathbf{V}\boldsymbol{\beta}_{\mathcal{M}}$. The posterior distribution of $\boldsymbol{\beta}$ under BMA is

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = \sum_{\mathcal{M}} p(\mathcal{M} \mid \mathbf{Y})\ p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}), \tag{2.51}$$

where the marginal posterior distribution under model $\mathcal{M}$ can be calculated as

$$p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}) = \int p\left(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, g, \mathcal{M}\right) p\left(\frac{g}{1+g} \mid \mathbf{Y}, \mathcal{M}\right) d\frac{g}{1+g}. \tag{2.52}$$

To study the parameter estimation performance of the BMA estimates asymptotically, we propose the following estimation consistency.

Definition 2 (consistency for parameter estimation)**.** *The parameter estimation under BMA is consistent if the posterior of $\boldsymbol{\beta}$ converges to the true parameter in probability as n increases, i.e.*

$$plim_{n\to\infty}\ p(\boldsymbol{\beta} \mid \mathbf{Y}) = \delta_{\boldsymbol{\beta}^*_{\mathcal{M}_T}}(\boldsymbol{\beta}). \tag{2.53}$$

Under BMA, the posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{Y})$ can be decomposed as a weighted average of the posterior under $\mathcal{M}_T$ and under other models,

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = p(\mathcal{M}_T \mid \mathbf{Y})\ p(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T) + \sum_{\mathcal{M} \neq \mathcal{M}_T} p(\mathcal{M} \mid \mathbf{Y})\ p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}), \tag{2.54}$$

To verify the BMA estimation consistency under the CH-g prior, we can use the results on model selection consistency in Section 2.4.1. When the selection consistency holds, i.e., $p(\mathcal{M}_T \mid \mathbf{Y})$ converges to 1, the second term in (2.54) diminishes in the limit. Hence in this case, we just need to focus on the posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}_T}$. On the other hand, when the selection consistency does not hold, which only occurs where $\mathcal{M}_T = \mathcal{M}_\emptyset$ and $a = O(1)$, we need to examine the limit distribution of $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M})$ under every $\mathcal{M}$. Fortunately, in this case the true parameter $\boldsymbol{\beta}^*_{\mathcal{M}_T} = \mathbf{0}$. Although shrinkage always exists, the limit of $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M})$ remains $\mathbf{0}$. Therefore, we have the following theorem.

Theorem 3. *With hyper parameters $b > 0, s \geqslant 0$, and $\lim_{n\to\infty} a/n = a^* \geqslant 0$, the CH-g prior* (2.28) *is consistent for parameter estimation of the coefficient under BMA* (2.53). *In addition, this result also holds with model specific hyper parameters $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$.*

Proof: see Appendix A.2.9

2.5.3 BMA Estimation for a New Case

In addition to the current data $\{\mathbf{Y}, \mathbf{V}\}$, if we have a new case and know the values of its exploratory variables $\mathbf{v} \in \mathbb{R}^p$, we want to estimate the mean of it response variable $\mu = \mathbb{E}(Y)$ under BMA. Under model $\mathcal{M}$, suppose $\mathbf{x}_{\mathcal{M}} = \mathbf{v}_{\mathcal{M}} - \tilde{\mathbf{v}}_{\mathcal{M}}$ is the vector of new predictors after the "centering" step, where $\tilde{\mathbf{v}}_{\mathcal{M}}$ is a $p_{\mathcal{M}}$-dim vector consists of $\tilde{v}_{j,\mathcal{M}}$ as in (2.11). The BMA estimate for the mean of the new response is

$$\begin{aligned}\mu &= \mathbb{E}(Y \mid \mathbf{v}, \mathbf{Y}, \mathbf{V}) \\ &= \sum_{\mathcal{M}} p(\mathcal{M} \mid \mathbf{Y})\ \mathbb{E}\left[b'(\alpha_{\mathcal{M}} + \mathbf{x}_{\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M})\right] \\ &= \sum_{\mathcal{M}} p(\mathcal{M} \mid \mathbf{Y}) \int b'(\alpha_{\mathcal{M}} + \mathbf{x}_{\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}}) p(\alpha_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}) p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M}) d(\alpha_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}})\end{aligned}$$

where the conditional posteriors of $\alpha_{\mathcal{M}}$ and $\boldsymbol{\beta}_{\mathcal{M}}$ are approximated by (2.26) and (2.52). Similarly to the prediction consistency criterion introduced in Liang et al. (2008), we define the estimation consistency under BMA for a new case for GLMs.

Definition 3. *We say that the BMA estimation μ for a new case $\mathbf{v}$ is consistent if*

$$plim_n\ \mu = b'(\alpha^*_{\mathcal{M}_T} + \mathbf{x}^T_{\mathcal{M}_T} \boldsymbol{\beta}^*_{\mathcal{M}_T}), \tag{2.55}$$

where $\mathcal{M}_T$ is the true model, $\mathbf{x}_{\mathcal{M}_T}$ is the sub-vector of the "centered" new exploratory variable corresponding to $\mathcal{M}_T$, and $\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T}$ are the true intercept and coefficients.

We again decompose the BMA estimate μ into the sum of two terms

$$\mu = p(\mathcal{M}_T \mid \mathbf{Y})\ \mathbb{E}\left[b'(\alpha_{\mathcal{M}_T} + \mathbf{x}_{\mathcal{M}_T}^T \boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T)\right] \\ + \sum_{\mathcal{M} \neq \mathcal{M}_T} p(\mathcal{M} \mid \mathbf{Y})\ \mathbb{E}\left[b'(\alpha_{\mathcal{M}} + \mathbf{x}_{\mathcal{M}}^T \boldsymbol{\beta}_{\mathcal{M}} \mid \mathbf{Y}, \mathcal{M})\right],$$

In the following theorem, we find that the BMA estimation consistency for a new case holds under the CH-g prior.

Theorem 4. *With hyper parameters $b > 0, s \geqslant 0$, and $\lim_{n\to\infty} a/n = a^* \geqslant 0$, the BMA estimation for a new case under the CH-g prior* (2.28) *is consistent. In addition, this result also holds with model specific hyper parameters $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$.*

Proof: see Appendix A.2.10

2.6 Simulation and Real Examples

In Section 2.3.3, we have established theoretical connections between our CH-g prior and some of the commonly used prior distributions on the hyper parameter g proposed for the g-prior, such as the uniform prior (Wang and George, 2007), the Hyper-g prior (Liang et al., 2008), the Beta prior (Maruyama and George, 2011) and the Robust prior (Bayarri et al., 2012). In this section, using both simulation studies and a real example, we will compare model selection and parameter estimation performance across these hyperpriors of g under our extension of the g-prior for GLMs (2.16), (2.17). In addition to the above-mentioned approaches, we also examine the Jeffrey's prior on g (Celeux et al., 2012), the local empirical Bayes (EB) (Hansen and Yu, 2001) method, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

In the local EB approach, the estimate of g under each model $\mathcal{M}$ is the maximizer of the marginal likelihood $p(\mathbf{Y} \mid g, \mathcal{M})$. Under the Laplace approximation (2.18),

$\hat{g}_{\mathcal{M}}^{\mathrm{EB}}$ is estimated as

$$\hat{g}_{\mathcal{M}}^{\mathrm{EB}} = \arg\max_{g} p(\mathbf{Y} \mid g, \mathcal{M}) = \max\left(\frac{Q_{\mathcal{M}}}{p_{\mathcal{M}}} - 1, 0\right), \tag{2.56}$$

and the marginal likelihood is obtained by plugging in this estimate $p_{\mathrm{EB}}(\mathbf{Y} \mid \mathcal{M}) = p(\mathbf{Y} \mid \hat{g}_{\mathcal{M}}^{\mathrm{EB}}, \mathcal{M})$.

Table 2.2: For GLMs: methods to be compared.

	a	b	s	comments
CH-g	$n/2$	0.5	0	
	2	2	0	Uniform prior
	2	1	0	Hyper-g
	2	0	0	Jeffrey's prior
	$n - p_\gamma - 1.5$	0.5	0	Beta (Maruyama and George, 2011)
Robust prior, $a_r = 0.5, b_r = 1, \rho_r = 1/(1 + p_{\mathcal{M}})$				
Local EB (Hansen and Yu, 2001)				
AIC				
BIC				

We summarize all these methods to be compared in Table 2.2. For AIC and BIC, we select the model with smallest AIC and BIC; while for all other methods, we select the model $\mathcal{M}$ with the highest posterior probability, i.e., maximum a posterior (MAP) estimate. In order to take into account the model uncertainty, for both fully Bayes and empirical Bayes methods, we use Bayesian model averaging (BMA) estimates for the parameter $\boldsymbol{\beta}$ and exceptions of new responses $\mu = \mathbb{E}(Y)$. While for AIC and BIC, these estimates are calculated only based on the selected model.

2.6.1 Default Choice of Hyper Parameters $\{a, b, s\}$ in the CH-g Prior

Before exhibiting the examples, we first give our recommendation on values of the hyper parameters a, b, s in the CH-g prior. In general, when a or s is large, or when b is small, the prior concentration of the shrinkage factor $z = g/(1 + g)$ is high near 1.

In this case, little shrinkage is imposed on the posterior estimates of $\boldsymbol{\beta}_{\mathcal{M}}$. Meanwhile, this corresponds to high prior concentration of large g, which implies a flat prior on $\boldsymbol{\beta}_{\mathcal{M}}$ that favors simple models in model selection, and hence is desirable for sparse problems.

We choose a to be proportional with the sample size n, to allow the CH-g prior to be consistent for model selection in all circumstances including when $\mathcal{M}_T = \mathcal{M}_{\o}$ (see Theorem 2). Some popular methods in linear regression such as Zellner and Siow (1980), the Hyper-g/n prior (Liang et al., 2008), the Beta prior (Maruyama and George, 2011) also recommend $g = O(n)$. Actually under the mild assumption (2.45), the expected information matrix based on all n sample points $\mathcal{I}_n(\boldsymbol{\beta}_{\mathcal{M}}) = O(n)$. This suggests that the g-prior on $\boldsymbol{\beta}_{\mathcal{M}}$ depends implicitly on n, and degenerates to a point mass at zero in the limit. Hence the choice of $g = O(n)$ is essential to avoid having the g-prior to dominate the likelihood. To eliminates the dependency of the prior distribution on specific features of model including the sample size n, Bayarri et al. (2012) proposes the intrinsic consistency of model selection priors, which suggests that as n increase, the prior distribution $p(\boldsymbol{\beta}_{\mathcal{M}} \mid \boldsymbol{\alpha}_{\mathcal{M}}, \mathcal{M})$ should be proper. In the context of g-prior (both for normal linear regression and our extension for GLMs), the intrinsic consistency means proper prior distribution on g/n in the limit. With $a = O(n)$, the CH-g prior yields an implicit $g = O(n)$ choice, in the sense that the prior expectation

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

To choose the default prior rate $a^* = a/n$, empirical experience indicates no significant difference in parameter estimation between $a^* = 0.5$ and 1. To remain objective and perform well in model selection under both sparse and non-sparse models, we recommend to use $a^* = 0.5$, i.e. $a = n/2$.

According to the approximate conditional posterior distribution of the shrinkage factor z (2.31), the parameter b in the CH-g prior is updated to $b + p_{\mathcal{M}}$ after incorporating the data. In addition, b controls the tail behavior of the prior distribution $p(\boldsymbol{\beta}_{\mathcal{M}})$. More specifically, $p(\boldsymbol{\beta}_{\mathcal{M}})$ has tails similar to a multivariate Student distribution with degrees of freedom b. Our choice $b = 0.5$ corresponds to a distribution with even heavier tails than Cauchy. Under this choice, the CH-g prior has vanishing prior influence on the estimation of $\boldsymbol{\beta}$, and thus is capable of preserving large signals. Maruyama and George (2011) also recommends a prior distribution with flatter tails than Cauchy as their default choice. Bayarri et al. (2012) recommends a prior with Cauchy tails but not heavier, because they think otherwise it strongly favors the smaller model, even with minimal sample size. Between the choices $b = 0.5$ and 1, the following simulation examples reveals no significant difference in BMA estimation.

The parameter s is updated by the data to $s + Q_{\mathcal{M}}$, and therefore serves as a prior RSS. We recommend $s = 0$, which implies no information or variation a priori. In addition, according to our empirical experience, when the parameter $a = O(n)$, different values of s yield no significant difference in both model selection and parameter estimation.

The parameter c is chosen according to the inverse variance of the response that has mean zero, i.e. $c = 1/\mathbb{V}(y_0)$ where $\mathbb{E}(y_0) = b'(\theta(\eta_0 = 0))$. For example, for both logistic regression and Probit regression, we let $c = 4$; while for Poisson regression, $c = 1$. Note that as n increases, since the prior variance of the intercept nc goes to infinity, i.e., choice of c hardly makes a difference with large samples.

2.6.2 Simulations: Logistic and Poisson Regressions

The logistic regression simulation study is based on the simulation example introduced in Hansen and Yu (2003), and the Poisson regression example is based on the one in Chen et al. (2008). To explain the output $\mathbf{Y}$, $p = 5$ potential predictors

$\mathbf{V}_1, \dots, \mathbf{V}_p$ are considered to be included in the logistic regression model, and $p = 3$ in the Poisson regression model. Each predictor is drawn from a standard normal distribution, with pairwise correlation

$$\text{cor}(\mathbf{V}_i, \mathbf{V}_j) = r^{|i-j|}, \quad 1 \leqslant i < j \leqslant p$$

Here we consider two cases: independent predictors ($r = 0$) and correlated predictors ($r = 0.75$). For each realization, $n = 100$ and 500 independent samples are generated for logistic regression and Poisson regression respectively, according to 4 scenarios of different sparsity of the true underlying models (see Table 2.3 for the intercepts and coefficients of the true models). For all Bayesian methods, we assign uniform prior distribution to the model space, i.e., $p(\mathcal{M}) = 1/2^p$. We repeat the simulation for $N = 100$ times, and compare their performance in model selection and parameter estimation.

Table 2.3: GLM simulation: four scenarios of true models that generate the simulation data, each represented by the true values of intercept and coefficients $(\alpha *, \boldsymbol{\beta}^*)$.

scenario	logistic regression	Poisson regression
null	(0 0 0 0 0 0)	(-0.3 0 0 0)
sparse	(0 2 0 0 0 0)	(-0.3 0.3 0 0)
medium	(0 3 2 2 0 0)	(-0.3 0.3 0.2 0)
full	(0 5 1 1 1 1)	(-0.3 0.3 0.2 -0.15)

To access the performance of the MAP estimates in model selection, we examine the their selection accuracy under the 0-1 loss, by reporting the number of times the correct underlying models being selected in Table 2.4 and 2.5. The results of both logistic regression and Poisson regression yield similar trend in comparison across all methods. In general, we find that the nine methods being compared can be roughly divided into two groups. The CH-g prior, the Beta prior, the Robust prior and BIC form the first group, since all of them prefer parsimonious models and hence outperform the rest of methods when the true model is sparse, or more extremely, the null model. In contrast, the second group consists of the uniform prior, the Hyper-g, Jeffrey's prior, EB and AIC, all of which prefer complex models and yield more accurate selection when the true model is the full model. The fact that AIC favors larger models and BIC smaller models is well studied. Since the model complexity penalty in the marginal likelihood under EB depends on the model fitting $Q_{\mathcal{M}}$, the EB tends to favor large models. Among the fully Bayes methods, the different preference in model complexity is mainly contributed by the different prior concentration of g. Large g corresponds to preference of small models. Since methods in the first group (except the BIC) indicate $g = O(n)$ a priori, they achieve model consistency, even when the true model is the null. However, the Bayesian approaches in the second group such as the Hyper-g perform poorly in this case, which confirms the theoretical results in Liang et al. (2008). On the other hand, in reality the information about the underlying true model is usually unavailable, good selection method should be able to adapt to a wide spectrum of sparsity. Among the methods in the first group, the CH-g prior performs the most accurately in model selection when data are generated from the full model.

To evaluate the estimation performance, we report the median $\text{SSE}(\boldsymbol{\beta}) = \sum_{j=1}^{p}(\tilde{\beta}_j - \beta^*_{j,\mathcal{M}_T})^2$ in Table 2.6 and 2.7. Here $\tilde{\beta}_j$ represents either the BMA estimates of the j-th coefficient for all mixtures of g-prior methods, or the MLE of it under the selected model by AIC and BIC; and $\beta^*_{j,\mathcal{M}_T}$ is the value corresponding coefficient in the true model that generates the data. In particular, $\beta^*_{j,\mathcal{M}_T} = 0$ if the j-th predictor is excluded in the true model. An overall trend of parameter estimation accuracy among these methods is that the models perform better in model selection also yield smaller estimation error. Note that the CH-g prior outperforms most methods in the second group except EB where the true model is the full model in logistic regres-

Table 2.4: Logistic regression: model selection accuracy under 0-1 loss. Number of times the true model are selected out of 100. Column-wise largest value is in bold type.

scenario	r	CH-g(50, 0.5, 0)	Uniform	Hyper-g	Jeffreys	Beta	Robust	EB	AIC	BIC
null	0	94	33	50	0	**95**	93	0	40	86
	0.75	93	36	55	0	**94**	93	0	48	90
sparse	0	83	52	59	72	**86**	75	60	46	**86**
	0.75	83	54	64	74	**84**	79	65	52	**84**
medium	0	72	43	51	58	78	77	51	55	**89**
	0.75	48	44	49	50	46	50	48	**57**	41
full	0	38	**69**	67	61	30	28	67	51	15
	0.75	1	**17**	13	9	0	0	15	1	0

sion. Furthermore, to evaluate their performance in estimating new cases, we use the $(i+1)$-th dataset as the test set for the model studies by the i-th dataset, where $i = 1, \dots, N$. We examine the median SSE loss in the expectation of the response $\sum_{i=1}^{n}(\hat{\mu}_i - \mu_{i,\mathcal{M}_T})^2$, which we omit here since it shows vey similar pattern to SSE($\boldsymbol{\beta}$). Furthermore, we also examine the out-of-sample classification error for logistic regression, which we also omit here since it reveals almost no difference across methods for this example.

2.6.3 Pima Indians Diabetes Data

We apply the CH-g prior to a real-world problem, the Pima Indians diabetes data, along with other state of the art approaches that being compared with in Section 2.6.2. The dataset is previously studied using Bayesian model selection approaches in Bové and Held (2011). It consists of $n = 532$ independent patients' records, and contains information including a binary response of diabetes signs y, and $p = 7$

Table 2.5: Poisson regression: model selection accuracy under 0-1 loss. Number of times the true model are selected out of 100. Column-wise largest value is in bold type.

scenario	r	CH-g(250, 0.5, 0)	Uniform	Hyper-g	Jeffreys	Beta	Robust	EB	AIC	BIC
null	0.00	97	53	64	0	**99**	96	1	55	96
	0.75	99	59	76	0	**100**	99	1	61	97
sparse	0.00	**95**	81	86	90	**95**	94	86	71	**95**
	0.75	**97**	86	89	94	**97**	**97**	89	78	**97**
medium	0.00	86	84	88	86	**89**	88	88	81	**89**
	0.75	53	**65**	61	56	47	54	61	70	49
full	0.00	72	**90**	88	87	68	71	88	**97**	63
	0.75	17	**46**	41	34	12	18	41	**61**	12

potential explanatory variables $\{X_1, \ldots, X_7\}$ such as number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skin fold thickness, BMI, diabetes pedigree function and age. To account for multiplicity adjustment, instead of uniform prior on the model space, we use the Beta-Binomial$(1, 1)$ prior as suggested by Scott and Berger (2010), i.e., $p(\mathcal{M}) = \frac{1}{p}\binom{p}{p_{\mathcal{M}}}^{-1}$. We enumerate all 2^p possible models. In Table 2.8, we show the marginal posterior inclusion probability for each predictor $p(\beta_j \neq 0 \mid \mathbf{Y})$ for $j = 1, \ldots, p$. For the two information criteria methods, similarly as in Bové and Held (2011), we use $e^{-\mathrm{AIC}/2}$ and $e^{-\mathrm{BIC}/2}$ in the place of the approximate marginal likelihood and average the posterior marginal inclusion over all 2^p models under the same prior of the model space.

