Paradoxes and Priors in Bayesian Regression

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy in the Graduate School of The Ohio State
University

By


Graduate Program in Statistics

The Ohio State University

2014

Dissertation Committee:

Dr. Christopher M. Hans, Advisor
Dr. Steven N. MacEachern, Co-advisor
Dr. Mario Peruggia
© Copyright by
Agniva Som
2014
Abstract

The linear model has been by far the most popular and most attractive choice of a statistical model over the past century, ubiquitous in both frequentist and Bayesian literature. The basic model has been gradually improved over the years to deal with stronger features in the data like multicollinearity, non-linear or functional data patterns, violation of underlying model assumptions etc. One valuable direction pursued in the enrichment of the linear model is the use of Bayesian methods, which blend information from the data likelihood and suitable prior distributions placed on the unknown model parameters to carry out inference.

This dissertation studies the modeling implications of many common prior distributions in linear regression, including the popular $g$ prior and its recent ameliorations. Formalization of desirable characteristics for model comparison and parameter estimation has led to the growth of appropriate mixtures of $g$ priors that conform to the seven standard model selection criteria laid out by Bayarri et al. (2012). The existence of some of these properties (or lack thereof) is demonstrated by examining the behavior of the prior under suitable limits on the likelihood or on the prior itself.

The first part of the dissertation introduces a new form of an asymptotic limit, the conditional information asymptotic, driven by a situation arising in many practical problems when one or more groups of regression coefficients are much larger than the rest. Under this asymptotic, many prominent “$g$-type” priors are shown to suffer from two new unsatisfactory behaviors, the Conditional Lindley’s Paradox and Essentially
Least Squares estimation. The cause behind these unwanted behaviors is the existence of a single, common mixing parameter in these priors that induces mono-shrinkage. The novel block $g$ priors are proposed as a collection of independent $g$ priors on distinct groups of predictor variables and improved further through mixing distributions on the multiple scale parameters. The block hyper-$g$ and block hyper-$g/n$ priors are shown to overcome the deficiencies of mono-shrinkage, and simultaneously display promising performance on other important prior selection criteria. The second part of the dissertation proposes a variation of the basic block $g$ prior, defined through a reparameterized design, which has added computational benefits and also preserves the desirable properties of the original formulation.

While construction of prior distributions for linear models usually focuses on the regression parameters themselves, it is often the case that functions of the parameters carry more meaning to a researcher than do the individual regression coefficients. If prior probability is not apportioned to the parameter space in a sensible manner, the implied priors on these summaries may clash greatly with reasonable prior knowledge. The third part of the dissertation examines the modeling implications of many traditional priors on an important model summary, the population correlation coefficient that measures the strength of a linear regression. After detailing deficiencies of standard priors, a new, science driven prior is introduced that directly models the strength of the linear regression. The resulting prior on the regression coefficients belongs to the class of normal scale mixture distributions in particular situations. Utilizing a fixed-dimensional reparameterization of the model, an efficient MCMC strategy that scales well with model size and requires little storage space is developed for posterior inference.
Dedicated to my father, Mr. Arup Kumar Som, my late mother, Mrs. Sanchaita Som and my grandmother, Mrs. Lila Som.
I would like to express my gratitude to everyone who has made it possible for me to arrive at this momentous occasion in my life. First and foremost, I wish to thank my advisors Dr. Chris Hans and Dr. Steve MacEachern for their endless support and amazing mentorship during my four years of doctoral studies. I am greatly indebted to them for their unwavering patience in me, which helped me immensely to overcome my limitations whenever I faltered. I consider myself to be very fortunate that I got the opportunity to learn about the discipline from such brilliant minds and to enhance my knowledge and understanding of statistics under their watchful guidance. My advisors are greatly responsible for instilling in me the tenacity and passion toward research work that is needed to be successful at this high level. Fond memories of our meetings, including the serious discussions that frequently opened my mind to new perspectives on approaching unresolved questions as well as the light hearted chats on non-academic topics, will remain with me forever.

I am grateful to Dr. Mario Peruggia for serving on my doctoral and candidacy exam committees and and to Dr. Xinyi Xu for serving on my candidacy exam committee. Their comments and suggestions helped to broaden the scope of my dissertation research. A word of thanks should go to Dr. Radu Herbei as well for introducing me to new areas of statistics in my first year of graduate school and for helping me
to develop a fascination for research in this field. I am thankful to NSF for partially supporting my research under grant numbers DMS-1007682, DMS-1209194 and DMS-1310294.