The marginal posterior inclusion probabilities provide us with the knowledge whether each predictor has significant impact on predicting the binary response. Comparing different methods, we notice the same trend in overall selection perfor-

Table 2.6: Logistic regression: median SSE of $\boldsymbol{\beta}$: $\sum_{j=1}^{p}(\hat{\beta}_i - \beta_i^*)^2 \times 10$ based on 100 realizations. Column-wise smallest value is in bold type. Friedman test shows that (1) CH-g(50, 0.5, 0) is significantly different from the Beta-prime in all scenarios, and (2) CH-g(50, 0.5, 0) is significantly different from the robust prior except for Row 4.

scenario	r	CH-g(50, 0.5, 0)	Uniform	Hyper-g	Jeffreys	Beta-prime	robust	EB	AIC	BIC
null	0	0.08	0.08	0.10	0.17	0.04	0.09	0.01	1.04	**0.00**
	0.75	0.08	0.11	0.15	0.24	0.04	0.09	0.01	0.96	**0.00**
sparse	0	1.70	2.37	2.10	1.93	1.55	1.67	2.13	3.06	**1.18**
	0.75	1.97	3.17	2.67	2.37	1.78	1.94	2.67	4.22	**1.06**
medium	0	8.21	10.20	8.53	**6.72**	8.19	8.61	6.82	10.47	7.68
	0.75	35.59	25.47	21.40	24.97	37.23	37.81	**21.28**	42.35	51.16
full	0	23.75	29.70	25.30	**22.06**	24.41	24.35	22.46	26.36	34.52
	0.75	67.25	45.30	**38.87**	40.50	71.08	72.80	38.39	101.58	108.97

mance as in the previous simulation studies. The methods in the first group (the CH-g, the Beta-prime prior, the robust prior and BIC) prefer smaller models while those in the second group (the Uniform prior, the Hyper-g, Jeffreys' prior, EB and AIC) are in favor of larger models. As to individual predictors, all different methods agree to include X_1, X_2, X_5 and X_6. According to most methods, X_7 also should be included, while X_4 can be excluded. For X_3, it is not clear whether it should be included.

We also examine the out-of-sample BMA estimation performance by ten-fold cross validation. Due to the somewhat high variability of Bernoulli distribution, almost no significant difference in classification error can be revealed. In this case, we recommend to use our CH-g prior, since most of the methods we compared here are actually special cases of it.

Table 2.7: Poisson regression: median SSE of $\boldsymbol{\beta}$: $\sum_{j=1}^{p}(\hat{\beta}_i - \beta_i^*)^2 \times 1000$ based on 100 realizations. Column-wise smallest value is in bold type. Friedman test shows that (1) CH-g(50, 0.5, 0) is significantly different from the Beta-prime in all scenarios; (2) CH-g(50, 0.5, 0) is significantly different from the robust prior except for Row 4, 5; (3) AIC is highly right skewed in Row 1, 2.

scenario	r	CH-g(50, 0.5, 0)	Uniform	Hyper-g	Jeffreys	Beta-prime	robust	EB	AIC	BIC
null	0	0.03	0.28	0.37	0.93	0.02	0.04	0.04	**0.00**	**0.00**
	0.75	0.02	0.32	0.42	1.00	0.01	0.03	0.03	**0.00**	**0.00**
sparse	0	1.58	3.06	2.67	2.05	1.42	1.67	2.44	2.93	**1.00**
	0.75	1.87	4.69	3.75	3.09	1.95	1.99	3.28	1.98	**1.42**
medium	0	4.56	5.14	4.90	4.81	4.54	4.52	4.77	5.89	**3.79**
	0.75	26.05	18.40	18.69	22.48	29.74	24.73	18.94	**9.66**	41.85
full	0	10.26	8.25	8.44	8.67	11.36	10.42	8.45	**6.07**	10.62
	0.75	48.66	35.03	34.20	37.01	51.96	48.14	33.94	**27.04**	66.11

2.7 Conclusion

In this chapter, we present a wide class of mixtures of g-priors, the CH-g prior, which extends several commonly used mixtures of g-priors to GLMs. We show in theoretical studies that the CH-g prior satisfies asymptotic criteria such as model selection consistency and parameter estimation consistency under specific choices of the hyper parameters.

We also propose a more generalized framework using the CCH prior, which encompasses but only CH-g prior itself, but also some hyper priors on g that are not special cases of CH distribution as well. One direction of our future work is to understand the theoretical and empirical performance of for GLMs with the CCH hyper prior.

Since our CH-g prior yields marginal likelihoods in tractable form, it has the ad-

Table 2.8: Pima Indian diabetes data: marginal posterior inclusion probability.

	CH-g(266, 0.5, 0)	Uniform	Hyper-g	Jeffreys	Beta-prime	robust	EB	AIC	BIC
X_1	0.966	0.981	0.980	0.978	0.960	0.958	0.980	0.990	0.946
X_2	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
X_3	0.290	0.536	0.507	0.471	0.221	0.196	0.506	0.684	0.100
X_4	0.281	0.516	0.488	0.453	0.216	0.193	0.487	0.662	0.103
X_5	0.998	0.999	0.999	0.999	0.998	0.998	0.999	0.999	0.997
X_6	0.995	0.998	0.998	0.998	0.993	0.993	0.998	0.999	0.987
X_7	0.580	0.785	0.764	0.737	0.503	0.479	0.764	0.884	0.334

vantage of computational efficiency in comparing models based on Bayes factors. We study its selection and estimation performance empirically using data with relatively small p, where enumerating the entire model space is feasible. However, when p increases (e.g. larger than 25), it is almost impossible to visit all potential models. In this case, we plan to incorporate stochastic search algorithms such as Bayesian adaptive sampling (Clyde et al., 2011), that may incorporate the approximate marginal likelihoods, and thus avoid computationally expensive model search alternatives such as reversible jump MCMC (Green, 1995).

3

The Local Rotation Invariant Prior

3.1 Introduction

In real world applications, the number of potential predictors in linear regression can be very large, while the response may be related to only a small proportion of all the predictors. Selecting one model and making inferences solely based on it ignores model uncertainty. Bayesian model averaging (BMA) (Hoeting et al., 1999) addresses model uncertainty (Clyde and George, 2004) by averaging the quantity of interest across all possible models, and thus often achieves more precise parameter estimation and prediction.

Zellner's g-prior (Zellner, 1986) and mixtures of g-priors (Liang et al., 2008) are commonly used for Bayesian model selection and model averaging, because of their computational efficiency and consistency (under regulatory conditions). Among mixtures of g-priors, the Zellner-Siow prior (Zellner and Siow, 1980) is considered a benchmark for BMA (Carvalho et al., 2009). The g-priors and mixtures of g-priors have an advantage of being invariant to linear transformations of the linear predictors. However, their inherent instability from ordinary least squares estimate (OLS) leads to their poor estimation accuracy when $\mathbf{X}^T\mathbf{X}$ is nearly singular. Moreover, according to Maruyama and George (2011), the g-prior imposes unwanted shrinkage towards zero along larger principle components, which is counter-intuitive.

Ridge regression (Hoerl and Kennard, 1970), the lasso (Tibshirani, 1996) and Bayesian shrinkage approaches such as the horseshoe estimator (Carvalho et al., 2010) use penalization methods to handle highly correlated design matrices. Simulation studies in Tibshirani (1996) suggest that ridge regression outperforms the lasso when regression coefficients are small and covariates are highly correlated. Ridge regression is invariant to orthogonal rotation of the coordinate system, while the lasso and the horseshoe prior are not, which requires that we should specify a coordinate system of interest. In terms of model selection, the lasso has the advantage over ridge regression, as it can shrink coefficients to exact zeros through modal estimators. However, lasso's selection procedure cannot be validated by optimizing any explicit loss or utility function, since the estimate of its tuning parameter λ is obtained via cross validation. The same issue exists for all the continuous Bayesian shrinkage priors without positive masses at zeros, including the horseshoe.

Since ridge regression and the lasso cannot uniformly dominate each other, the elastic net (Zou and Hastie, 2005) has been proposed to combine their strengths by using a mixture of L_1 and L_2 penalties on the coefficients. The elastic net can be considered as a stabilized generalization of the lasso, which is able to shrink coefficients to exact zeros while resolving the problems the lasso has with highly correlated predictors. Although globally predicting more accurately than the lasso, the elastic net loses to ridge regression empirically when oracles are non-sparse. In addition to the elastic net, some other penalization methods also incorporate both L_1 and L_2 penalties. For example, the group lasso (Yuan and Lin, 2006) targets regression problems with known group structures among covariates, such as multilevel factors. By imposing an intermediate between L_1 and L_2 penalties, the group lasso enjoys the desirable property of selecting predictors in the same group together.

Motivated by these methods, we propose an alternative fully Bayes approach that shrinks coefficients to zero more efficiently than the lasso in sparse cases, yet performs as well as ridge regression in non-sparse cases. We assign a Dirichlet Process (DP) mixture hyperprior, which naturally induces groups among coefficients, and uses a rotation invariant prior within each group. Compared with the group lasso, this approach does not require pre-specified group structure, and takes into account the uncertainty of groups.

Section 2 introduces our Local Rotation Invariant (LoRI) prior. We illustrate its adaptivity to both sparse and non-sparse regressions by examining its marginal and joint shrinkage properties. We also demonstrate that LoRI has bounded influence, which ensures its ability to preserve large signals. Section 3 details the Markov chain Monte Carlo procedure we implement for posterior computation. Section 4 compares parameter estimation accuracy of LoRI and other widely used methods including the horseshoe, the Bayesian lasso, g-prior, mixtures of g-prior, the lasso and ridge regression on two simulation examples. Section 5 shows LoRI's prediction accuracy on a protein activity dataset. Section 6 contains a discussion and direction of future work.

3.2 The Local Rotation Invariant Prior

In linear regressions, responses y_i are predicted by a linear combination of p-dimensional explanatory variables $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^T$ with an independent Gaussian noise:

$$y_i = \alpha + \sum_{j=1}^{p} x_{i,j}\beta_j + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \frac{1}{\phi}\right), \quad i = 1, \dots, n \tag{3.1}$$

where α is the intercept and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are the regression coefficients. The precision parameter ϕ equals the inverse variance of the errors. Without loss of generality, we assume that $\mathbf{Y} = (y_1, \ldots, y_n)^T$ and columns of design matrix $\mathbf{X}_j = (x_{1,j}, \ldots, x_{n,j})^T$, $j = 1, \ldots, p$ are centered and furthermore $\mathbf{X}_j$ are also scaled to have norm 1.

According to Diaconis and Ylvisaker (1985), in the exponential families, any prior distribution can be well approximated by finite mixtures of conjugate priors. In linear regressions, scale mixtures of normals priors can be denoted as mixtures of (normal-gamma) conjugate priors, with a distribution placed on the precision of the normal distribution. A popular alternative is to assign Dirichlet process (DP) priors on hyper parameters, which automatically induces discrete mixtures of conjugate priors. These DP mixture models achieve flexible mixtures by circumventing parametric specification of hyperpriors. Furthermore, the DP induces a discrete structure, which yields an automatic grouping among coefficients. For example, MacLehose and Dunson (2010) proposes a DP mixture model to shrink coefficients into a small number of clusters.

We propose a Local Rotation Invariant (LoRI) prior. This semi-parametric shrinkage prior can be considered as a mixture of normals with a DP hyperprior on the normal variances, i.e.,

$$\beta_j \mid \omega_j \overset{\text{ind}}{\sim} \mathrm{N}\left(0, \omega_j\right) \tag{3.2}$$

$$\omega_j \sim D \tag{3.3}$$

$$D \mid m, D_0 \sim \mathrm{DP}(m\ D_0) \tag{3.4}$$

$$D_0(\omega \mid \phi, \rho) = \int_0^\infty \mathrm{IG}\left(\omega; \frac{1}{2}, \frac{\eta^2}{2}\right) \cdot \left\{(1-\rho)\delta_0(\omega) + \rho\ \mathrm{C}^+\left(\eta; 0, \frac{1}{\sqrt{\phi}}\right)\right\} d\eta \tag{3.5}$$

Each dimension β_j has conditionally independent normal prior with mean zero and variance ω_j. The hyper variance parameters ω_j are assigned a random prior probability measure D. This unknown measure D is further assigned a prior measure DP($m\ D_0$), where the base measure D_0 corresponds to our best prior guess for D and the DP precision parameter m controls the similarity between D and D_0.

3.2.1 Independent Cauchy versus Multivariate Cauchy

After marginalizing D out, our DP mixture prior has the Polya Urn (Blackwell and MacQueen, 1973) representation:

$$\omega_{k+1} \mid \omega_1, \ldots, \omega_k, m, D_0 \sim \frac{m}{m+k}\ D_0 + \sum_{h=1}^{k} \frac{1}{m+k}\ \delta_{\omega_h}, \tag{3.6}$$

which iteratively gives the prior distribution of ω_{k+1} conditional on the previous parameters $\{\omega_1, \ldots, \omega_k\}$, for any $k = 0, \ldots, p-1$. Because some ω_j can take the same values, the Polya Urn scheme (3.6) implies the prior dependency among $\{\omega_1, \ldots, \omega_p\}$. In this sense, $\{\beta_1, \ldots, \beta_p\}$ are also dependent a priori. Suppose there are K distinct values $\{\omega_1^*, \ldots, \omega_K^*\}$, where each of them has independent D_0 prior; then the original p parameters can be clustered to K different groups, such that all the parameters in the k-th group $\{\omega_{k_1}, \ldots, \omega_{k_{m_k}}\}$ take the value ω_k^*. Denote a vector of group indicators as $\mathbf{c} = (c_1, \ldots, c_p)^T$, where $c_j = k$ if and only if $\omega_j = \omega_k^*$, for $j = 1, \ldots, p$.

Given the group structure $\mathbf{c}$, the marginal prior of the jth regression coefficient β_j can be decomposed as the following hierarchical form after introducing a latent variable $\eta_{c_j}^*$:

$$\beta_j \mid \mathbf{c}, \boldsymbol{\omega}^* \overset{\text{ind}}{\sim} \mathrm{N}(0, \omega_{c_j}^*) \tag{3.7}$$

$$\omega_{c_j}^* \mid \eta_{c_j}^* \overset{\text{ind}}{\sim} \mathrm{IG}\left(1/2, \eta_{c_j}^{*2}/2\right) \tag{3.8}$$

with hyperprior

$$\eta_{c_j}^* \overset{\text{iid}}{\sim} (1-\rho)\ \delta_0 + \rho\ \mathrm{C}^+\left(0, \frac{1}{\sqrt{\phi}}\right) \tag{3.9}$$

Here we generalize the Inverse Gamma distribution with scale parameter being zero to represent degenerate distribution of positive point mass at zero. After integrating $\omega^*_{c_j}$ out in (3.7) and (3.8), the marginal prior on β_j becomes a univariate Cauchy distribution with scale parameter $\eta^*_{c_j}$:

$$\beta_j \mid \mathbf{c}, \eta^*_{c_j} \sim \mathrm{C}(\beta_j; 0, \eta^*_{c_j}) \tag{3.10}$$

With tails heavier than a normal distribution, Cauchy priors, along with other prior distributions in the Student t family, are considered more robust and can better adapt to large signals. Jeffreys (1961) justifies the use of the Cauchy prior on normal location parameters in terms of information consistency, which suggests that the Bayes factor on testing location against zero goes to infinity if the observations are overwhelmingly far from zero. Dawid (1973) shows that under the Cauchy prior, the posterior mean of a normal location parameter converges to the observation as the observation goes to infinity. Therefore, in Bayesian model selection and model averaging, Student t distributions, especially the Cauchy distribution, are used conventionally as prior distributions on regression coefficients. For example, Zellner and Siow (1980) proposes a multivariate Cauchy prior distribution on regression coefficients. Tipping (2001) applies independent Student t prior distributions whose scale parameters and degrees of freedom are small to sparse problems.

We find that as shrinkage priors for multi-dimemsional coefficients, independent univariate Cauchy distribution performs differently from multivariate Cauchy distribution. We will illustrate this property of the shrinkage prior $p(\boldsymbol{\beta})$ in the framework of penalized regression, whose estimate is obtained by minimizing the sum of squared errors (SSE) and a penalty $f(\boldsymbol{\beta})$,

$$\hat{\boldsymbol{\beta}}_f = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + f(\boldsymbol{\beta}) \right\}. \tag{3.11}$$

In particular, the maximum a posteriori (MAP) estimate under prior $p(\boldsymbol{\beta})$ equals

$\hat{\boldsymbol{\beta}}_f$ if $f(\boldsymbol{\beta}) = -\frac{2}{\phi}\log p(\boldsymbol{\beta})$. For example, independent normal priors and independent double exponential (Laplace) priors are Bayesian counterparts to ridge regression and the lasso respectively.

We first show the bivariate contour plots of the negative logarithm of prior densities of independent double exponentials and independent normals in the two upper panels of Figure 3.1. Between ridge regression and the lasso, the latter can yield sparse solutions while the former cannot. From a Bayesian point of view, this difference of their posterior solutions lies in the shapes of their prior densities. The diamond-shaped contours indicate that the double exponential priors place more probability mass along the axes, where one regression coefficients is set to zero. In contrast, the circular contours imply that the independent normal priors place equivalent probability in all directions rather than favoring the directions along the axes.

With respect to shrinking all directions equally, the difference between univariate independent Cauchy and multivariate Cauchy distributions resembles that between the lasso and ridge regression (see the lower two panels in Figure 3.1). The contours of independent Cauchy priors are somewhat round near the origin, which is similar to the contours of ridge regression and the lasso combined. However, as the norm $\|\boldsymbol{\beta}\|$ increases, it gradually becomes star-shaped, which suggests that the independent Cauchy prior distribution imposes even stronger shrinkage than the lasso towards the axes. On the contrary, the contour shape of a bivariate Cauchy distribution remains circular, which indicates equal shrinkage along all directions.

When considered as scale mixtures of normal distributions, the independent Cauchy priors have different hyper parameters governing the prior variance of every dimension. These different parameters contributes to heterogeneous amounts of shrinkage along each regression coefficients. Given the group structure induced by DP, for any pair of parameters $(\beta_j, \beta_{j'})$, if they belong to different groups, their have

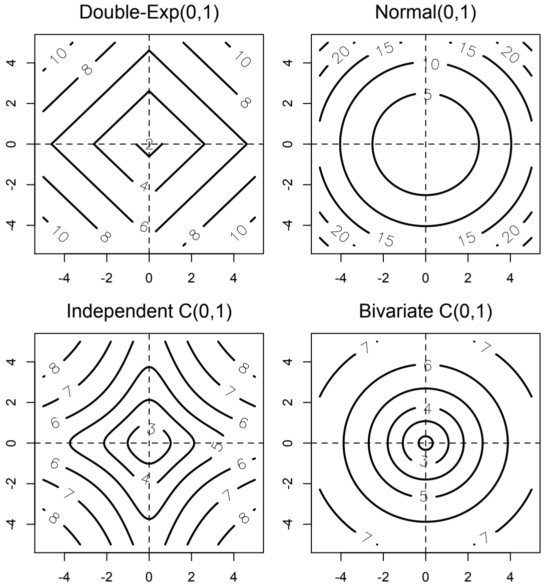


FIGURE 3.1: Contour plots of $-\log p(\beta_1)$ versus $-\log p(\beta_2)$. From upper left to lower right, β_1 and β_2 are independent and identically distributed as Laplace$(0, 1)$, Normal$(0, 1)$, independent Cauchy$(0, 1)$, bivariate Cauchy$(0, 1)$.

conditionally independent Cauchy prior:

$$p(\beta_j, \beta_{j'} \mid \boldsymbol{\eta}^*, c_j \neq c_{j'}) = \int p(\beta_j \mid \omega^*_{c_j}) p(\omega^*_{c_j} \mid \eta^*_{c_j}) d\omega^*_{c_j} \cdot \int p(\beta_{j'} \mid \omega^*_{c_{j'}}) p(\omega^*_{c_{j'}} \mid \eta^*_{c_{j'}}) d\omega^*_{c_{j'}}$$

$$= \mathrm{C}(\beta_j; 0, \eta^*_{c_j}) \cdot \mathrm{C}(\beta_{j'}; 0, \eta^*_{c_{j'}})$$

On the other hand, if regression coefficients β_j and $\beta_{j'}$ belong to the same group $(c_j = c_{j'})$, then their conditional joint prior is a bivariate Cauchy:

$$p\left(\begin{pmatrix}\beta_j \\ \beta_{j'}\end{pmatrix} \mid \boldsymbol{\eta}^*, c_j = c_{j'}\right) = \int \mathrm{N}\left(\begin{pmatrix}\beta_j \\ \beta_{j'}\end{pmatrix}; \mathbf{0}, \omega^*_{c_j}\mathbf{I}\right) \mathrm{IG}\left(\omega^*_{c_j}; 1/2, \eta^{*2}_{c_j}/2\right) d\omega^*_{c_j}$$

$$= \mathrm{C}_2\left(\begin{pmatrix}\beta_j \\ \beta_{j'}\end{pmatrix}; \mathbf{0}, \eta^{*2}_{c_j}\mathbf{I}\right)$$

Because this bivariate Cauchy has circular contours, it does not favor sparse models a priori. Furthermore, this prior can be considered as a scale mixture of normals with a single variance parameter, which alone controls the magnitudes of shrinkage in all dimensions.

3.2.2 Local Rotation Invariance

To rigorously differentiate between the independent and multivariate Cauchy prior distributions, we adopt the concept of rotation invariance. We note that the sparsity of regression problems changes under transformation of design matrices. For example, an orthogonal rotation $\mathbf{U}$ of the coordinate systems $\mathbf{X}$

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}\mathbf{U})(\mathbf{U}^T\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{3.12}$$

transforms the predictors $\mathbf{X}$ to $\mathbf{XU}$, and the vector of coefficients $\boldsymbol{\beta}$ to $\mathbf{U}^T\boldsymbol{\beta}$. If under the original design the true model is sparse, i.e., some of the true coefficients $\boldsymbol{\beta}_0$ are zeros, then after the rotation all dimensions in $\mathbf{U}^T\boldsymbol{\beta}_0$ are probably nonzero. We say a prior distribution $p(\boldsymbol{\beta})$ is rotation invariant if $p(\mathbf{U}^T\boldsymbol{\beta})$ has the same prior density. Such priors place similar amount of shrinkage before and after the rotation of coordinate systems. In contrast, rotation variant priors such as the independent double exponential cannot achieve sparse solution in all coordinate systems. In particular, in group selection problems, rotation invariant priors are assigned to regression coefficients in the same group, so that all the predictors in the same group are imposed similar amounts of shrinkage. For example, the group lasso (Yuan and Lin, 2006) has a local L_2 penalty within each group.

Conditional on the group structure $\mathbf{c}$, we denote $\boldsymbol{\beta}_{(k)} = (\beta_{k_1}, \ldots, \beta_{k_{m_k}})^T$ as the vector that consists all coefficients in the kth group. Rewrite LoRI prior (3.2)-(3.5) in a hierarchical form for the kth group

$$\boldsymbol{\beta}_{(k)} \mid \omega_k^* \sim \mathrm{N}\left(\mathbf{0}, \omega_k^*\mathbf{I}_{m_k}\right)$$

$$\omega_k^* \mid \eta_k^* \sim \mathrm{IG}\left(\frac{1}{2}, \frac{\eta_k^{*2}}{2}\right)$$

After integrating ω_k^* out, we obtain a m_k dimensional multivariate Cauchy distribution as the prior for the coefficient in group k, i.e.,

$$p\left(\boldsymbol{\beta}_{(k)} \mid \eta_k^*\right) \propto \frac{1}{|\boldsymbol{\Sigma}_{(k)}|^{1/2}\left[1+\boldsymbol{\beta}'_{(k)}\boldsymbol{\Sigma}_{(k)}^{-1}\boldsymbol{\beta}_{(k)}\right]^{(1+p_k)/2}}$$

with $\boldsymbol{\Sigma}_{(k)} = \eta_k^{*2}\mathbf{I}_{m_k}$. For any positive η_k^*, this corresponding penalty term of $\boldsymbol{\beta}_{(k)}$ has spherical contours. This implies that the multivariate Cauchy prior with a single scale parameter is rotation invariant. Therefore, conditional on the group structures, our method assigns a spherical multivariate Cauchy prior to the vector of coefficients in each group, which has a "local" rotation invariance property. Without loss of generality, suppose the the order of regression coefficients and their corresponding predictors are rearranged according to the groups,

$$\boldsymbol{\beta} = \left(\boldsymbol{\beta}_{(1)}^T, \ldots, \boldsymbol{\beta}_{(K)}^T\right)^T$$
$$\mathbf{X} = \left(\mathbf{X}_{(1)}, \ldots, \mathbf{X}_{(K)}\right)$$

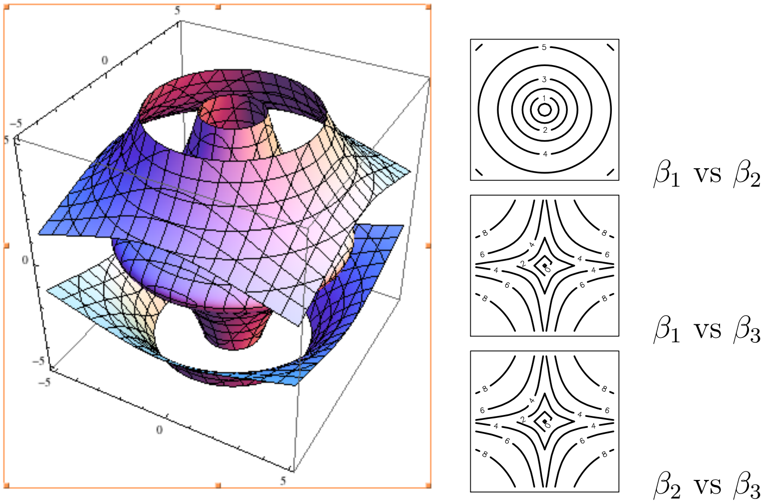
If we rotate the predictors in each group $\boldsymbol{\beta}_{(k)}$ by an $m_k \times m_k$ orthogonal matrix $\mathbf{U}_{(k)}$, which transforms the regression model (3.14) to

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}_{(1)}\boldsymbol{\beta}_{(1)} + \ldots + \mathbf{X}_{(K)}\boldsymbol{\beta}_{(K)} + \boldsymbol{\epsilon}\\ &= \left(\mathbf{X}_{(1)}\mathbf{U}_{(1)}\right)\left(\mathbf{U}_{(1)}^T\boldsymbol{\beta}_{(1)}\right) + \ldots + \left(\mathbf{X}_{(K)}\mathbf{U}_{(K)}\right)\left(\mathbf{U}_{(K)}^T\boldsymbol{\beta}_{(K)}\right) + \boldsymbol{\epsilon},\end{aligned}$$

then after the rotation, $p(\mathbf{U}_{(k)}^T\boldsymbol{\beta}_{(k)} \mid \eta_k^*)$ remains the same multivariate Cauchy density. Thus our prior can be considered locally rotation invariant in this sense.

We visualize the prior contours of $\boldsymbol{\beta}$ in a simple case that only contains three covariates. Suppose the first two coefficients are in the same group and the third coefficient is in a different group, $\omega_1 = \omega_2 = \omega_1^*$ and $\omega_3 = \omega_2^*$. Table 3.1 shows its 3D contour plot, where x, y, z axes represent $\beta_1, \beta_2, \beta_3$ respectively. Specifically, 2D contours on the horizontal hyper plain of β_1 vs β_2 have circular shapes and 2D contours on the vertical hyper plain of β_1 vs β_3 or β_2 vs β_3 have the contour shapes of 5 Cauchy distributions, round in the inside and star-shaped in the outside.

Table 3.1: 3D Contour plots of $(\beta_1, \beta_2, \beta_3)$ with $\eta_1^* = 1$ and $\eta_2^* = 1$.



The local rotation invariant structure of LoRI can also be shown from the penalized regression perspective. The corresponding penalty term in (3.11) for LoRI can be written as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta}) + \sum_{k=1}^{K} h\left(\boldsymbol{\beta}_{(k)}, \eta_k^*\right) \right\}, \tag{3.13}$$

where $h(\mathbf{x}, \eta)$ is the negative logarithm of multivariate Cauchy density with $\boldsymbol{\Sigma} = \eta^2 \mathbf{I}$. The penalty term in (3.13) has a similar form to the group lasso's penalty term. However, our model is different from the group lasso in the following three aspects.

First, the group lasso is designed for group selection with pre-specified group structures among covariates, while LoRI aims to solve more general questions, including the ones without known group information. In fact, LoRI is even capable of revealing group structures among covariates, according to the strength of their impacts on the response variable, rather than their correlation with each other. (See the simulation example in Section 3.4.2.)

Second, LoRI takes into account the uncertainty of group structures. Consider two very extreme cases: i) all covariates exist in the same group or ii) each covariate forms its own group. In the former case, LoRI is equivalent to a mixture of a ridge estimator with hyperprior $G(1/2, 1/2)$ assigned on the normal precision parameter and a point mass at the origin. This case appears when all covariates have similar strength in predicting the response. LoRI degenerates to a rotation invariant prior in this case, which is desirable since it imposes the same amount of shrinkage on each dimension. In the latter case, LoRI performs the same as independent mixtures of univariate Cauchy distributions and point masses at zeros. Therefore, when covariates are heterogeneous in prediction, LoRI treats each dimension differently and is capable of yielding sparse solutions.

Finally, while the group lasso only has one tuning parameter for all the groups, LoRI has different parameters η_k^* to control shrinkage for each group. This flexibility yields different amounts of shrinkage that can better adapt to the data.

3.2.3 Point Mass at Zero

According to the base measure D_0 (3.5), ω_j has a "spike and slab" type marginal prior, which is a mixture of point mass at zero with weight $1 - \rho$ and a continuous distribution with weight ρ. Since the positive probability mass at zero on ω_j implies a positive probability mass at zero on β_j, each regression coefficient β_j can be considered as having a "spike and slab" prior marginally. The positive point mass at zero component in the prior leads to multiple shrinkage (George, 1986) in the Bayesian model averaging estimator. Given the group structure $\mathbf{c}$, the point mass at zero in the prior of ω_k^* yields to a positive posterior probability on the models which do not contain the whole vector of the coefficients in the kth group $\boldsymbol{\beta}_{(k)}$.

Furthermore, the point mass at zero enable LoRI to have valid selection rules that can be justified by optimizing certain loss functions. For example, by using the posterior median estimator which minimizes the L_1 loss, some predictors can be excluded from the model if the majority of samples drawn from their posterior distributions are zeros. In this way, LoRI prior can be considered as a variable selection prior. On the other hand, according to Tipping (2001), independent Cauchy priors with small scale parameters can also yield strong shrinkage. However, without the point mass at zero, the shrinkage prior that only consists of continuous distributions ignores uncertainty in the model space, even though it may achieve sparse point estimates under the posterior mode.

3.2.4 Robustness to Large Coefficients

LoRI achieves a balance between shrinking trivial coefficients to zeros and preserving large ones. We will demonstrate in the following special case of orthonormal design that LoRI's prior influence is bounded and vanishes in the limit.

Orthonormal designs have tractable analytical forms under many penalized regression methods, and can be obtained in real world applications from well-designed experiments or in signal processing applications using $\mathbf{X}\mathbf{X}^T = \mathbf{I}$ wavelets, for example Clyde and George (2004). Suppose $\mathbf{X}$ is a squared orthogonal matrix ($\mathbf{X}^T\mathbf{X} = \mathbf{I}$). Then the regression model (3.1) can be transformed into the following form by a rotation

$$\mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \mathbf{X}^T\boldsymbol{\epsilon}, \tag{3.14}$$

where $\mathbf{X}^T\boldsymbol{\epsilon}$ remains a vector of independent Gaussian errors with the same variance $1/\phi$. Notice that the response vector after the transformation $\mathbf{X}^T\mathbf{Y}$ equals the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$. Thus (3.14) can be rewritten as independent normal observations $\hat{\beta}_j$ with different locations β_j and a common precision parameter ϕ

$$\hat{\beta}_j = \beta_j + e_j,\ e_j \overset{\text{iid}}{\sim} \mathrm{N}\left(0, \frac{1}{\phi}\right),\ \text{for } j = 1, \ldots, p. \tag{3.15}$$

Therefore, if the number of potential predictors p in an orthogonal-designed linear regression equals the sample size n, it can be represented in the form of (3.15). We call this special case the normal means case, which includes many common models. For example, a random effect model

$$z_{j,r} \sim N(\beta_j, \sigma^2),$$

where $z_{j,r}$ is the rth observation within the jth subject, for $r = 1, \dots, m$, can be obtained by substituting $\hat{\beta}_j$ in (3.15) with the sufficient statistics $\bar{z}_j$, the sample mean of $\{z_{j,1}, \dots, z_{j,m}\}$, and changing the variance $1/\phi$ to σ^2/m accordingly. In addition, nonparametric regressions that naturally have orthonormal bases such as splines and wavelets can also be represented in the format of model (3.15).

For the normal means case, Bayesian shrinkage methods such as the empirical Bayes approach (Clyde and George, 2000; Johnstone and Silverman, 2004) and the horseshoe (Carvalho et al., 2010) were originally created to estimate sparse signals among β_j's while eliminating the disturbance caused by background noise.

We can use the normal means case (3.15) to illustrate the shrinkage mechanism of LoRI. Under standard normal errors, i.e. $\phi = 1$, the marginal conditional posterior mean is

$$\mathbb{E}\left[\beta_j | \hat{\boldsymbol{\beta}}, \boldsymbol{\omega}\right] = \left(1 - \frac{1}{\omega_j + 1}\right) \hat{\beta}_j \tag{3.16}$$

The term enclosed in the parentheses in (3.16) takes value on interval $[0, 1)$ and can be considered a shrinkage factor. Small ω_j indicates a high prior concentration around zero, which results in a small shrinkage factor which suggests strong shrinkage. In particular, $\omega_j = 0$ shrinks the posterior mean of the location parameter to exact zero. On the other hand, large ω_j indicates a large dispersion in prior density, which associates with a shrinkage factor close to 1 and thus avoids over-shrinking a large signal.

According to the Polya Urn scheme (3.6), the marginal prior on the first parameter ω_1 equals to exactly the base measure D_0. Because of exchangeability, (3.6) holds under all permutation of the order among ω_j. Therefore, for any $j = 1, \ldots, p$, the marginal prior measure of LoRI for ω_j is also D_0.

For the normal means case, Dawid (1973), Pericchi and Smith (1992), Pericchi and Sanso (1995), and Carvalho et al. (2010) show that certain heavy tail priors such as the double exponential prior, the Student t prior and the horseshoe have bounded influence. According to the following theorem, marginally, LoRI also has bounded prior influence. Furthermore, LoRI's prior influence vanishes for large observation in the limit.

Theorem 5. *Suppose $\hat{\beta}_j = \beta_j + e_j$ and $e_j \sim N(0, 1/\phi)$, where the location parameters β_j are unknown and the precision parameter ϕ is known. Then according to LoRI, the marginal prior on β_j*

$$(1-\rho)\delta_0(\beta_j) + \rho \int_0^\infty \int_0^\infty N(\beta_j; 0, \omega) \cdot IG\left(\omega; \frac{1}{2}, \frac{\eta^2}{2}\right) \cdot C^+\left(\eta; 0, \frac{1}{\sqrt{\phi}}\right) d\eta d\omega, \quad (3.17)$$

1) has bounded prior influence $E(\beta_j|\hat{\beta}_j) - \hat{\beta}_j$, for any $\hat{\beta}_j \in \mathbb{R}$;

2) Prior influence vanishes for large $\hat{\beta}_j$:

$$\lim_{|\hat{\beta}_j| \to \infty} E(\beta_j|\hat{\beta}_j) - \hat{\beta}_j = 0 \quad (3.18)$$

Proof: see Appendix B.1.

3.2.5 Hyper Priors and Parameters

On the choice of hyperpriors, Gelman (2006) suggests using a half-t prior on the hierarchical normal standard deviation parameter; the horseshoe has half-Cauchy priors on both the local and the global scale parameters (Polson and Scott, 2012) . According to (3.5), we take the priors of the scale parameters η_j in the continuous component of D_0 to be half-Cauchy distributions with the common scale $\frac{1}{\sqrt{\phi}}$, which can adapt to different variations in the observation errors. Because of the flexibility achieved by the DP structure, LoRI does not require another hierarchy of a global shrinkage parameter.

The marginal inclusion probability ρ controls the model size. We assign a hyperprior $\rho \sim \text{Beta}(a_\rho, b_\rho)$ with hyper parameters $a_\rho = 1$ and

$$b_\rho = p^c,$$

where c takes value between 0 and 1. Choices of c reflect different prior beliefs in model sparsity. In the case of $c = 0$, the prior on ρ degenerates to a uniform distribution on $(0, 1)$, which implies that on average half of the covariates should be included. In the case of $c = 1$, the prior mean of ρ decreases to $p/(1 + p)$, which yields a more sparse solution with a model size close to 1. This choice is desirable to solve sparse problems, where expected model sizes do not increase with p. Without prior knowledge of the sparsity of the true model, we avoid specifying extreme values such as 0 or 1 on hyper parameter c. Instead, we recommend default value $c = 1/2$, which allows LoRI to better adapt to different model size p and sparsity.

The Polya Urn Schemes (3.6) indicates that the DP precision parameter m controls the number of different values among $\{\omega_1, \ldots, \omega_p\}$. Larger m yields more clusters, since ω_{k+1} is more likely to differ from $\omega_1, \ldots, \omega_k$; and vice versa. In the absence of prior information on the number of clusters, we recommend a hyperprior $m \sim$Gamma(a_m, b_m) with $a_m = b_m = 1$.

3.3 Posterior Computation

In a stick-breaking representation, the posterior of ω_j:

$$P(\omega_j) = \sum_{k=1}^{\infty} p_k \cdot \delta_{\omega_k^*}(\omega_j) \tag{3.19}$$

$$p_k = v_k \prod_{l<k} (1 - v_l) \tag{3.20}$$

$$v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), k = 1, 2, \ldots \tag{3.21}$$

where ω_k^* are posterior samples of ω_j as if under the same prior in the base measure D_0. Conventional sampling algorithms for the DP prior such as the blocked Gibbs sampler (Ishwaran and James, 2001) truncate (3.19) to a finite number of components, which treat the DP models as finite mixture models. Some new DP sampling approaches circumvent the unnecessary finite truncation step and remain simple to implement. We apply the exact block Gibbs sampler algorithm proposed by Papaspiliopoulos (2008), which combines the ideas of two efficient algorithms for non-parametric model sampling: retrospective sampling (Papaspiliopoulos and Roberts, 2008) and slice sampling (Walker, 2007). To draw posterior samples of the DP precision α, we apply the Gibbs sampler by introducing an auxiliary variable (Escobar and West, 1995).

We marginalize $\boldsymbol{\beta}$ out to improve MCMC mixing. Since closed forms of full conditionals are not available, we use the Metropolis-Hastings algorithm within the Gibbs sampler. We use a Gaussian random walk proposal to sample the conditional posterior of (ω_k^*, η_k^*) and ϕ. To obtain appropriate value of proposal variance, we apply the adaptive Metropolis (AM) algorithm (Haario et al., 2001), whose choice of proposal variance depends on historical draws of posterior samples. Although the adaptive Metropolis is not Markovian, this tuning free algorithm still achieves ergodicity.

Because parameters are updated univariately, the AM chain can get stuck in local modes, especially when covariates are highly correlated. To resolve this problem, a simple random swap step (Ghosh and Clyde, 2011) is added in each iteration of the AM chain. For a pair of highly correlated covariates $\{X_i, X_j\}$ where only one of their η being zero, we propose to swap the values of their corresponding parameters (ω_i, η_i) and (ω_j, η_j) with a small probability.

Detailed posterior sampling steps are listed in Appendix B.2.