My family has always been a great source of love and support and without their immense influence this would never have been possible. Maa, Baba, Dida and Dada have always been there for me whenever I needed them, especially during the tough times that I experienced in my first year living so far away from home for the first time in my life. It was my parents’ dream that I get a doctorate degree, but now that I am so close to receiving one, I feel heartbroken that my mother is no longer with us to enjoy this moment with the rest of the family. I dedicate this dissertation to my dad, who has been a constant source of my inspiration, and my mother and my grandmother, both of whom sacrificed so much for me to achieve this cherished dream.

I would also like to thank all my wonderful friends who made life in graduate school and in Columbus enjoyable and worth cherishing all my life. There is a long list of friends without whom my life in the last four years would not have been as much fun. Thank you Abrez, Aritra, Arjun, Arka, Casey, Dani, Grant, Indrajit, Jingjing, John, Matt, Parichoy, Prasenjit, Puja, Steve and the many others who I did not mention here (but have been equally important to me in recent years), for helping me get through the thick and thin of life as a graduate student. And most importantly, thank you Sarah, the most special person in my life, for our wonderful times together. I am highly appreciative of Sarah’s support, who stood by me throughout, in happiness and in sorrow, and put up with my craziness the past few years.
Vita

August 4, 1987 .......................... Born - Kolkata, India.

2005-2008 .............................. B. Stat. (Hons.), Indian Statistical Institute, Kolkata, India.

2008-2010 .............................. M. Stat., Indian Statistical Institute, Kolkata, India.

2010-2014 .............................. Graduate Teaching Associate,
                               The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Statistics.
# Table of Contents

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiv</td>
</tr>
</tbody>
</table>

## Chapters

<table>
<thead>
<tr>
<th>1. Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Prior Specification in the Linear Model</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Problems with Mono-Shrinkage</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Implications of the Prior on Other Regression Summaries</td>
<td>9</td>
</tr>
<tr>
<td>2. Review of Popular Priors in Bayesian Regression Problems</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Variable Selection Priors</td>
<td>13</td>
</tr>
<tr>
<td>2.1.1 “Spike and Slab”-type Priors</td>
<td>14</td>
</tr>
<tr>
<td>2.1.2 Stochastic Search Algorithms</td>
<td>16</td>
</tr>
<tr>
<td>2.1.3 Priors on the Model Space</td>
<td>17</td>
</tr>
<tr>
<td>2.2 Shrinkage Priors</td>
<td>19</td>
</tr>
<tr>
<td>2.2.1 Scale Mixture Representation</td>
<td>19</td>
</tr>
<tr>
<td>2.3 Classification Based on Prior Knowledge</td>
<td>24</td>
</tr>
<tr>
<td>2.3.1 Subjective Priors</td>
<td>24</td>
</tr>
<tr>
<td>2.3.2 Objective Priors</td>
<td>25</td>
</tr>
<tr>
<td>2.4 Other Discussions on Bayesian Regression</td>
<td>26</td>
</tr>
</tbody>
</table>
3. Block $g$ Priors .................................................. 27
   3.1 Review of $g$ Priors ....................................... 27
      3.1.1 Zellner’s $g$ Prior .................................. 28
      3.1.2 Mixtures of $g$ Priors .............................. 29
   3.2 Asymptotic Evaluations: Paradoxes, Old and New ... 31
      3.2.1 A Conditional Information Asymptotic .......... 33
      3.2.2 A Conditional Lindley’s Paradox .............. 34
      3.2.3 The CLP in Other Mixtures of $g$ Priors ... 36
      3.2.4 Avoiding ELS and CLP Behaviors ............ 39
   3.3 Block $g$ Priors ........................................... 40
      3.3.1 The Block Hyper-$g$ Prior ....................... 42
      3.3.2 Sampling from the Block Hyper-$g$ Prior ...... 42
      3.3.3 Expansion with respect to Least Squares Estimates under the Block $g$ prior ................. 46
   3.4 Asymptotic Evaluations of the Block Hyper-$g$ Prior ... 52
   3.5 Consistency of the Block Hyper-$g$ Prior .......... 57
      3.5.1 Information Consistency .......................... 58
      3.5.2 Conditions and Assumptions ..................... 59
      3.5.3 Model Selection Consistency ..................... 60
      3.5.4 Prediction Consistency ........................... 61
   3.6 The Block Hyper-$g/n$ Prior ........................... 64
      3.6.1 The Conditional Information Asymptotic in the Block Hyper-$g/n$ Prior .................. 65
      3.6.2 Consistency of the Block Hyper-$g/n$ Prior .... 69
   3.7 Application to Real Data ............................... 71
      3.7.1 US Crime Data ...................................... 71
      3.7.2 Ozone Data .......................................... 80
   3.8 Summary .................................................. 87