3.4 Simulation Studies

3.4.1 Normal Means

The empirical Bayes method proposed by Johnstone and Silverman (2004) is considered a benchmark among shrinkage priors for detecting sparse signals. Carvalho et al. (2010) compares it with the horseshoe prior in the following simulation design. Suppose $n = 250$ independent observations $\hat{\beta}_i$ are drawn independently from $N(\beta_i, 1)$. The true values of the location parameters β_i are generated from independent mixtures of a Student t distribution $t_\xi(0, 3)$ with weight w and a point mass at zero with weight $1 - w$. Combinations of different levels of model sparsity $w \in \{0.05, 0.2, 0.5\}$ and signal strength $\xi \in \{2, 10\}$ are examined. For each combination 500 simulated datasets are generated.

We compare LoRI with both the horseshoe and empirical Bayes on the above simulated data. Empirical results suggest that different choices of priors on model precision ϕ do not lead to significant differences in posterior inferences. In fact, the reference prior $p(\phi) \propto 1/\phi$ and half-Cauchy prior on the inverse squared root of ϕ yield almost identical posteriors. Without loss of generality, we report the results under the half-Cauchy prior, which is a proper prior.

Similar to the horseshoe approach, we first use the posterior means as our default estimates for β_i. Table 3.2 shows the L_2 loss from the 500 independent simulations.

Table 3.2: Simulation study of normal means case: median sum of squared errors (SSE), $\sum_{i=1}^{n}(\beta_i - \tilde{\beta}_i)^2$, from 500 simulations. $\tilde{\beta}_i$ are posterior mean estimators for LoRI and the horseshoe. We use bootstrap with 500 samplings to estimate the corresponding standard errors of medians and report them in parentheses. Columnwise smallest error is in bold type.

	$w = 0.05$		$w = 0.2$		$w = 0.5$	
ξ	2	10	2	10	2	10
LoRI	**27** (0.6)	**30** (0.7)	**90** (1.3)	**91** (1.2)	175(1.6)	**174** (1.4)
Horseshoe	31 (0.5)	**30** (0.8)	98 (0.8)	94 (1.0)	**174** (0.9)	199 (3.3)
Empirical Bayes	29 (0.8)	33 (1.0)	113 (1.8)	124 (2.0)	388 (4.5)	441 (6.4)

The empirical Bayes method almost systematically yields largest estimation errors. Between LoRI and the horseshoe, LoRI performs as accurately as the horseshoe in the sparse small signal scenario ($w = 0.05, \xi = 10$) and non-sparse large signal scenario ($w = 0.5, \xi = 2$), and achieves smaller errors than the horseshoe in all other scenarios. Both methods have heavy tailed priors with high concentrations near zero, and thus are able to both shrink noises and preserve large signals. The left panel of Figure 3.2 compares these three methods in the sparse large signal scenario ($w = 0.05, \xi = 2$). Similar to the empirical Bayes and the horseshoe, LoRI estimates for large signals remain almost identical to the observed values, which agrees with the bounded influence property. For small observations, LoRI shrinks them severely to almost zeros. Thanks to the positive mass at zero, LoRI has flatter posterior slopes near the origin, which indicates better handling of noise in sparse scenarios. On the other hand, in non-sparse scenarios, LoRI's local rotation invariant property makes it perform similarly to ridge regression, and thus avoids over sparse solutions.

The posterior median estimate used by the empirical Bayes method is not optimal for L_2 loss, which explains the systematically large errors from the empirical Bayes method in Table 3.2. For a fair comparison, we also explore posterior median estimates for the two fully Bayes methods, LoRI and the horseshoe, and report

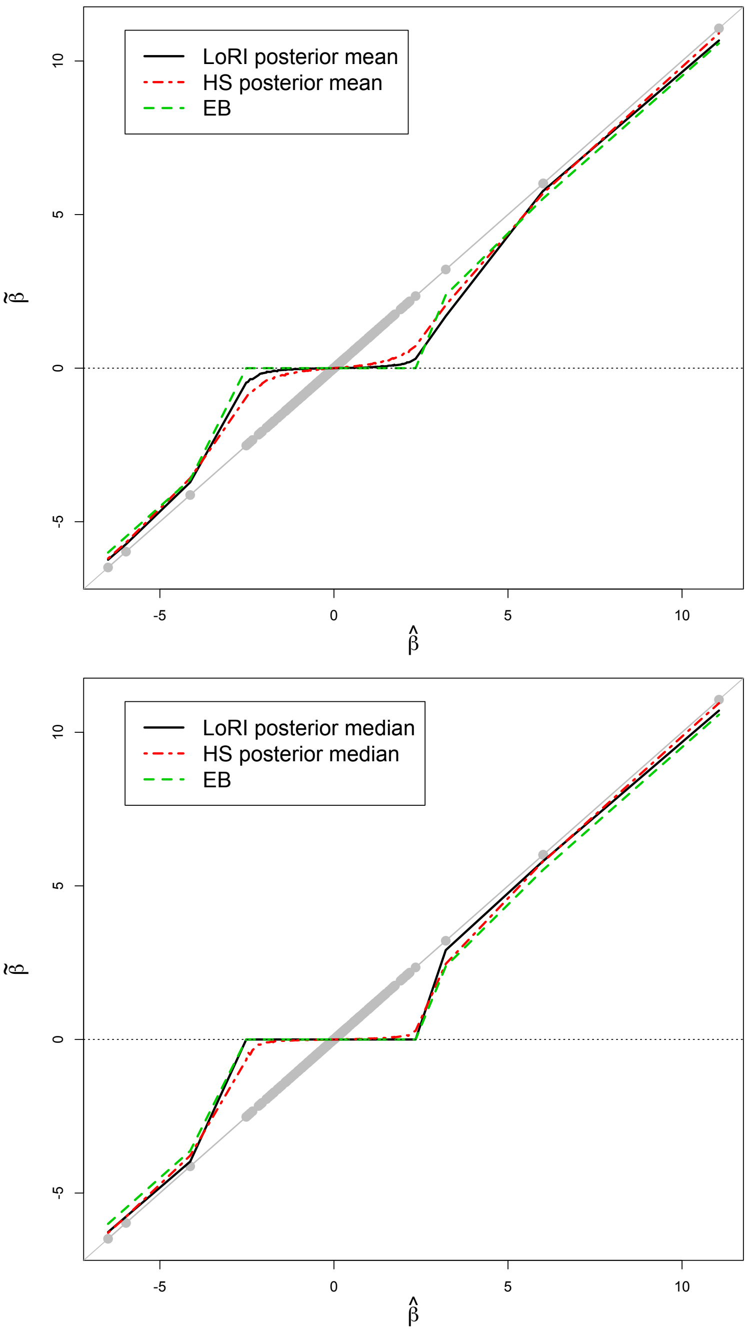


FIGURE 3.2: Simulation study of normal means case: observations $\hat{\beta}_j$ vs posterior estimates of location parameters $\tilde{\beta}_j$, from one simulation in scenario $w = 0.05, \xi = 2$. Grey line and dots are 45-degree diagonal line and values of observations. Left: posterior mean estimator for LoRI and the horseshoe. Right: posterior median estimator for LoRI and the horseshoe.

Table 3.3: Simulation study of normal means case: median sum of absolute errors (SAE), $\sum_{i=1}^{n} |\beta_i - \tilde{\beta}_i|$, from 500 simulations. $\tilde{\beta}_i$ are posterior median estimators for LoRI and the horseshoe. We use bootstrap with 500 samplings to estimate the corresponding standard errors of medians and report them in parentheses. Columnwise smallest error is in bold type.

	$w = 0.05$		$w = 0.2$		$w = 0.5$	
ξ	2	10	2	10	2	10
LoRI	**16** (0.3)	**17** (0.4)	**59** (0.7)	**61** (0.7)	**130** (0.7)	**132** (0.8)
Horseshoe	21 (0.5)	19 (0.4)	89 (1.1)	78 (0.7)	156 (0.5)	149 (0.8)
Empirical Bayes	**16** (0.3)	**17** (0.4)	62 (0.6)	64 (0.7)	179 (1.6)	193 (2.0)

the L_1 loss in Table 3.3. LoRI beats the horseshoe systematically, and outperforms the empirical Bayes in moderately sparse and non-sparse scenarios. In contrast, the horseshoe yields the largest L_1 errors in all scenarios, probably due to its lack of complete shrinkage to zero. Furthermore, the right panel of Figure 3.2 illustrates that the posterior median estimates of LoRI can reach exact zeros for small observations, while the horseshoe cannot.

3.4.2 Regression

To understand LoRI's performance in linear regressions with different correlation structures among covariates and different values of true coefficients, we compare LoRI with several commonly used Bayesian and non-Bayesian methods: the Zellner-Siow prior, the unit information prior (Kass and Wasserman, 1995), the horseshoe, the Bayesian lasso, the lasso, ridge regression, the elastic net and the OLS estimator on the simulation example originally designed in the lasso paper (Tibshirani, 1996). In this example, observations are simulated according to

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathrm{N}_n(0, \sigma^2\mathbf{I}),$$

with sample size $n = 200$, number of covariates $p = 8$, standard deviation in normal likelihood $\sigma = 3$ and pairwise correlation $\mathrm{corr}(\mathbf{X}_i, \mathbf{X}_j) = r^{|i-j|}$. Measurement for correlation r is originally set to 0.5 in the lasso paper. In order to compare different correlation levels, we also consider the cases $r \in \{0, 0.99\}$. The following two scenarios represent different structures of the true values of coefficients $\boldsymbol{\beta}$:

$$\text{Scenario 1: } \boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)$$

$$\text{Scenario 2: } \beta_j = 0.85, \text{ for all } j = 1, \ldots, 8.$$

Table 3.4 shows the estimation accuracy measured by the sum of squared errors for $\boldsymbol{\beta}$ under combinations of all three correlation levels and two coefficient structure scenarios. Among all the methods, the Zellner-Siow and the unit information prior are variations of g-prior. The Zellner-Siow has a multivariate Cauchy prior density and can be represented as a mixture of g-prior with inverse Gamma hyperprior on g. The unit information prior is the g-prior with $g = n$ and acts similarly to the BIC criteria. The Bayesian lasso is the posterior mean estimate on coefficients under independent Laplace prior. For all Bayesian approaches, we use the default estimates, which are posterior means.

In Scenario 1, the true model is sparse and the nonzero coefficients have comparatively large values. For cases with independent or moderately correlated predictors ($r = 0, 0.5$), the two g-prior methods result in the most precise estimations. LoRI slightly underperforms in relation to them but outperforms all other methods. Another Bayesian method, the horseshoe, also outperforms all non-Bayesian methods. Among the three non-Bayesian methods, the elastic net acts similarly to the lasso, and both of them outperform ridge regression, which is consistent with its reputation in sparse regression. These three methods all outperform the Bayesian lasso as they may shrink coefficients exactly to zero while the Bayesian lasso estimate retains all coefficients. However, in a highly correlated case ($r = 0.99$), we notice an obvious change in estimation performance. The Bayesian lasso yields the most accurate estimation, while the two g-prior related methods become the least reliable, due to their

Table 3.4: Simulation study of regression: median sum of squared error $\sum_{j=0}^{8}(\beta_j - \tilde{\beta}_j)^2$, from 500 simulations. $\tilde{\beta}_j$ are posterior mean estimates for Bayesian methods. We use bootstrap with 500 samplings to estimate the corresponding standard errors of medians and report them in parentheses. Column-wisely, smallest error is in bold type and second smallest error in italic type.

Scenario 1			
r	0	0.5	0.99
LoRI	*0.14* (0.01)	*0.17* (0.01)	*5.85* (0.22)
Zellner-Siow	**0.13** (0.01)	**0.15** (0.01)	8.76 (0.42)
Unit Info	**0.13** (0.01)	**0.15** (0.01)	8.35 (0.35)
Horseshoe	0.20 (0.01)	0.26 (0.01)	6.52 (0.28)
Bayesian lasso	0.29 (0.01)	0.38 (0.01)	**4.88** (0.13)
Lasso	0.27 (0.01)	0.32 (0.01)	7.08 (0.28)
Ridge	0.34 (0.01)	0.50 (0.01)	6.13 (0.16)
Elastic net	0.28 (0.01)	0.33 (0.01)	6.69 (0.21)
OLS	0.34 (0.01)	0.53 (0.01)	26.92 (1.18)

Scenario 2			
r	0	0.5	0.99
LoRI	*0.33* (0.01)	*0.46* (0.02)	4.64 (0.16)
Zellner-Siow	0.47 (0.02)	1.20 (0.04)	8.30 (0.31)
Unit Info	0.57 (0.02)	1.41 (0.03)	8.15 (0.36)
Horseshoe	0.44 (0.02)	0.66 (0.02)	4.71 (0.22)
Bayesian lasso	0.42 (0.01)	0.51 (0.02)	1.76 (0.08)
Lasso	*0.33* (0.01)	0.52 (0.02)	7.44 (0.14)
Ridge	**0.32** (0.01)	**0.35** (0.01)	*0.27* (0.02)
Elastic net	0.34 (0.01)	0.49 (0.01)	**0.04** (0.00)
OLS	*0.33* (0.01)	0.52 (0.02)	27.83 (1.05)

inherent instability from nearly singular designs. Notably, LoRI remains the second best in this highly correlated case.

In Scenario 2, the true model is non-sparse while all true coefficients are small. When $r = 0$ or 0.5, ridge regression outperforms all other methods, which is generally consistent with the comparison between the lasso and ridge regression. Similar to Scenario 1, LoRI yields the second smallest estimation error by slightly underperforming compared to ridge regression. Interestingly, when $r = 0.99$, the elastic net becomes the best method. One possible reason is that the L_2 penalty dominates and thus the elastic net becomes a soft thresholding method on univariate regression coefficients, which ignores the dependence among predictors. Since the oracle is the full model, pure shrinkage methods solely based on the full model, such as the Bayesian lasso and the horseshoe, perform more accurately than the methods which average all sub-models, such as the Zellner-Siow and the unit information prior. In addition, these two g-prior methods also show instability when predictors are highly correlated. We notice that although LoRI also averages all sub-models, it generally performs more accurately than the pure shrinkage methods.

Table 3.5: Simulation study of regression: marginal inclusion probability of LoRI for $r = 0.5$, averaged over 500 simulations.

Scenario 1								
β_j	3	1.5	0	0	2	0	0	0
$P(\beta_j \neq 0 \mid \mathbf{Y})$	1.00	1.00	0.22	0.22	1.00	0.22	0.21	0.21
Scenario 2								
β_j	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
$P(\beta_j \neq 0 \mid \mathbf{Y})$	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99

LoRI systematically performs with accuracy in both scenarios, which makes it a good default method, since in real word applications, people usually lack the knowledge about true sparsity. To better understand the contribution of the positive mass at zero component in LoRI, we compare the posterior marginal inclusion probability in the above two different scenarios (see Table 3.5). In LoRI, although the continuous component with high concentration around zero leads to severe shrinkage, point mass at zero component alone sets the coefficient to exact zero, or equivalently, excludes the corresponding predictor. In the sparse scenario, predictors in the true model are 100% included, while the rest predictors only have about 20% inclusion probabilities. In the non-sparse scenario, all predictors are included 99% of the time. The DP

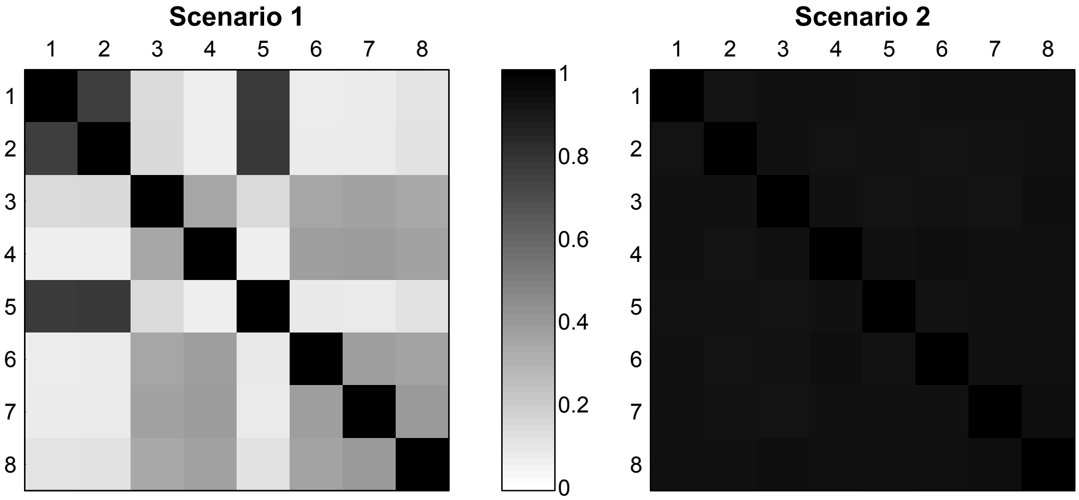


FIGURE 3.3: Simulation study of regression: pairwise posterior probability of being assigned to the same group by LoRI. For $r = 0.5$, averaged over 500 simulations. Values in diagonal cells equal 1.

hierarchical prior in LoRI introduces more flexibility than any parametric prior, and thus allows LoRI's point mass at zero component to adapt to different sparsity levels. If a variable selection procedure is of interest, under the median probability model (Barbieri and Berger, 2004) that includes the variables whose posterior marginal inclusion probabilities exceed 0.5, LoRI selects the correct models in both sparse and non-sparse scenarios.

In addition, grouping structures among coefficients induced by the DP in LoRI correspond to the characteristics of true coefficients. In Figure 3.3, we use heat maps to represent pairwise posterior probability of β_i and β_j in the same group, for $i \neq j$. In Scenario 1, eight coefficients seem to be divided into two groups $\{\beta_1, \beta_2, \beta_5\}$ and $\{\beta_3, \beta_4, \beta_6, \beta_7, \beta_8\}$, which are consistent with the covariates to be included and excluded according to the true model. Coefficients within the same group have a higher probability of being selected together than coefficients between groups. In

Scenario 2, LoRI assigns all coefficients to the same group, which suggests that in non-sparse scenario LoRI performs similarly as rotation invariant methods such as ridge regression. Notice that the correlation structures of design matrices **X** are the same across different scenarios. Therefore, the difference in LoRI's grouping reflects the structures of coefficients rather than correlations. This property of LoRI further differentiates it from the group lasso.

3.5 Protein Activity Example

We apply LoRI to a protein activity dataset, which has been previously studied by Clyde and Parmigiani (1998) and Clyde et al. (2011). This dataset was collected from a well-designed experiment studying the relationship between the protein activity level and different factors of storage conditions as well as their two-way interactions. This dataset consists of $n = 96$ observations and $p = 88$ potential exploratory variables. The heat map of the correlation matrix among predictors (Figure 3.4) suggests that some of the variables are highly correlated. 348 pairs (9.1%) of predictors have absolute correlations larger than 0.5, and among them, 19 pairs have absolute correlations larger than 0.95.

We apply LoRI on this dataset and report the posterior mean estimate and marginal posterior inclusion probability for each dimension in Figure 3.5. By averaging all sub-models, the six variables that have the largest absolute values of posterior mean are the main effects protein concentration (con), two detergent levels (detT, detN), two-way interactions between buffer and temperature (bufPO4.temp), buffer and detergent (bufTRS.detN), concentration and detergent (con.detT). These predictors also have the highest posterior inclusion probabilities, and thus form the median probability model. The grouping pattern (Figure 3.6) of LoRI for this dataset seems similar to the one in Section 3.4.2, since these six predictors are more likely to form their own group rather other than join the other 82 predictors.

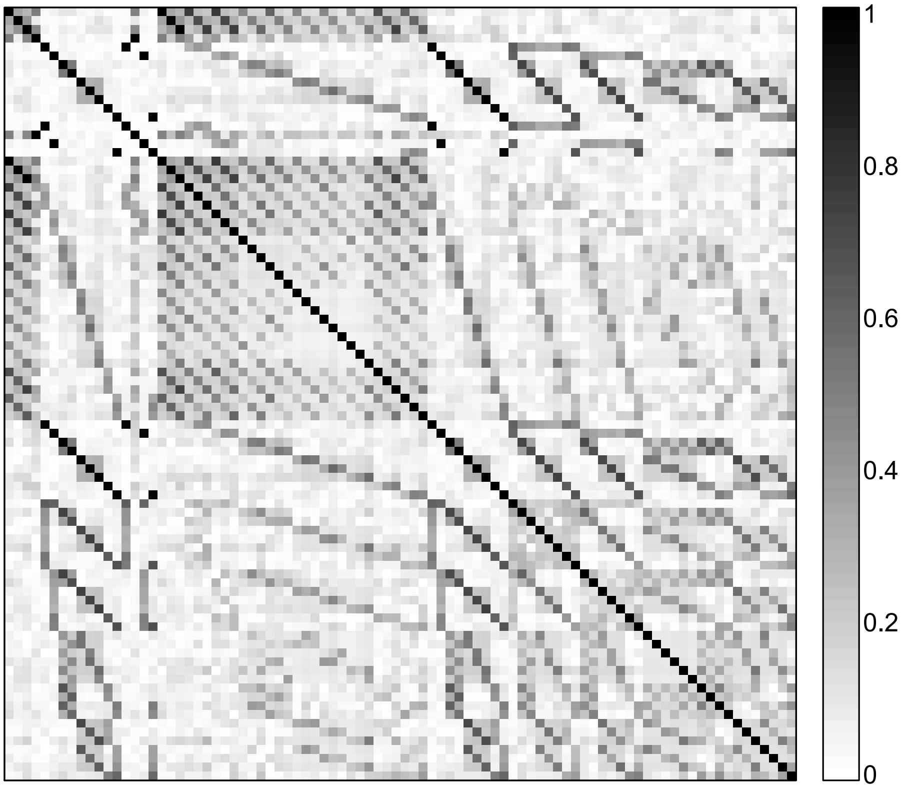


FIGURE 3.4: Protein activity data: heat map of correlation matrix among predictors. The (i, j) cell corresponds to the correlation between the i-th and the j-th predictor, $1 \leqslant i, j \leqslant 88$. Diagonal cells have value 1.

We also compare LoRI with all other methods mentioned in Section 3.4.2. To assess the prediction accuracy across different methods, we conduct leave-one-out cross validation. For each observation i, we put it aside and use the other 95 observations in the dataset to estimate the regression parameters and obtain a predicted value $\tilde{y}_{(i)}$ for the i-th observation. The RootMSE (Table 3.6), squared root of mean squared prediction error, are computed to measure prediction accuracy:

$$\text{RootMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{y}_{(i)}\right)^2}.$$

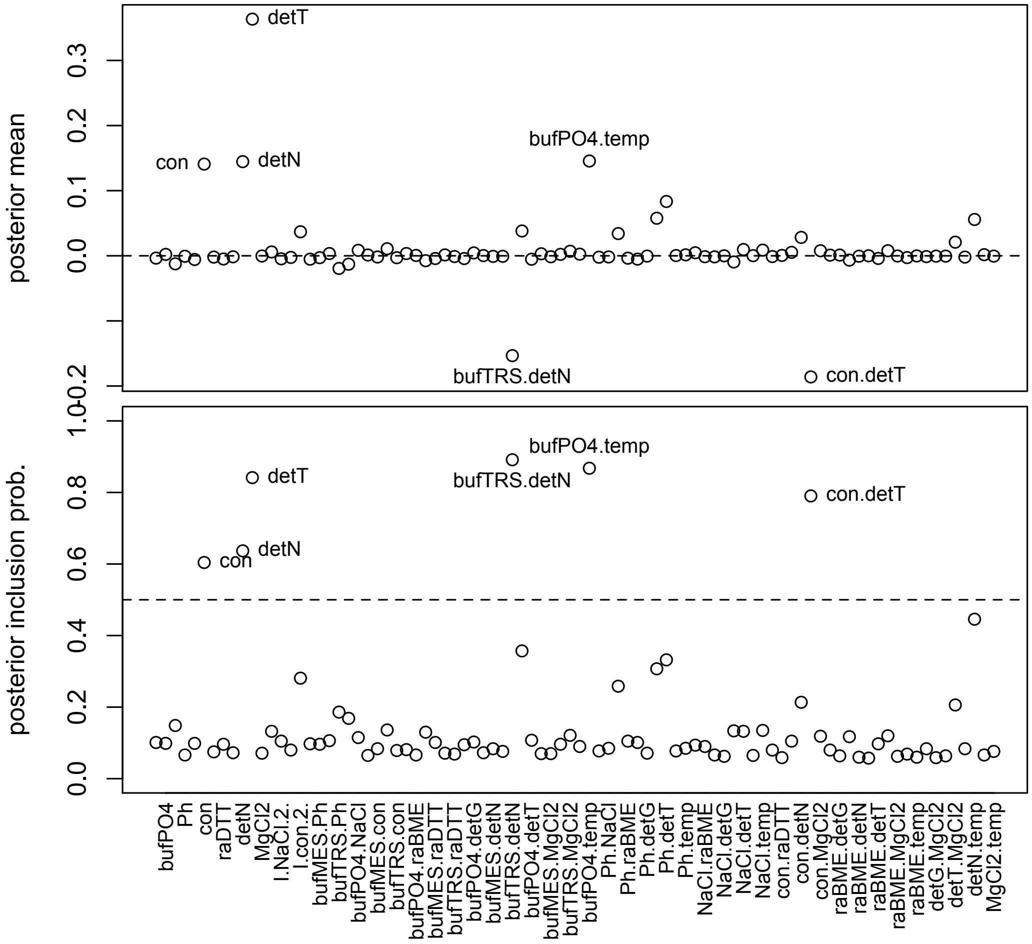


FIGURE 3.5: Applying LoRI on protein activity dataset; from upper to lower: posterior means of coefficients, marginal posterior inclusion probability.

Due to the high correlation among variables, predictions from g-prior methods and the lasso are not as reliable as those from other Bayesian shrinkage methods along with ridge regression. Notably, LoRI yields the smallest prediction error, which confirms LoRI as an ideal option of default approach.

3.6 Discussion

We have proposed LoRI as a novel semi-parametric shrinkage prior for Bayesian model averaging and recommended it as a default method. Both simulation and real

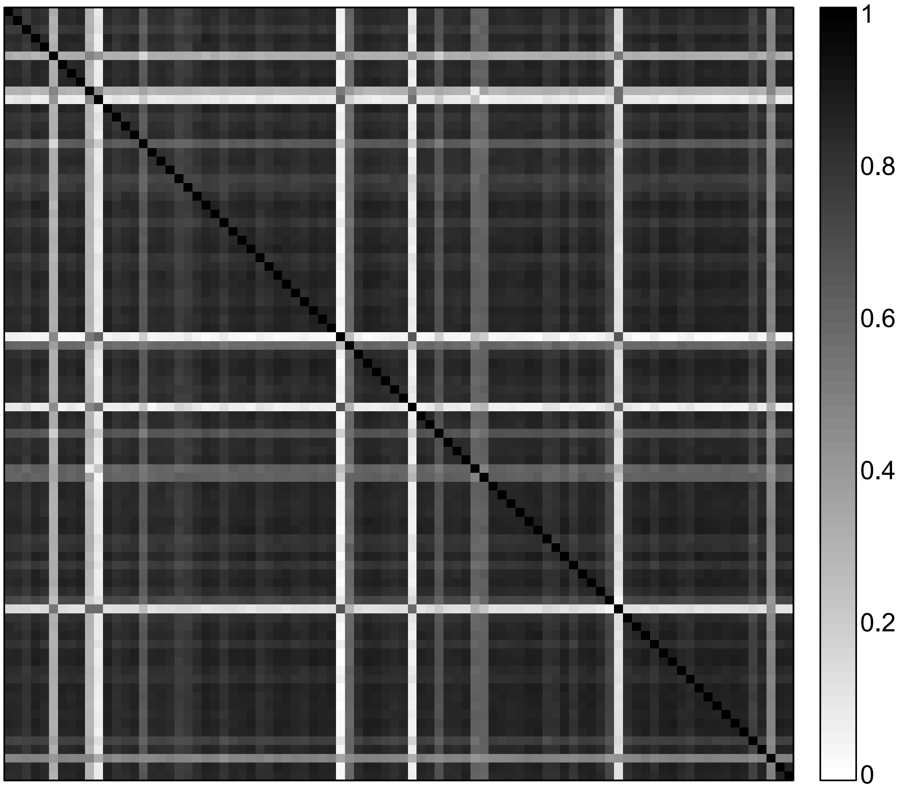


FIGURE 3.6: Applying LoRI on protein activity dataset: heat map of pairwise posterior probability of being assigned to the same group. The (i, j) cell corresponds to the pairwise posterior between the i-th and the j-th predictor, $1 \leqslant i, j \leqslant 88$.

examples show that LoRI adapts to model sparsity, values of true coefficients as well as correlation structures among predictors, and yields accurate parameter estimation and prediction. Thanks to its Dirichlet Process hyperprior, LoRI exhibits flexibility as well as yields groupings. When the true model is sparse, LoRI performs similar to independent mixtures of Cauchy priors and its point masses at zeros components further contribute to sparse solutions. LoRI's bounded prior influence allows it to preserve large coefficients. When the true model is non-sparse, LoRI groups most variables together and performs similar to a rotation invariant prior.

Table 3.6: Protein activity dataset: prediction errors measured by RootMSE. Column-wise smallest error is in bold type.

	RootMSE
LoRI	**0.484**
Zellner-Siow	0.646
Unif Info	0.552
Horseshoe	0.494
Bayesian lasso	0.499
Lasso	0.547
Ridge	0.502
Elastic net	0.507
OLS	1.743

In this article, our focus lies on model averaging rather than model selection, since utilizing information contained in all sub-models can avoid bias and lead to more accurate predictions, as well as measures of uncertainty. However, if model selection is of interest, the point mass at zero component in LoRI provides coherent model selection procedures related to optimizing certain loss functions. For example, we have shown in the simulation study that by using the posterior median estimator, which minimizes the L_1 loss, LoRI can recover the true model. One area of future research is to propose detailed selection rules for LoRI, and assess their performance in the framework of model selection.

4

Discussion

We have developed two new hierarchical prior distributions for normal linear regression and Generalized Linear Models (GLMs) respectively. Both of them have positive probabilities at zero for each coefficient, which are capable of yielding sparse solutions under valid variable selection criteria. They both can be considered as scale mixtures of normal distributions with heavy tails that are robust to large signals in coefficients while accommodating many zero coefficients. For the LoRI prior, we incorporate a non-parametric hyper prior, through a Dirichlet process prior to gain extra flexibility. The essential discreteness of the DP prior reveals group structure among predictors, and thus makes LoRI adaptive to datasets with different densities of sparsity. For the CH-g prior used in GLMs, we assign a generalized Beta distribution as hyper prior, which is very flexible to encompass most conventional hyper priors on g in mixtures of g-priors.

4.1 Future Directions

We think a major difference between LoRI and the CH-g prior is the incorporation of the correlation structure among predictors in the prior dependence of coefficients.

Recall that LoRI prior is a scale mixtures of independent normals. Each predictor is standardized so that the prior is invariant to scale and location transformations of the predictors but not more general linear transformations. On the other hand, the CH-g prior is based on the g-prior, whose prior precision matrix is proportional to $\mathbf{X}_{\mathcal{M}}^T\mathbf{X}_{\mathcal{M}}$ in normal linear regression, or the information matrix in GLM. Although this setting automatically adjusts for linear transformations of the design matrix, estimation may suffer greatly of coefficients under g-prior and mixtures of g-priors with nearly singular design matrices. This issue is verified empirically in the simulation example in Section 3.4.2. In contrast, with independent predictors, g-prior variants yield smaller estimation error than methods in the independent scale mixtures of normals family. One of our future directions is to conduct an in-depth comparison between these two types of model selection priors, to obtain a better understanding of their strength and weakness when being applied to different types of problems.

We also plan to extend LoRI prior to GLMs. Based on its ideal empirical performance in selection and prediction in linear models, we are interested to learn if it will become a good model selection prior for questions with binary or categorical responses. Note that LoRI prior is not conjugate for a normal likelihood. When extending it to GLMs, the computational expense will almost remain the same since an almost identical MCMC algorithm can be applied.

Appendix A

Appendix for Chapter 2

A.1 Confluent Hypergeometric Function

${}_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)} \int_0^1 z^{a-1}(1-z)^{b-a-1} \exp(sz)dz$ is the confluent hypergeometric function (Abramowitz and Stegun, 1970), for $a > 0$ and $b > 0$. Since Gamma function $\Gamma(x)$ does not converge for non-positive integer x, here we assume $b - a > 0$.

A.2 Proofs

A.2.1 Proof to Remark 1

Proof. Without loss of generality, we assume for first $p_{\mathcal{M}_T}$ columns of $\mathbf{X}_{\mathcal{M}}$ forms $\mathbf{X}_{\mathcal{M}_T}$. In addition, $\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T}) = \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) = \mathbf{I}_n$ identity matrix. Because

$$\begin{aligned}\mathrm{BF}_{\mathcal{M}_T:\mathcal{M}} &= \frac{f_{\mathcal{M}_T}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_T}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{f_{\mathcal{M}}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})} (1+g)^{\frac{p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}} \exp\left\{\frac{\mathrm{RSS}_{\mathcal{M}} - \mathrm{RSS}_{\mathcal{M}_T}}{2(1+g)}\right\} \\ &= (1+g)^{\frac{p_{\mathcal{M}} - p_{\mathcal{M}_T}}{2}} \left[\frac{f_{\mathcal{M}_T}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}_T}, \hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{f_{\mathcal{M}}(\mathbf{Y}|\hat{\alpha}_{\mathcal{M}}, \hat{\boldsymbol{\beta}}_{\mathcal{M}})}\right]^{\frac{g}{1+g}}\end{aligned}$$

can be written as a function of the maximized likelihood ratio, which is in the order of $O(1)$ in the case of $\mathcal{M}_T \subset \mathcal{M}$. Therefore, the Bayes factor is also in the order of

$O(1)$. □

A.2.2 Proof to Proposition 2

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

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

We will show that as $\|\boldsymbol{\beta}_{\mathcal{M}}\|$, or equivalently, $\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{I}_n} \longrightarrow \infty$, both the lower bound and upper bound of (A.1) are proportional to $\left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}\right)^{-\frac{b+p_{\mathcal{M}}}{2}}$.

We first find a lower bound of (A.1) up to a constant. Since $s \geqslant 0$,

$$\begin{aligned}
\text{(A.1)} &\geqslant \int_0^\infty g^{-\frac{p_{\mathcal{M}}}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2g}\right] g^{\frac{a}{2}-1}\left(\frac{1}{1+g}\right)^{\frac{a+b}{2}} dg \\
&= \int_0^\infty \left(\frac{g}{1+g}\right)^{\frac{a+b}{2}} \left(\frac{1}{g}\right)^{\frac{b+p_{\mathcal{M}}-2}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2g}\right] d\left(\frac{1}{g}\right),
\end{aligned}$$

then according to the Watson's Lemma (see (Olver, 1997), p71), the limit of the lower bound

$$\begin{aligned}
&\lim_{\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{I}_n} \to \infty} \int_0^\infty \left(\frac{g}{1+g}\right)^{\frac{a+b}{2}} \left(\frac{1}{g}\right)^{\frac{b+p_{\mathcal{M}}-2}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2g}\right] d\left(\frac{1}{g}\right) \\
&\qquad \propto \left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}\right)^{-\frac{b+p_{\mathcal{M}}}{2}}
\end{aligned}$$

Next we will find an upper bound.

$$\begin{aligned}
\text{(A.1)} &\leqslant \int_0^\infty g^{-\frac{p_{\mathcal{M}}}{2}} \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2(1+g)}\right] g^{\frac{a}{2}-1}\left(\frac{1}{1+g}\right)^{\frac{a+b}{2}} \exp\left[\frac{sg}{2(1+g)}\right] dg \\
&= e^{-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2}} \int_0^1 \left(\frac{g}{1+g}\right)^{\frac{a-p_{\mathcal{M}}-2}{2}} \left(\frac{1}{1+g}\right)^{\frac{b+p_{\mathcal{M}}-2}{2}} e^{\frac{(s+\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n})g}{2(1+g)}} d\left(\frac{g}{1+g}\right) \\
&= e^{-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2}} B\left(\frac{a-p_{\mathcal{M}}}{2}, \frac{b+p_{\mathcal{M}}}{2}\right) {}_1F_1\left(\frac{a-p_{\mathcal{M}}}{2}, \frac{a+b}{2}, \frac{s+\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2}\right)
\end{aligned}$$

According to (Abramowitz and Stegun, 1970) formula (13.1.4), the limit behavior of ${}_1F_1(a, b, s)$ function for large positive s is

$$ {}_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(a)} \exp(s) s^{a-b} [1 + O(|s|^{-1})], \text{ when } s > 0. \tag{A.2} $$

Hence the limit of the upper bound

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

$$ = \exp\left[-\frac{\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2} \right] \Gamma\left(\frac{b + p_{\mathcal{M}}}{2} \right) \exp\left[\frac{s + \|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2} \right] \cdot \left(\frac{s + \|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n}}{2} \right)^{-\frac{b + p_{\mathcal{M}}}{2}} $$

$$ \propto \left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n} \right)^{-\frac{b + p_{\mathcal{M}}}{2}} $$

Therefore, as $\|\boldsymbol{\beta}_{\mathcal{M}}\|_{\mathcal{I}_n}$ 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 $\left(\|\boldsymbol{\beta}_{\mathcal{M}}\|^2_{\mathcal{I}_n} \right)^{-\frac{b + p_{\mathcal{M}}}{2}}$. □

A.2.3 Proof to Lemma 1

Proof. Note that for any model $\mathcal{M}$, $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) = \mathbf{1}_n^T \, \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \, \mathbf{1}_n$ equals the first diagonal element of $[\mathbf{1}_n, \mathbf{X}]^T \, \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \, [\mathbf{1}_n, \mathbf{X}]$. According to the Assumption 1, there exists a positive constant $c_{\mathcal{M}}$ such that

$$ \text{plim}_{n \to \infty} \frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})}{n} = c_{\mathcal{M}} $$

Therefore, we have

$$ \text{plim}_{n \to \infty} \left[\frac{1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T}) nc}{1 + \mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc} \right]^{-\frac{1}{2}} e^{-\frac{1}{2} \left[\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T}) \hat{\alpha}^2_{\mathcal{M}_T}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T}) nc + 1} - \frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) \hat{\alpha}^2_{\mathcal{M}}}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) nc + 1} \right]} = \left(\frac{c_{\mathcal{M}_T}}{c_{\mathcal{M}}} \right)^{-\frac{1}{2}} < \infty, $$

and

$$ \text{plim}_{n \to \infty} \left[\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}_T})}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})} \right]^{-\frac{1}{2}} = \left(\frac{c_{\mathcal{M}_T}}{c_{\mathcal{M}}} \right)^{-\frac{1}{2}} < \infty $$

Hence the asymptotic behaviors of both $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ and $\Lambda^{c=\infty}_{\mathcal{M}_T:\mathcal{M}}$ are dominated by the likelihood ratio. Next we will study the asymptotic property of the likelihood ratio in two difference cases: 1) $\mathcal{M}_T \subset \mathcal{M}$ and 2) $\mathcal{M}_T \not\subset \mathcal{M}$.

In the first case where $\mathcal{M}_T \subset \mathcal{M}$, from the well-known results of likelihood ratio test, the logarithm of ratio of maximized likelihoods has a central chi-square distribution $\chi^2_{p_{\mathcal{M}}-p_{\mathcal{M}_T}}$. This suggests that the log-likelihood ratio does not depend on n, i.e., $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O(1)$.