4. Block $g$ Priors in Non-Orthogonal Designs ............ 89
   4.1 Introduction .............................................. 89
   4.2 Block $g$ Prior on a Transformed Design ............ 90
      4.2.1 Transformation to a Block Orthogonal Design .. 91
      4.2.2 The Block Orthogonalized Block $g$ Prior .... 93
   4.3 Properties of the BOB Hyper-$g$ Prior ............ 95
      4.3.1 Effect of the CLP and ELS ....................... 96
      4.3.2 Significance of the Order of Orthogonalization 105
   4.4 Consistency of the BOB Hyper-$g$ and BOB Hyper-$g/n$ Priors 111
      4.4.1 Information Consistency .......................... 112
### A.5 Proof of Invariance of Block $g$ Priors to Reparameterizations by Blockwise Affine Transformations

196

### A.6 Proof of Lemma 3.4.2

197

### A.7 Proof of Theorem 3.4.1

199

### A.8 Proof of Corollary 3.4.1

201

### A.9 Proof of Corollary 3.4.2

203

### A.10 Proof of Theorem 3.5.1

204

### A.11 Proof of Theorem 3.5.2

206

### A.12 Proof of Lemma 3.5.1

211

### A.13 Proof of Theorem 3.6.3

213

### A.14 Proof of Theorem 3.6.4

214

### A.15 Proof of Theorem 3.6.5

217

### A.16 Proof of Theorem 3.5.2 when $2 < a \leq 3$

218

### A.17 Description of the Ozone Data

220

### B. Appendix for Chapter 4

222

#### B.1 Proof of Lemma 4.3.2

222

#### B.2 Proof of Lemma 4.3.3

224

#### B.3 Proof of Proposition B.2.1

227

#### B.4 Proof of Theorem 4.3.6

229

#### B.5 Proof of Lemma 4.5.3

232

#### B.6 Proof of Lemma 4.5.4

233

#### B.7 Description of the Boston Housing Data

235

### C. Appendix for Chapter 5

236

#### C.1 Derivation of $\pi(\beta \mid \sigma^2$) in (5.4)

236

#### C.2 Derivation of the Scale Mixture Representation of $\beta$

239

#### C.3 Log-concavity of the Full Conditional of $1/\sigma^2$

240

#### C.4 MH Step for the joint update of $(M, \phi_{-1})$

241

### Bibliography

243
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Table showing entropy values for the model posterior under different priors. Hg and Hg/n stand for hyper-g and hyper-g/n respectively, BHg/n denotes the block hyper-g/n prior.</td>
</tr>
<tr>
<td>3.2</td>
<td>Table showing (A) entropy values for posterior distribution of model size and (B) average entropy values for posterior distributions of marginal inclusion under different priors. Hg, Hg/n and BHg/n stand for hyper-g, hyper-g/n and block hyper-g/n respectively.</td>
</tr>
<tr>
<td>3.3</td>
<td>Table showing marginal inclusion probabilities of variables under different posteriors. Predictors marked with asterisks belong to the corresponding median probability model.</td>
</tr>
<tr>
<td>3.4</td>
<td>Table displaying the cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{2}$.</td>
</tr>
<tr>
<td>3.5</td>
<td>Table displaying the cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{3}$.</td>
</tr>
<tr>
<td>3.6</td>
<td>Table from Yu et al. (2011) displaying the cross-validated predictive SSE under different methods.</td>
</tr>
<tr>
<td>3.7</td>
<td>Table displaying the cross-validated predictive SSE (under BMA) for different procedures using the entire training data.</td>
</tr>
<tr>
<td>3.8</td>
<td>Table displaying the cross-validated predictive SSE (under BMA) for different procedures after discarding the influential observation from the training data.</td>
</tr>
<tr>
<td>3.9</td>
<td>Marginal inclusion probabilities under the block hyper-g/n (Form 1) and the hyper-g posteriors.</td>
</tr>
</tbody>
</table>
4.1 Table showing cross-validated predictive SSE under different methods. The SSE values for the Bayesian models are obtained by model averaging. All five forms of the BOPB hyper-g/n prior have the hyperparameter value of $a = 3$.  