In the second case where $\mathcal{M}_T \not\subset \mathcal{M}$, we first examine the sub-case where $\mathcal{M} \subset \mathcal{M}_T$. Without loss of generality, we assume the space spanned by the first $p_{\mathcal{M}}$ columns of $\mathbf{X}_{\mathcal{M}_T}$ and $\mathbf{1}_n$ equals the space spanned by $\mathbf{X}_{\mathcal{M}}$ and $\mathbf{1}_n$, i.e.,

$$\mathcal{C}(\mathbf{1}_n, \mathbf{X}_{1,\mathcal{M}_T}, \ldots, \mathbf{X}_{p_{\mathcal{M}},\mathcal{M}_T}) = \mathcal{C}(\mathbf{1}_n, \mathbf{X}_{\mathcal{M}})$$

For notation simplicity, we denote the parameters as $\psi_{\mathcal{M}} = (\alpha_{\mathcal{M}}, \boldsymbol{\beta}_{\mathcal{M}})$, the log-likelihood as $l_{\mathcal{M}}(\psi_{\mathcal{M}}) = \log f_{\mathcal{M}}(\mathbf{Y} \mid \psi_{\mathcal{M}})$, and the i-th row of the original design matrix $[\mathbf{1}_n, \mathbf{V}_{\mathcal{M}}]$ as $\mathbf{v}_{i,\mathcal{M}}$. According to the power calculation results for GLM in (Self et al., 1992), when testing nested models, if the larger model is true, then we have that the logarithm of likelihood ratio converges in distribution to a non-central χ^2, that is

$$\Lambda_{\mathcal{M}_T:\mathcal{M}} \xrightarrow{d} \exp\{\chi^2_{p_{\mathcal{M}_T}-p_{\mathcal{M}}}(p_{\mathcal{M}} + 1 - \mathrm{tr}(\mathbf{M}_1^{-1}\mathbf{M}_2) + \Psi)\}, \tag{A.3}$$

where $\chi^2_k(m)$ is a non-central χ^2 distribution with degrees of freedom k and non-centrality parameter m; and

$$
\begin{aligned}
\mathbf{M}_1 &= \mathbb{E}\left[-\frac{\partial^2 l_{\mathcal{M}_T}(\psi_{\mathcal{M}_T})}{\partial(\psi_{\mathcal{M}})^2}\bigg|_{(\psi^*_{\mathcal{M}},\mathbf{0})} \right] \\
&= \sum_{i=1}^{n} a_i^{-1}(\phi) \left\{ b''(\theta^*_{i,\mathcal{M}}) \left(\frac{\partial \theta^*_{i,\mathcal{M}}}{\partial \eta^*_{i,\mathcal{M}}} \right)^2 - [b'(\theta^*_{i,\mathcal{M}_T}) - b'(\theta^*_{i,\mathcal{M}})] \frac{\partial^2 \theta^*_{i,\mathcal{M}}}{\partial \eta^{*2}_{i,\mathcal{M}}} \right\} \mathbf{v}_{i,\mathcal{M}} \mathbf{v}^T_{i,\mathcal{M}} \\
\mathbf{M}_2 &= \mathbb{E}\left[\left(\frac{\partial l_{\mathcal{M}_T}(\psi_{\mathcal{M}_T})}{\partial\ \psi_{\mathcal{M}}} \right) \left(\frac{\partial l_{\mathcal{M}_T}(\psi_{\mathcal{M}_T})}{\partial\ \psi_{\mathcal{M}}} \right)^T \bigg|_{(\psi^*_{\mathcal{M}},\mathbf{0})} \right] \\
&= \sum_{i=1}^{n} a_i^{-1}(\phi) b''(\theta^*_{i,\mathcal{M}_T}) \left(\frac{\partial \theta^*_{i,\mathcal{M}}}{\partial \eta^*_{i,\mathcal{M}}} \right)^2 \sum_{j=p_{\mathcal{M}}+1}^{p_{\mathcal{M}_T}} \mathbf{v}_{i,\mathcal{M}} \mathbf{v}^T_{i,\mathcal{M}} \\
\Psi &= \sum_{i=1}^{n} a_i^{-1}(\phi) \left\{ b'(\eta^*_{i,\mathcal{M}_T}) \left[\eta^*_{i,\mathcal{M}_T} - \eta^*_{i,\mathcal{M}} \right] - \left[b(\eta^*_{i,\mathcal{M}_T}) - b(\eta^*_{i,\mathcal{M}}) \right] \right\}
\end{aligned}
$$

where the expectation in $\mathbf{M}_1$ and $\mathbf{M}_2$ are taken with respect to the true parameters $\psi^*_{\mathcal{M}_T} = (\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T})$; $\theta^*_{i,\mathcal{M}}$ is the i-th canonical parameter and $\eta^*_{i,\mathcal{M}}$ is the i-th linear predictor under parameters $\psi^*_{\mathcal{M}}$ in model $\mathcal{M}$. Both (Self et al., 1992) and (Shieh, 2000) point out that empirical experience suggests that $\mathrm{tr}(\mathbf{M}_1^{-1}\mathbf{M}_2)$ is very close to $p_{\mathcal{M}} + 1$. Furthermore, if we treat the explanatory variables as random independent samples, then we can easily show that $\mathrm{tr}(\mathbf{M_1^{-1}M_2})$ only depends on $p_{\mathcal{M}}$, not n. Therefore, we can say that the non-centrality parameter in (A.3) $p_{\mathcal{M}} + 1 - \mathrm{tr}(\mathbf{M}_1^{-1}\mathbf{M}_2) + \Psi$ is dominated by Ψ. Because $\mathcal{M} \not\subset \mathcal{M}_T$, which means the limits of parameter in $\mathcal{M}$ do not equal the true parameters, i.e., $(\alpha^*_{\mathcal{M}}, \boldsymbol{\beta}^*_{\mathcal{M}}, \mathbf{0}) \neq (\alpha^*_{\mathcal{M}_T}, \boldsymbol{\beta}^*_{\mathcal{M}_T})$, it is reasonable to assume that $\lim_{n\to\infty} \Psi/n$ converges to a constant c. Since the non-centrality parameter in a χ^2 distribution must be positive, this limit c must be positive, which implies that $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O(e^{cn})$.

In the case where the two models $\mathcal{M}$ and $\mathcal{M}_T$ are not nested, we introduce a third model $\mathcal{M}'$ which includes all the predictors in both $\mathcal{M}$ and $\mathcal{M}_T$. Notice that using a similar method as in (Self et al., 1992), we can easily show that $\Lambda_{\mathcal{M}':\mathcal{M}}$ also has an 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: $\Lambda_{\mathcal{M}_T:\mathcal{M}'} = O(1)$ and $\Lambda_{\mathcal{M}':\mathcal{M}} = O\left(e^{cn}\right)$. Therefore, we can conclude that $\Lambda_{\mathcal{M}:\mathcal{M}} = O\left(e^{cn}\right)$ in this case. □

A.2.4 Proof to Lemma 2

Proof. We will show the asymptotic results about the RSS for GLM in two steps: under model $\mathcal{M}_T$ and under model $\mathcal{M}$. According to our Remark 2, under $\mathcal{M}_T$, the MLE estimator $\hat{\boldsymbol{\beta}}_{\mathcal{M}_T}$ converges in probability to the true coefficients $\boldsymbol{\beta}^*_{\mathcal{M}_T}$ as n increases. Furthermore, there exists the asymptotic normality,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_T} - \boldsymbol{\beta}^*_{\mathcal{M}_T}) \xrightarrow{\text{d}} \mathrm{N}\left(\mathbf{0}, [\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{M}_T})]^{-1}\right),$$

where $\mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{M}_T})$ is the expected information matrix based on a single observation evaluated at the true parameters. Note the columns of $\mathbf{X}_{\mathcal{M}_T}$ are in the space spanned by $\mathcal{C}(\mathbf{1}_n, \mathbf{X})$; so based on the Assumption 1,

$$\frac{\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{n} = \frac{\mathbf{X}^T_{\mathcal{M}_T}\ \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\ \mathbf{X}_{\mathcal{M}_T}}{n}$$

converges to a positive definite matrix in probability. The consistency of MLE suggests that this limit is

$$\mathrm{plim}_{n\to\infty}\frac{\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}_T})}{n} = \mathcal{I}(\boldsymbol{\beta}^*_{\mathcal{M}_T})$$

We apply Slutsky's theorem to rewrite the asymptotic normality as

$$\left[\mathbf{X}^T_{\mathcal{M}_T}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\mathbf{X}_{\mathcal{M}_T}\right]^{\frac{1}{2}}\hat{\boldsymbol{\beta}}_{\mathcal{M}_T} - \left[\mathbf{X}^T_{\mathcal{M}_T}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\mathbf{X}_{\mathcal{M}_T}\right]^{\frac{1}{2}}\boldsymbol{\beta}^*_{\mathcal{M}_T} \xrightarrow{\text{d}} \mathrm{N}\left(\mathbf{0}, \mathbf{I}_{p_{\mathcal{M}_T}}\right),$$

Therefore, the RSS for GLM under the true model

$$Q_{\mathcal{M}_T} = \hat{\boldsymbol{\beta}}^T_{\mathcal{M}_T}\left[\mathbf{X}^T_{\mathcal{M}_T}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\mathbf{X}_{\mathcal{M}_T}\right]\hat{\boldsymbol{\beta}}_{\mathcal{M}_T}$$

has a non-central χ^2 distribution with degrees of freedom $p_{\mathcal{M}_T}$, and non-centrality parameter $\boldsymbol{\beta}^{*T}_{\mathcal{M}_T}\left[\mathbf{X}^T_{\mathcal{M}_T}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\mathbf{X}_{\mathcal{M}_T}\right]\boldsymbol{\beta}^{*}_{\mathcal{M}_T}$. Its expectation is

$$\mathbb{E}(Q_{\mathcal{M}_T}) = p_{\mathcal{M}_T} + \boldsymbol{\beta}^{*T}_{\mathcal{M}_T}\left[\mathbf{X}^T_{\mathcal{M}_T}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}_T})\mathbf{X}_{\mathcal{M}_T}\right]\boldsymbol{\beta}^{*}_{\mathcal{M}_T},$$

so if the true coefficient $\boldsymbol{\beta}^{*}_{\mathcal{M}_T} \neq \mathbf{0}$, then the non-centrality parameter increases in the order of $O(n)$. On the other hand, the non-centrality parameter equals 0 if and only if $\boldsymbol{\beta}^{*}_{\mathcal{M}_T} = \mathbf{0}$, which only occurs under the null model $\mathcal{M}_{\emptyset}$. Therefore, we have the asymptotic behavior of $Q_{\mathcal{M}_T}$, that is, if $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$, then $Q_{\mathcal{M}_T} = O(n)$; if $\mathcal{M}_T = \mathcal{M}_{\emptyset}$, then $Q_{\mathcal{M}_T} = O(1)$.

For any model $\mathcal{M} \neq \mathcal{M}_T$, according to the asymptotic properties of the M-estimators (see (van der Vaart, 2000) Chapter 5), there also exists a limit of the MLE $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ and similar asymptotic normality results: $\hat{\boldsymbol{\beta}}_{\mathcal{M}} \xrightarrow{\text{P}} \boldsymbol{\beta}^{*}_{\mathcal{M}}$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*}) \xrightarrow{\text{d}}$$

$$\text{N}\left(\mathbf{0}, \left\{\mathbb{E}\left[\frac{\partial^2 l_{\mathcal{M}}}{\partial \boldsymbol{\beta}^2_{\mathcal{M}}}\Big|_{\boldsymbol{\beta}^{*}_{\mathcal{M}}}\right]\right\}^{-1}\mathbb{E}\left[\left(\frac{\partial l_{\mathcal{M}}}{\partial \boldsymbol{\beta}_{\mathcal{M}}}\right)\left(\frac{\partial l_{\mathcal{M}}}{\partial \boldsymbol{\beta}_{\mathcal{M}}}\right)^T\Big|_{\boldsymbol{\beta}^{*}_{\mathcal{M}}}\right]\left\{\mathbb{E}\left[\frac{\partial^2 l_{\mathcal{M}}}{\partial \boldsymbol{\beta}^2_{\mathcal{M}}}\Big|_{\boldsymbol{\beta}^{*}_{\mathcal{M}}}\right]\right\}^{-1}\right),$$

where $l_{\mathcal{M}}(\cdot) = \log f_{\mathcal{M}}(\cdot)$ and all the expectations are taken with respect to the true model $\mathcal{M}_T$ and the true parameters $(\alpha^{*}_{\mathcal{M}_T}, \boldsymbol{\beta}^{*}_{\mathcal{M}_T})$. Hence the above normal precision is not the Fisher's information matrix. To simplify the notification, we denote the above covariance matrix as $\mathbf{A}$. It is reasonable to assume that $\mathbf{A}$ is a positive definite. Denote $\boldsymbol{\xi} = \sqrt{n}\mathbf{A}^{-\frac{1}{2}}\hat{\boldsymbol{\beta}}_{\mathcal{M}}$, then

$$\boldsymbol{\xi} - \sqrt{n}\mathbf{A}^{-\frac{1}{2}}\boldsymbol{\beta}^{*}_{\mathcal{M}} \xrightarrow{\text{d}} \text{N}\left(\mathbf{0}, \mathbf{I}_{p_{\mathcal{M}}}\right),$$

and thus its quadratic form has a χ^2 distribution in the limit

$$\boldsymbol{\xi}^T\boldsymbol{\xi} \xrightarrow{\text{d}} \chi^2_{p_{\mathcal{M}}}\left(n\boldsymbol{\beta}^{*T}_{\mathcal{M}}\mathbf{A}\boldsymbol{\beta}^{*}_{\mathcal{M}}\right)$$

So when $\boldsymbol{\beta}^{*}_{\mathcal{M}} \neq \mathbf{0}$, $\boldsymbol{\xi}^T\boldsymbol{\xi} = O(n)$, and otherwise, $\boldsymbol{\xi}^T\boldsymbol{\xi} = O(1)$. Based on assumption (2.45), $\lim_{n\to\infty}\mathbf{X}^T_{\mathcal{M}}\mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{X}_{\mathcal{M}}/n$ converges to a positive definite matrix $\mathbf{C}$. This results in that the $p_{\mathcal{M}} \times p_{\mathcal{M}}$ matrix $\mathbf{A}^{\frac{1}{2}}\mathbf{C}\mathbf{A}^{\frac{1}{2}}$ is also positive definite. Suppose λ_1 and $\lambda_{p_{\mathcal{M}}}$ are its largest and smallest eigenvalues, then

$$\lambda_{p_{\mathcal{M}}}\boldsymbol{\xi}^T\boldsymbol{\xi} \leqslant Q_{\mathcal{M}} = \hat{\boldsymbol{\beta}}_{\mathcal{M}}^T \left[\mathbf{X}_{\mathcal{M}}^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}})\mathbf{X}_{\mathcal{M}}\right] \hat{\boldsymbol{\beta}}_{\mathcal{M}} = \boldsymbol{\xi}^T \left[\mathbf{A}^{\frac{1}{2}}\mathbf{C}\mathbf{A}^{\frac{1}{2}}\right] \boldsymbol{\xi} \leqslant \lambda_1 \boldsymbol{\xi}^T\boldsymbol{\xi} \qquad \text{(A.4)}$$

Therefore, if $\boldsymbol{\beta}_{\mathcal{M}}^* = \mathbf{0}$, then $Q_{\mathcal{M}} = O(1)$; In particular, when $\mathcal{M}_T = \mathcal{M}_\varnothing$, under any model $\mathcal{M}$, since $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ converges to a vector of zeros, $Q_{\mathcal{M}}$ does not increase with n. On the other hand, if $\boldsymbol{\beta}_{\mathcal{M}}^* \neq \mathbf{0}$, then $Q_{\mathcal{M}} = O(n)$. □

A.2.5 Proof to Theorem 1

Proof. We have shown in Lemma 1 that the asymptotic property of the first term $\Lambda_{\mathcal{M}_T:\mathcal{M}}$ in the approximate Bayes factor under the CH-g prior

$$\mathrm{BF}_{\mathcal{M}_T:\mathcal{M}} = \Lambda_{\mathcal{M}_T:\mathcal{M}} \cdot \Omega_{\mathcal{M}_T:\mathcal{M}}^{\mathrm{CH}} + O(n^{-1})$$

So here we focus on the asymptotic behavior of the second term

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

According to (Abramowitz and Stegun, 1970) formula (13.1.5), the limit of the Confluent Hypergeometric ${}_1F_1(a, b, s)$ function for large $|s|$ when s is negative can be approximated by

$${}_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(b-a)}(-s)^{-a}[1 + O(|s|^{-1})], \text{ when } s < 0. \qquad \text{(A.5)}$$

We will show the asymptotic result about $\Omega_{\mathcal{M}_T:\mathcal{M}}^{\mathrm{CH}}$ in two separate cases: 1) $\mathcal{M}_T \neq \mathcal{M}_\varnothing$ and 2) $\mathcal{M}_T = \mathcal{M}_\varnothing$.

In the first case where $\mathcal{M}_T \neq \mathcal{M}_\varnothing$, then according to Lemma 2, $Q_{\mathcal{M}_T} = O(n)$.

For any other model $\mathcal{M}$, if $Q_{\mathcal{M}} = O(n)$, then

$$\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} = \frac{\Gamma\left(\frac{b+p_{\mathcal{M}_T}}{2}\right)\left(\frac{s+Q_{\mathcal{M}_T}}{2}\right)^{-\frac{b+p_{\mathcal{M}_T}}{2}}(1+O(n^{-1}))}{\Gamma\left(\frac{b+p_{\mathcal{M}}}{2}\right)\left(\frac{s+Q_{\mathcal{M}}}{2}\right)^{-\frac{b+p_{\mathcal{M}}}{2}}(1+O(n^{-1}))} = O\left(n^{-\frac{p_{\mathcal{M}_T}-p_{\mathcal{M}}}{2}}\right)$$

Similarly, if $Q_{\mathcal{M}} = O(1)$, then $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} = O\left(n^{-\frac{p_{\mathcal{M}_T}}{2}}\right)$. Therefore, as to the Bayes factor, we can conclude that if $\mathcal{M}_T \subset \mathcal{M}$, then $p_{\mathcal{M}} > p_{\mathcal{M}_T}$, and

$$\text{BF}_{\mathcal{M}_T:\mathcal{M}} = O(1) \cdot O\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right) \xrightarrow{\text{P}} \infty$$

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

$$\text{BF}_{\mathcal{M}_T:\mathcal{M}} \geqslant O\left(e^{cn}\right) \cdot O\left(n^{-\frac{p_{\mathcal{M}_T}}{2}}\right) \xrightarrow{\text{P}} \infty$$

In the second case where $\mathcal{M}_T = \mathcal{M}_\varnothing$, Lemma 2 suggest both $Q_{\mathcal{M}_T}$ and $Q_{\mathcal{M}}$ are in the same order $O(1)$. In addition, since any model $\mathcal{M} \supset \mathcal{M}_T$, we have both $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O(1)$ and $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} = O(1)$. In this case the Bayes factor $\text{BF}_{\mathcal{M}_T:\mathcal{M}}$ is bounded, which suggests the selection consistency does not hold when $\mathcal{M}_T = \mathcal{M}_\varnothing$.

Additionally, this theorem also holds if we allow a, b, s to be model specific, since it is reasonable to let the hyper parameters depend on $p_{\mathcal{M}}$. In the case where $\mathcal{M}_T = \mathcal{M}_\varnothing$, the formula of $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}}$ does not change. The only difference is that all a, b, s are substituted with $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$. In the case where $\mathcal{M}_T \neq \mathcal{M}_\varnothing$, as long as for all model $\mathcal{M}$, $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$ do no diverge as n increase, then

$$\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} = \frac{B\left(\frac{b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}+b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, -\frac{s_{\mathcal{M}_T}+Q_{\mathcal{M}_T}}{2}\right)}{B\left(\frac{b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}+b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, -\frac{s_{\mathcal{M}}+Q_{\mathcal{M}}}{2}\right)}$$

$$\cdot \frac{B\left(\frac{b_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}+b_{\mathcal{M}}}{2}, -\frac{s_{\mathcal{M}}}{2}\right)}{B\left(\frac{b_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}+b_{\mathcal{M}_T}}{2}, -\frac{s_{\mathcal{M}_T}}{2}\right)}$$

is in the order of $O\left(n^{-\frac{p_{\mathcal{M}_T}+b_{\mathcal{M}_T}-p_{\mathcal{M}}-b_{\mathcal{M}}}{2}}\right)$.

□

A.2.6 Proof to Theorem 2

Proof. We have shown in Lemma 1 that the asymptotic property of the first term in the approximate Bayes factor under the CH-g prior

$$\mathrm{BF}_{\mathcal{M}_T:\mathcal{M}} = \Lambda_{\mathcal{M}_T:\mathcal{M}} \cdot \Omega^{\mathrm{CH}}_{\mathcal{M}_T:\mathcal{M}} + O(n^{-1})$$

So here we focus on the asymptotic behavior of the second term

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

We will show the asymptotic result about $\Omega^{\mathrm{CH}}_{\mathcal{M}_T:\mathcal{M}}$ in two separate cases: 1) $\mathcal{M}_T = \mathcal{M}_\emptyset$ and 2) $\mathcal{M}_T \neq \mathcal{M}_\emptyset$.

If $\mathcal{M}_T = \mathcal{M}_\emptyset$, then Lemma 1 shows $\Lambda_{\mathcal{M}_T:\mathcal{M}} = O(1)$, and Lemma 2 indicates that both $Q_{\mathcal{M}} = O(1)$ and $Q_{\mathcal{M}_T} = O(1)$. According to (Slater, 1960) formula (4.3.3): if b is large, and a, s are bounded, then the limit of ${}_1F_1$ function can be approximated as

$${}_1F_1(a, b, s) = 1 + O(|b|^{-1}). \tag{A.6}$$

Along with the Stirling's Formula

$$\Gamma(n) = e^{-n} n^{n-\frac{1}{2}} (2\pi)^{\frac{1}{2}} (1 + O(n^{-1})), \tag{A.7}$$

we can conclude that

$$\Omega^{\mathrm{CH}}_{\mathcal{M}_T:\mathcal{M}} = \frac{B\left(\frac{b}{2}, \frac{a}{2}\right)\ {}_1F_1\left(\frac{b}{2}, \frac{a+b}{2}, -\frac{s}{2}\right)}{B\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a}{2}\right)\ {}_1F_1\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)}$$
$$\longrightarrow C \cdot \frac{B\left(\frac{b}{2}, \frac{a}{2}\right)}{B\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a}{2}\right)} = C \cdot \frac{\Gamma\left(\frac{a+b+p_{\mathcal{M}}}{2}\right)}{\Gamma\left(\frac{a+b}{2}\right)} = O\left(n^{\frac{p_{\mathcal{M}}}{2}}\right),$$

which means that the Bayes factor

$$\mathrm{BF}_{\mathcal{M}_T:\mathcal{M}} = O(1) \cdot O\left(n^{\frac{p_{\mathcal{M}}}{2}}\right) \xrightarrow{\mathrm{P}} \infty.$$

On the other hand, if $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, then $Q_{\mathcal{M}_T} = O(n)$. According to (Slater, 1960) formulas (4.3.7): if b is large, $s = by$, and a, y are bounded, then

$$ {}_1F_1(a, b, s) = (1-y)^{-a} \left[1 - \frac{a(a+1)}{2b} \left(\frac{y}{1-y} \right)^2 + O(|b|^{-2}) \right]. \tag{A.8} $$

In both cases $Q_{\mathcal{M}} = O(1)$ or $O(n)$,

$$ \begin{aligned} \Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} &= \frac{B\left(\frac{b+p_{\mathcal{M}_T}}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b+p_{\mathcal{M}_T}}{2}, \frac{a+b+p_{\mathcal{M}_T}}{2}, -\frac{s+Q_{\mathcal{M}_T}}{2}\right)}{B\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a}{2}\right) \ {}_1F_1\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, -\frac{s+Q_{\mathcal{M}}}{2}\right)} \\ &\longrightarrow C \cdot \frac{B\left(\frac{b+p_{\mathcal{M}_T}}{2}, \frac{a}{2}\right)}{B\left(\frac{b+p_{\mathcal{M}}}{2}, \frac{a}{2}\right)} = O\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right). \end{aligned} $$

Therefore, as to the Bayes factor, we can conclude that if $\mathcal{M}_T \subset \mathcal{M}$, then $p_{\mathcal{M}} > p_{\mathcal{M}_T}$, and

$$ \text{BF}_{\mathcal{M}_T:\mathcal{M}} = O(1) \cdot O\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right) \xrightarrow{\text{P}} \infty $$

If $\mathcal{M}_T \not\subset \mathcal{M}$, then

$$ \text{BF}_{\mathcal{M}_T:\mathcal{M}} \geqslant O\left(e^{cn}\right) \cdot O\left(n^{-\frac{p_{\mathcal{M}_T}}{2}}\right) \xrightarrow{\text{P}} \infty $$

Additionally, this theorem also holds if we allow a, b, s to be model specific. It is reasonable to let the hyper parameters depend on $p_{\mathcal{M}}$, for example, in the Beta-prime prior on g (Maruyama and George, 2011), $a = n - p_{\mathcal{M}} - 1.5$. In the case of $\mathcal{M}_T = \mathcal{M}_\emptyset$, the formula of $\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}}$ does not change. The only difference is that all a, b, s are substituted with $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$. In the case of $\mathcal{M}_T \neq \mathcal{M}_\emptyset$, as long as for all model $\mathcal{M}$, $b_{\mathcal{M}}, s_{\mathcal{M}}$ do no diverge as n increase, and $a_{\mathcal{M}} = O(n)$, then

$$\Omega^{\text{CH}}_{\mathcal{M}_T:\mathcal{M}} = \frac{B\left(\frac{b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}+b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, -\frac{s_{\mathcal{M}_T}+Q_{\mathcal{M}_T}}{2}\right)}{B\left(\frac{b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}+b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, -\frac{s_{\mathcal{M}}+Q_{\mathcal{M}}}{2}\right)}$$

$$\cdot \frac{B\left(\frac{b_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}+b_{\mathcal{M}}}{2}, -\frac{s_{\mathcal{M}}}{2}\right)}{B\left(\frac{b_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right)\ {}_1F_1\left(\frac{b_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}+b_{\mathcal{M}_T}}{2}, -\frac{s_{\mathcal{M}_T}}{2}\right)}$$

$$\longrightarrow C \cdot \frac{B\left(\frac{b_{\mathcal{M}_T}+p_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right) B\left(\frac{b_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right)}{B\left(\frac{b_{\mathcal{M}}+p_{\mathcal{M}}}{2}, \frac{a_{\mathcal{M}}}{2}\right) B\left(\frac{b_{\mathcal{M}_T}}{2}, \frac{a_{\mathcal{M}_T}}{2}\right)} = O\left(n^{\frac{p_{\mathcal{M}}-p_{\mathcal{M}_T}}{2}}\right).$$

□

A.2.7 Proof to Proposition 3

Proof. We use the characteristic function to show that the degenerate distribution at 1 is the limit distribution of the conditional posterior of $(z \mid \mathbf{Y}, \mathcal{M})$. The characteristic function

$$\phi_z(t) = E\left(e^{itz}\right) \tag{A.9}$$

$$= \int \frac{z^{\frac{a}{2}-1}(1-z)^{\frac{b+p_{\mathcal{M}}}{2}-1} \exp\left[\left(\frac{s+Q_{\mathcal{M}}}{2} + it\right) z\right]}{B(\frac{a}{2}, \frac{b+p_{\mathcal{M}}}{2})\ {}_1F_1(\frac{a}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2})} dz \tag{A.10}$$

$$= \frac{{}_1F_1(\frac{a}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2} + it)}{{}_1F_1(\frac{a}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2})} \tag{A.11}$$

Lemma 2 shows that if $\boldsymbol{\beta}^*_{\mathcal{M}} \neq \mathbf{0}$, then $Q_{\mathcal{M}} = O(n)$. According to (Abramowitz and Stegun, 1970) formula (13.1.4),

$${}_1F_1(a, b, s) = \frac{\Gamma(b)}{\Gamma(a)} \exp(s) s^{a-b}[1 + O(|s|^{-1})], \text{ when } Re(s) > 0. \tag{A.12}$$

the characteristic function

$$\phi_z(t) \longrightarrow \frac{\exp(\frac{s+Q_{\mathcal{M}}}{2} + it) \cdot (\frac{s+Q_{\mathcal{M}}}{2} + it)^{-\frac{b+p_{\mathcal{M}}}{2}}}{\exp(\frac{s+Q_{\mathcal{M}}}{2}) \cdot (\frac{s+Q_{\mathcal{M}}}{2})^{-\frac{b+p_{\mathcal{M}}}{2}}} = \exp(it),$$

where $\exp(it)$ is the characteristic function of the degenerated distribution at 1. □

A.2.8 Proof to Proposition 4

Proof. We also use the characteristic function (A.9) here.

$$\phi_z(t) = \frac{{}_1F_1(\frac{a}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2} + it)}{{}_1F_1(\frac{a}{2}, \frac{a+b+p_{\mathcal{M}}}{2}, \frac{s+Q_{\mathcal{M}}}{2})}$$

Since $a^* > 0$, under model $\mathcal{M}$ if $Q_{\mathcal{M}} = O(1)$, we can use (Slater, 1960) formulas (4.3.6): when a, b are large, and $b - a, s$ are bounded,

$$ {}_1F_1(a, b, s) = \mathrm{e}^s \left[1 + O(|b|^{-1})\right]; \tag{A.13}$$

if $Q_{\mathcal{M}} = O(n)$, we can use (Slater, 1960) formulas (4.3.7): when a, b are large, and $b - a, s/b$ are bounded,

$$ {}_1F_1(a, b, s) = \mathrm{e}^s \left(1 + \frac{s}{b}\right)^a \left[1 + O(|b|^{-1})\right] \tag{A.14}$$

Similarly as in the proof of Proposition 3, we find that $\phi_z(t) \longrightarrow \exp(it)$. □

A.2.9 Proof to Theorem 3

Proof. For notation simplicity, we omit the subscript $\mathcal{M}$ in $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$ where these is no ambiguity and denote $\boldsymbol{\Sigma}_{n,\mathcal{M}} = \left[\mathcal{I}_n(\hat{\boldsymbol{\beta}}_{\mathcal{M}})\right]^{-1} = \left[\mathbf{X}_{\mathcal{M}}^T \mathcal{I}_n(\hat{\boldsymbol{\eta}}_{\mathcal{M}}) \mathbf{X}_{\mathcal{M}}\right]^{-1}$. Then (2.27) and (2.31) can be simplified to

$$\boldsymbol{\beta}_{\mathcal{M}} \mid z, \mathcal{M}, \mathbf{Y} \xrightarrow{\mathrm{d}} \mathrm{N}(z\, \hat{\boldsymbol{\beta}}_{\mathcal{M}},\ z\, \boldsymbol{\Sigma}_{n,\mathcal{M}})$$

$$z \mid \mathcal{M}, \mathbf{Y} \sim \mathrm{CH}\left(\frac{a}{2}, \frac{a + b + p_{\mathcal{M}}}{2}, -\frac{s + Q_{\mathcal{M}}}{2}\right)$$

We will prove this theorem in two steps: 1) $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$ and 2) $\mathcal{M}_T = \mathcal{M}_{\emptyset}$.

When $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$, the model selection consistency holds, so we just need to show the estimation consistency under the true model $\mathcal{M}_T$. According to (2.54) , it is sufficient to focus on the true model. We again use the characteristic function of the posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}}$. Notice that the integrand $e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}}}$ has a bounded modulus, so according to Fubini's Theorem, the two integral can be interchanged.

$$\phi_{\boldsymbol{\beta}_{\mathcal{M}}}(\mathbf{t}) = \int e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}}}\ p(\boldsymbol{\beta}_{\mathcal{M}} \mid \mathcal{M}, \mathbf{Y})\ d\boldsymbol{\beta}_{\mathcal{M}} \tag{A.15}$$

$$= \int e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}}} \left\{ \int p(\boldsymbol{\beta}_{\mathcal{M}} \mid z, \mathcal{M}, \mathbf{Y})\ p(z \mid \mathcal{M}, \mathbf{Y}) dz \right\} d\boldsymbol{\beta}_{\mathcal{M}} \tag{A.16}$$

$$= \int \left\{ \int e^{i\mathbf{t}^T\boldsymbol{\beta}_{\mathcal{M}}}\ p(\boldsymbol{\beta}_{\mathcal{M}} \mid z, \mathcal{M}, \mathbf{Y})\ d\boldsymbol{\beta}_{\mathcal{M}} \right\} p(z \mid \mathcal{M}, \mathbf{Y}) dz \tag{A.17}$$

$$= \int e^{z(i\mathbf{t}^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_{n,\mathcal{M}}\mathbf{t})}\ p(z \mid \mathcal{M}, \mathbf{Y}) dz \tag{A.18}$$

Similar to the proof of Propositions 3, 4, when $Q_{\mathcal{M}} = O(n)$, the limit of (A.18) is

$$\lim_{n\to\infty} \int e^{z(i\mathbf{t}^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_{n,\mathcal{M}}\mathbf{t})}\ p(z \mid \mathcal{M}, \mathbf{Y}) dz = \exp\left(i\mathbf{t}^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_{n,\mathcal{M}}\mathbf{t} \right)$$

Since under the true model $\hat{\boldsymbol{\beta}}_{\mathcal{M}_T} \to \boldsymbol{\beta}^*_{\mathcal{M}_T}$, and $\boldsymbol{\Sigma}_{n,\mathcal{M}} = O(n^{-1}) \to \mathbf{0}$, hence

$$\phi_{\boldsymbol{\beta}_{\mathcal{M}_T}}(\mathbf{t}) \longrightarrow \exp\left(i\mathbf{t}^T\boldsymbol{\beta}^*_{\mathcal{M}_T} \right),$$

which is the characteristic function of a degenerated distribution at $\boldsymbol{\beta}^*_{\mathcal{M}_T}$.

On the other hand, when $\mathcal{M}_T = \mathcal{M}_{\emptyset}$, the model selection consistency does not hold. Hence we need to examine the limit of posterior distribution of $\boldsymbol{\beta}_{\mathcal{M}}$ under all models. Under model $\mathcal{M}$, the MLE of the coefficient $\hat{\boldsymbol{\beta}}_{\mathcal{M}}$ converges to the true parameters $\mathbf{0}$. Since the modulus of (A.18) is bounded by a constant 1, which is integrable if regarded as a function of z, so according to the dominated convergence theorem,

$$\lim_{n\to\infty} \int e^{z(i\mathbf{t}^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}_{n,\mathcal{M}}\mathbf{t})}\ p(z \mid \mathcal{M}, \mathbf{Y}) dz$$

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

Therefore, the posterior of $\boldsymbol{\beta}_{\mathcal{M}}$ under any model converges to $\mathbf{0}$. □

A.2.10 Proof to Theorem 4

Proof. For notation simplicity, we omit the subscript $\mathcal{M}$ in $a_{\mathcal{M}}, b_{\mathcal{M}}, s_{\mathcal{M}}$ where these is no ambiguity. We will show this consistency in two steps: 1) $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$ and 2) $\mathcal{M}_T = \mathcal{M}_{\emptyset}$.

When $\mathcal{M}_T \neq \mathcal{M}_{\emptyset}$, the model selection consistency holds. In this case it is sufficient to prove this consistency under $\mathcal{M}_T$. According to the consistency of the MLE, as $n \to \infty$, $\hat{\boldsymbol{\beta}}_{\mathcal{M}_T}$ converges in probability to the true coefficients $\boldsymbol{\beta}^*_{\mathcal{M}_T}$, and $\hat{\alpha}_{\mathcal{M}_T}$ converges to the true intercept $\alpha^*_{\mathcal{M}_T}$. According to Assumption 1, $\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) = O(n)$. The approximate posterior mean of of $\alpha_{\mathcal{M}_T}$ in (2.26) converges to the MLE

$$\frac{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}})}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) + \frac{1}{nc}}\, \hat{\alpha}_{\mathcal{M}} \longrightarrow \hat{\alpha}_{\mathcal{M}}$$

and its posterior variance converges to zero

$$\frac{1}{\mathcal{I}_n(\hat{\alpha}_{\mathcal{M}}) + \frac{1}{nc}} = O(n^{-1})$$

Hence the estimation of $\alpha_{\mathcal{M}_T}$ is consistent, i.e.

$$\mathrm{plim}_{n\to\infty}\, p(\alpha_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T) = \delta_{\alpha^*_{\mathcal{M}_T}}(\alpha_{\mathcal{M}_T})$$

Similarly, under the flat prior, the posterior of $\alpha_{\mathcal{M}_T}$ also converges to the true value $\alpha^*_{\mathcal{M}_T}$ asymptotically. The proof of Theorem 3 indicates that the estimation of $\boldsymbol{\beta}_{\mathcal{M}_T}$ is also consistent,

$$\mathrm{plim}_{n\to\infty}\, p\left(\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T\right) = \delta_{\boldsymbol{\beta}^*_{\mathcal{M}_T}}(\boldsymbol{\beta}_{\mathcal{M}_T})$$

Therefore, the estimation μ under $\mathcal{M}_T$ is consistent, that is

$$\begin{aligned} \mathrm{plim}_{n\to\infty}\, \mu &= \mathrm{plim}_n\, \mathbb{E}\left[b'(\alpha_{\mathcal{M}_T} + \mathbf{x}^T_{\mathcal{M}_T}\boldsymbol{\beta}_{\mathcal{M}_T} \mid \mathbf{Y}, \mathcal{M}_T)\right] \\ &= b'(\alpha^*_{\mathcal{M}_T} + \mathbf{x}^T_{\mathcal{M}_T}\boldsymbol{\beta}^*_{\mathcal{M}_T}). \end{aligned}$$

Next we discuss in the case where $\mathcal{M}_T = \mathcal{M}_\emptyset$. In this case, the selection consistency does not hold, and a sufficient condition for estimation consistency under BMA is estimation consistency for μ under each model $\mathcal{M}$. Since for each model $\mathcal{M}$, the true model $\mathcal{M}_T$ is nested in it. Hence the MLE of intercept and coefficients under $\mathcal{M}$ converge to the true value $\alpha^*_{\mathcal{M}_T}$ and $\mathbf{0}$. Therefore, similar method that proves the consistency in the previous case is also applicable here. $\square$

Appendix B

Appendix for Chapter 3

B.1 Proof to Theorem 5

Proof. For notation simplicity, we use β to represent β_j and z to represent $\hat{\beta}_j$ in the following proof. Denote $\lambda = \eta^2/\omega$, then prior $p(\beta \mid \eta)$ has two equivalent hierarchical representations

i. latent parameter ω:

$$\beta \mid \omega \sim \mathrm{N}(0, \omega) \tag{B.1}$$

$$\omega \mid \eta \sim \mathrm{IG}(1/2, \eta^2/2) \tag{B.2}$$

ii. latent parameter λ:

$$\beta \mid \eta, \lambda \sim \mathrm{N}(0, \eta^2/\lambda) \tag{B.3}$$

$$\lambda \sim \mathrm{G}(1/2, 1/2) \tag{B.4}$$

In the following proof, we will use the second representation, i.e. transform ω to λ. Without loss of generality, assume $\phi = 1$. We first show that the theorem holds if the prior just contains the continuous component:

$$\tilde{\pi}(\beta) = \int_0^\infty \int_0^\infty \mathrm{N}\left(\beta; 0, \frac{\eta^2}{\lambda}\right) \cdot \mathrm{C}^+\left(\eta; 0, \frac{1}{\sqrt{\phi}}\right) \cdot \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\lambda d\eta.$$

Let $m(z)$ denote the marginal likelihood under prior $\tilde{\pi}(\beta)$:

$$m(z) = \frac{1}{\sqrt{2\pi^3}} \int_0^\infty \int_0^\infty \exp\left(-\frac{z^2/2}{1+\eta^2/\lambda}\right) \frac{1}{(1+\eta^2)\sqrt{1+\eta^2/\lambda}} \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\eta d\lambda$$

Its easy to show that the $m(z) > 0$ for all $z \in \mathbb{R}$. According to (Carvalho et al., 2010) Theorem 2,

$$E(\beta|z) - z = \frac{\mathrm{d}}{\mathrm{d}z} \log m(z). \tag{B.5}$$

According to (Carvalho et al., 2010) proof of Theorem 3,

$$m(z) = \frac{2\sqrt{\lambda}}{\sqrt{2\pi^3}} \exp\left(-\frac{z^2}{2}\right) \int_0^\infty \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{z^2}{2}, 1-\lambda\right) \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\lambda$$

$$\frac{\mathrm{d}}{\mathrm{d}z} m(z) = -\frac{4z\sqrt{\lambda}}{3\sqrt{2\pi^3}} \exp\left(-\frac{z^2}{2}\right) \int_0^\infty \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{z^2}{2}, 1-\lambda\right) \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\lambda$$

Therefore (B.5) becomes:

$$\frac{\mathrm{d}}{\mathrm{d}z} \log m(z) = -\frac{2z \int_0^\infty \Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{z^2}{2}, 1-\lambda\right) \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\lambda}{3 \int_0^\infty \Phi_1\left(\frac{1}{2}, 1, \frac{3}{2}, \frac{z^2}{2}, 1-\lambda\right) \mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\lambda} \tag{B.6}$$

In the numerator, when $0 < \lambda \leqslant 1$,

$$\begin{aligned}
&\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{z^2}{2}, 1-\lambda\right) \\
&= \exp\left(\frac{z^2}{2}\right) \sum_{n=0}^\infty \frac{\left(\frac{1}{2}\right)_n (1)_n}{\left(\frac{5}{2}\right)_n} \frac{(1-\lambda)^n}{n!} \; {}_1F_1\left(2, \frac{5}{2}+n, -\frac{z^2}{2}\right) \\
&= \exp\left(\frac{z^2}{2}\right) \sum_{n=0}^\infty \frac{\left(\frac{1}{2}\right)_n (1)_n}{\left(\frac{5}{2}\right)_n} \frac{(1-\lambda)^n}{n!} \frac{\Gamma(2)}{\Gamma\left(\frac{1}{2}+n\right)} \left(\frac{z^2}{2}\right)^{-2} \left\{1 + O(z^{-2})\right\}
\end{aligned}$$

and when $\lambda > 1$

$$\begin{aligned}
&\Phi_1\left(\frac{1}{2}, 1, \frac{5}{2}, \frac{z^2}{2}, 1-\lambda\right) \\
&= \exp\left(\frac{z^2}{2}\right)\lambda^{-1}\Phi_1\left(2, 1, \frac{5}{2}, -\frac{z^2}{2}, \frac{\lambda-1}{\lambda}\right) \\
&= \exp\left(\frac{z^2}{2}\right)\lambda^{-1}\exp\left(-\frac{z^2}{2}\right)\sum_{n=0}^{\infty}\frac{(2)_n\,(1)_n}{\left(\frac{5}{2}\right)_n}\frac{\left(\frac{\lambda-1}{\lambda}\right)^n}{n!}\;{}_1F_1\left(\frac{1}{2}, \frac{5}{2}+n, \frac{z^2}{2}\right) \\
&= \sum_{n=0}^{\infty}\frac{(2)_n\,(1)_n}{\left(\frac{5}{2}\right)_n}\frac{\frac{(\lambda-1)^n}{\lambda^{n+1}}}{n!}\frac{\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{5}{2}+n\right)}\exp\left(\frac{z^2}{2}\right)\cdot\left(\frac{z^2}{2}\right)^{-2-n}\left\{1+O(z^{-2})\right\}
\end{aligned}$$

where $(a)_n$ is rising factorial. As $|z| \to \infty$, the numerator in (B.6) converges to:

$$2z\exp\left(\frac{z^2}{2}\right)\cdot\left(\frac{z^2}{2}\right)^{-2}\cdot\left[\frac{3}{4\sqrt{2\pi}}\int_0^1\sum_{n=0}^{\infty}\frac{(\lambda-1)^n}{\Gamma\left(\frac{5}{2}+n\right)}\lambda^{-\frac{1}{2}}\mathrm{e}^{-\frac{\lambda}{2}}d\lambda + \frac{4}{3}\int_1^{\infty}\lambda^{-\frac{3}{2}}\mathrm{e}^{-\frac{\lambda}{2}}d\lambda\right] \tag{B.7}$$

It is easy to show that in (B.7), the second integral is finite. For the first integral, according to the monotone convergence theorem, we can exchange limit and integral, and

$$\sum_{n=0}^{\infty}\int_0^1\frac{(\lambda-1)^n}{\Gamma\left(\frac{5}{2}+n\right)}\lambda^{-\frac{1}{2}}\mathrm{e}^{-\frac{\lambda}{2}}d\lambda = \sum_{n=0}^{\infty}\frac{\Gamma(1+n)\,\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{5}{2}+n\right)\Gamma\left(\frac{3}{2}+n\right)}{}_1F_1\left(\frac{1}{2}, \frac{3}{2}+n, -\frac{1}{2}\right) \tag{B.8}$$

By expanding

$${}_1F_1\left(\frac{1}{2}, \frac{3}{2}+n, -\frac{1}{2}\right) = \sum_{k=0}^{\infty}\frac{\left(\frac{1}{2}\right)_k}{\left(\frac{3}{2}+n\right)_k}\frac{\left(-\frac{1}{2}\right)^k}{k!},$$

we find ${}_1F_1\left(\frac{1}{2}, \frac{3}{2}+n, -\frac{1}{2}\right)$ decreases with n. Therefore, (B.8) converges. The numerator in (B.6) can be simplified as

$$2z\exp\left(\frac{z^2}{2}\right)\cdot\left(\frac{z^2}{2}\right)^{-2}\cdot C_1,$$

where C_1 is a constant. Similarly, the denominator in (B.6) can be simplified as

$$3 \exp\left(\frac{z^2}{2}\right) \cdot \left(\frac{z^2}{2}\right)^{-1} \cdot C_2,$$

where C_2 is also a constant. Therefore, $\frac{\mathrm{d}}{\mathrm{d}z} \log m(z) \to 0$ as $|z| \to 0$. This means (3.18) holds under $\tilde{\pi}(\beta)$.

Next we show that this results still holds after introducing a component of point mass at zero, i.e. $\pi(\beta) = (1-\rho)\ \delta_0(\beta) + \rho\ \tilde{\pi}(\beta)$ for any $0 < \rho \leqslant 1$. Since

$$\mathbb{E}(\beta \mid z) = P(\beta = 0 \mid z) \cdot 0 + P(\beta \neq 0 \mid z)\mathbb{E}(\beta \mid z, \beta \neq 0)$$

where $\mathbb{E}(\beta \mid z, \beta \neq 0)$ is the posterior mean of β under prior density $\tilde{\pi}(\beta)$, to prove (3.18) it is sufficient to show

$$\lim_{|z| \to \infty} P(\beta = 0 \mid z) = 0 \tag{B.9}$$

According to the Bayes rule,

$$\begin{aligned}
&P(\beta = 0 \mid z) \\
&= \frac{P(z \mid \beta = 0)P(\beta = 0)}{P(z \mid \beta = 0)P(\beta = 0) + P(z \mid \beta \neq 0)P(\beta \neq 0)} \\
&= \frac{N(z; 0, 1)\ (1-\rho)}{\mathrm{N}(z; 0, 1)\ (1-\rho) + \rho \int_0^\infty \int_0^\infty N(z; 0, 1 + \eta^2/\lambda)\mathrm{C}^+(\eta; 0, 1)\mathrm{G}(\lambda; \frac{1}{2}, \frac{1}{2}) d\eta d\lambda} \\
&= \frac{1}{1 + \frac{\rho}{1-\rho} \int_0^\infty \int_0^\infty \frac{\mathrm{N}(z; 0, 1+\eta^2/\lambda)}{\mathrm{N}(z; 0, 1)} \mathrm{C}^+(\eta; 0, 1)\mathrm{G}(\lambda; \frac{1}{2}, \frac{1}{2}) d\eta d\lambda} \\
&= \left\{ 1 + \frac{\rho}{1-\rho} \int_0^\infty \int_0^\infty \frac{1}{\sqrt{1 + \frac{\eta^2}{\lambda}}} \exp\left[\frac{z^2 \eta^2}{2(\eta^2 + \lambda)}\right] \mathrm{C}^+(\eta; 0, 1)\mathrm{G}\left(\lambda; \frac{1}{2}, \frac{1}{2}\right) d\eta d\lambda \right\}^{-1}
\end{aligned}$$

In the above integral, the integrand is positive and increases with $|z|$, so we can interchange integral and limit. Apparently, this integral reaches infinity in the limit $|z| \to \infty$. Therefore, (B.9) holds.

Thus we have proved that the prior influence $\mathbb{E}(\beta \mid z) - z$ vanishes as $|z|$ goes to infinity. Notice that prior influence is 0 when $z = 0$. According to continuity, we can conclude that prior influence is bounded. □

B.2 Posterior Sampling Steps

Similar as in Section B.1, we reparametrize the regression problem by transforming parameters $(\beta_j, \omega_j, \eta_j)$ to $(\beta_j, \lambda_j, \eta_j)$, see (B.1) - (B.4). Suppose in the current iteration of MCMC, m_k is the number of pairs (η_j, λ_j) in group k, i.e. $m_k = \sum_{j=1}^{p} \mathbf{1}(c_j = k)$ for $k = 1, 2, \ldots$ According to the slice sampling idea, we introduce an latent variable u_j for each $j = 1, \ldots, p$; and to eliminate confusion, in this section we denote α as the precision parameter of DP, instead of m which we use previously. thus the class indicator c_j can only take value from a finite set. We update all the model parameters according to the following scheme. For each iteration:

1. Update v_k, $k = 1, 2, \ldots$

$$v_k \sim \text{Beta}(1 + m_k, \alpha + p - \sum_{l=1}^{k} m_l)$$

2. Update w_k, $k = 1, 2, \ldots$

$$w_k = v_k \prod_{l<k} (1 - v_l)$$

3. Update u_j, $j = 1, \ldots, p$

$$u_j \sim \text{Unif}(0, w_{c_j})$$

4. Update η_k^*, $k = 1, 2, \ldots$

   Let $\eta_k^* = \xi_k^* \psi_k^*$, where $\xi_k^* = \mathbf{1}(\eta_k^* \neq 0)$. When $\xi_k^* = 1$, $\psi_k^* = \eta_k^*$; when $\xi_k^* = 0$, the value of ψ_k^* does not affect η_k^*. After this decomposition, a priori,

$$\xi_k^* \sim \text{Bernoulli}(\rho)$$
$$\psi_k^* \sim \text{Cauchy}^+\left(0, \frac{1}{\sqrt{\phi}}\right)$$

We use the Metropolis-Hastings algorithm to update ξ_k^* according to

$$p(\xi_k^* \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\xi}_{(-k)}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi) \propto p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi) \cdot p(\xi_k^*)$$

where the $p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi)$ is the likelihood after marginalized out β_0 and $\boldsymbol{\beta}$ as shown in (2.1).

To update ψ_k^*, we use the adaptive Metropolis algorithm on $\log(\psi_k^*)$ according to

$$p(\psi_k^* \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}_{(-k)}^*, \boldsymbol{\lambda}^*, \phi) \propto p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi) \cdot p(\psi_k^*)$$

5. Update c_j: for $j = 1, \ldots, p$

$$c_j = k \quad \text{with probability } \mathbf{1}(w_k > u_j) p(\mathbf{Y} \mid c_j = k, \mathbf{c}_{(-j)}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi),$$

for $k = 1, \ldots, k^*$, where $k^* = \arg\min_k \left\{ \sum_{l=1}^{k} p_l > 1 - \min_{1 \leqslant j \leqslant p} (u_j) \right\}$.

6. Update λ_k^*, $k = 1, 2, \ldots$

Apply the adaptive Metropolis algorithm on $\log(\lambda_k^*)$ according to

$$p(\lambda_k^* \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}_{(-k)}^*, \phi) \propto p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi) \cdot p(\lambda_k^*)$$

7. Simple random swap: if both sets $A = \{i : \eta_i \neq 0\}$ and $B = \{j : \eta_j = 0\}$ are nonempty, randomly draw index $i \in A$ with equivalent weights, and draw $j \in B$ with weight

$$\frac{|\text{Corr}(\mathbf{X}_i, \mathbf{X}_j)|}{\sum_{j' \in B} |\text{Corr}(\mathbf{X}_i, \mathbf{X}_{j'})|}$$

With probability p_{swap}, use the Metropolis-Hasting algorithm to propose to swap (c_i, η_i, λ_i) with (c_j, η_j, λ_j).

8. Update ρ: the Gibbs sampler

$$\rho \sim \text{Beta}\left(a_\rho + \sum_{k=1}^{k*} \delta(\eta_k^* \neq 0), b_\rho + \sum_{k=1}^{k*} \delta(\eta_k^* = 0)\right)$$

9. Update α: the Gibbs sampler by introducing an auxiliary variable x:

$$\alpha \mid x, d \sim \pi_x \mathrm{G}(a_\alpha + d, b_\alpha - \log x) + (1 - \pi_x)\mathrm{G}(a_\alpha + d - 1, b_\alpha - \log x)$$
$$x \mid \alpha \sim \text{Beta}(\alpha + 1, n)$$

where d is the number of non-empty classes: $d = \sum_{k=1}^{k*} \delta(m_k \geqslant 1)$, and $\pi_x = \frac{a_\alpha + d - 1}{p(b_\alpha - \log x)}$.

10. Update ϕ: the adaptive Metropolis algorithm on $\log(\phi)$ according to

$$p(\phi \mid \mathbf{Y}, \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*) \propto p(\mathbf{Y} \mid \mathbf{c}, \boldsymbol{\xi}^*, \boldsymbol{\psi}^*, \boldsymbol{\lambda}^*, \phi) \cdot p(\phi)$$

Bibliography

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

Barbieri, M. and Berger, J. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897.

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.

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhya: The Indian Journal of Statistics. Series A (1961-2002)*, 60, 307–321.

Bernardo, J. M. and Smith, A. F. (2000), *Bayesian Theory*, Wiley Bayesian theory. Vol. 405. Wiley, 2009.

Blackwell, D. and MacQueen, J. (1973), "Ferguson Distributions Via Polya Urn Schemes," *The Annals of Statistics*, 1, 353–355.

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

Carvalho, C., Polson, N., and Scott, J. (2009), "Handling Sparsity via the Horseshoe," *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics.*

Carvalho, C., Polson, N., and Scott, J. (2010), "The horshshoe estimator for sparse signals," *Biometrika.*

Celeux, G., Anbari, M. E., Marin, J., 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., Huang, L., Ibrahim, J. G., and Kim, S. (2008), "Bayesian Variable Selection and Computation for Generalized Linear Models with Conjugate Priors," *Bayesian Analysis*, 3, 585–614.

Clyde, M. and Parmigiani, G. (1998), "Protein Construct Storage: Bayesian Variable Selection and Prediction with Mixtures," *Journal of Biopharmaceutical Statistics*, 8, 431–443.

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

Clyde, M. A. and George, E. I. (2000), "Flexible Expirical Bayes Estimation for Wavelets," *J. R. Statist. Soc. B*, 62, 681–698.

Clyde, M. A. and George, E. I. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94.

Dawid, A. (1973), "Posterior Expectations for Large Observations," *Biometrika*, 60, 664–666.

Diaconis, P. and Ylvisaker, D. (1985), "Quantifying Prior Opinion," *Bayesian Statistics*, 2, 133–156.

Escobar, M. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

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

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1, 515–533.

George, E. I. (1986), "Minimax Multiple Shrinkage Estimation," *The Annals of Statistics*, 14, 188–205.

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

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

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

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

Griffin, J. and Brown, P. (2005), “Alternative prior distributions for variable selection with very many more variables than observations,” *University of Kent Technical Report*.

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.

Haario, H., Saksman, E., and Tamminen, J. (2001), “An Adaptive Metropolis Algorithm,” *Bernoulli*, 7, 223–242.

Hans, C. (2009), “Bayesian lasso regression,” *Biometrika*, 96, 835–845.

Hansen, M. and Yu, B. (2003), “Minimum Description Length Model Selection Criteria for Generalized Linear Models,” *Lecture Notes-Monograph Series*, pp. 145–163.

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.

Hoerl, A. and Kennard, R. (1970), “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” *Technometrics*, 12, 55–67.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999), “Bayesian Model Averaging: A Tutorial,” *Statistical Science*, 14, 382–417.

Ishwaran, H. and James, L. (2001), “Gibbs Sampling Methods for Stick-Breaking Priors,” *Journal of the American Statistical Association*, 96, 161–173.

Ishwaran, H. and Rao, J. S. (2005), “Spike and slab variable selection: frequentist and bayesian strategies,” *The Annals of Statistics*, 33, 730–773.

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

Johnstone, I. and Silverman, B. (2004), “Needles and straw in haystracks: empirical Bayes estimates of possibly sparse sequences,” *The Annals of Statistics*, 32, 1594–1649.

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.

Kass, R. E., Tierney, L., and Kadane, J. (1990), “The Validity of Posterior Expansions Based on Laplace’s Method,” *Essays in Honor of George Bernard, eds. S. Geisser, J.S. Hodges, S.J. Press, and A. Zellner, Amsterdam: North Holland*, pp. 473–488.

Ley, E. and Steel, M. F. (2012), “Mixtures of g-priors for Bayesian Model Averaging with Economic Applications,” *Journal of Econometrics*, 171, 251–26.

Li, Q. and Lin, N. (2010), “The Bayesian Elastic Net,” *Bayesian Analysis*, 5, 151–170.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (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.

MacLehose, R. and Dunson, D. (2010), “Bayesian semi-parametric multiple shrinkage,” *Biometrics*, 66, 455–462.

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*, Capman and Hall.

Mitchell, T. and Beauchamp, J. (1988), “Bayesian Variable Selection in Linear Regression,” *Journal of the American Statistical Association*, 83, 1023–1032.

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

Papaspiliopoulos, O. (2008), “A Note on Posterior Sampling from Dirichlet Mixture Models,” *Working Paper.*

Papaspiliopoulos, O. and Roberts, G. (2008), “Retrospective Markov Chain Monte Carlo Methods for Dirichlet Process Hierarchical Model,” *Biometrika*, 95, 169–186.

Park, T. and Casella, G. (2008), “The Bayesian Lasso,” *Journal of the American Statistical Association*, 103, 681–686.

Pericchi, L. and Sanso, B. (1995), “A Note on Bounded Influence in Bayesian Analysis,” *Biometrika*, 82, 223–225.

Pericchi, L. and Smith, A. (1992), “Exact and Approximate Posterior Moments for a Normal Location Parameter,” *J. R. Statist. Soc. B*, 54, 793–804.

Polson, N. and Scott, J. (2012), “On the Half-Cauchy Prior for a Global Scale Parameter,” *Bayesian Analysis*, 7, 1–16.

Scott, J. and Berger, J. (2006), “An Exploration of Aspects of Bayesian Multiple Testing,” *Journal of Staticial Planning and Inference*, 136, 2144–2162.

Scott, J. and Berger, J. (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 Powe 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.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *J. R. Statist. Soc. B*, 58, 267–288.

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

Tipping, M. (2001), "Sparse Bayesian Learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, 1, 211–244.

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

Walker, S. (2007), "Sampling the Dirichlet Mixture Model with Slices," *Communications in Statistics - Simulation and Computation*, 36, 45–54.

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

West, M. (1987), "On Scale Mixtures of Normal Distributions," *Biometrika*, 74, 646–648.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *J. R. Statist. Soc. B*, 68, 49–67.

Zellner, A. (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*, pp. 233–243, North-Holland/Elsevier.

Zellner, A. and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Helf in Valencia (Spain)*, pp. 585–603, Valencia, Spain, University of Valencia Press.

Zou, H. and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *J. R. Statist. Soc. B*, 67, 301–320.

Biography

Yingbo Li was born and grew up in Shanghai, China. She studied in the School of Mathematical Sciences in Peking University, and received a Bachelor of Science degree in Statistics in 2008. After that, she continued her study in the Department of Statistical Science in Duke University as a graduate student. She earned a Master of Science degree in 2011 and is expected to receive her Ph.D. degree in Statistics in 2013.