4.2 Table showing cross-validated predictive SSE (under BMA) for the hyper-g, hyper-g/n and BOPB hyper-g/n priors. The model averaging is done over all possible permutations and combinations of the set of design blocks appearing in the specific prior form.  

5.1 Prior mean and standard deviation of $\rho^2$ as a function of model size when $\beta \sim N(0, \sigma_{\epsilon}^2 I_p)$.  

5.2 Prior mean and standard deviation of $\rho^2$ as a function of model size when (A) $\beta \sim N(0, I_p)$ and $\sigma_{\epsilon}^2 \sim IG(3, \frac{1}{2})$, and (B) $\beta \sim N(0, I_p)$ and $\sigma_{\epsilon}^2 \sim \text{Gamma}(1, 1)$.  

5.3 Prior mean and standard deviation of $\rho^2$ as a function of model size when $n = 20$ and (A) $\beta \sim N(0, \sigma_{\epsilon}^2 (X^T X)^{-1})$, and (B) $\beta \sim N(0, p^2 \sigma_{\epsilon}^2 (X^T X)^{-1})$.  

5.4 Cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{2}$.  

5.5 Cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{3}$.  

5.6 Cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{4}$.
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The implied prior on $\rho^2$ as a function of model size $p$.</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines and expansion of $\hat{\beta}_2$ occurs in areas shaded with horizontal lines.</td>
<td>48</td>
</tr>
<tr>
<td>3.2</td>
<td>Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines and expansion of $\hat{\beta}_2$ occurs in areas shaded with horizontal lines.</td>
<td>49</td>
</tr>
<tr>
<td>3.3</td>
<td>Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines and expansion of $\hat{\beta}_2$ occurs in areas shaded with horizontal lines.</td>
<td>50</td>
</tr>
<tr>
<td>3.4</td>
<td>Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines. $\hat{\beta}_2$ does not expand since the region of expansion is negligible.</td>
<td>51</td>
</tr>
<tr>
<td>3.5</td>
<td>Comparison of $\log(P(M_\gamma \mid y))$ under the two priors. The figures on the upper panel correspond to prior probability of inclusion of a predictor equal to $\frac{1}{2}$ and the ones on the lower panel correspond to prior probability of inclusion equal to $\frac{1}{3}$. The red line denotes the $y = x$ line.</td>
<td>73</td>
</tr>
<tr>
<td>3.6</td>
<td>Distance $d = 1 - \frac{|\hat{\beta} - \hat{\beta}<em>{LS}|}{|\hat{\beta}</em>{LS}|}$ plotted against log posterior probability. Prior inclusion probability $w = \frac{1}{2}$ in the upper panel and $w = \frac{1}{3}$ in the lower panel.</td>
<td>76</td>
</tr>
<tr>
<td>3.7</td>
<td>Comparison of $\log(P(M_\gamma \mid y))$ under the hyper-$g$ and block hyper-$g/n$ (Form 1) priors. The red line denotes the $y = x$ line.</td>
<td>87</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.1</td>
<td>Kernel density estimates of the posteriors of $\delta, \sigma^2$ and $\rho^2$ from the full model</td>
<td>174</td>
</tr>
<tr>
<td>5.2</td>
<td>Kernel density estimates of the posterior of $\rho^2$ from the full model when the prior mean is fixed</td>
<td>176</td>
</tr>
<tr>
<td>5.3</td>
<td>Kernel density estimates of the posterior of $\rho^2$ from the full model when $a + b$ is fixed</td>
<td>177</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

The linear model has been a mainstay of statistical practice for over 200 years, dating to Gauss and Legendre’s work on least squares estimation and Laplace’s work on least absolute deviation regression. Early use of the model was propelled by its clear focus on the relationship between a set of covariates and the mean or median of the response variable which lends a natural interpretation to the regression coefficients and connects the fitted model with prediction of future responses. The model’s scope can be extended considerably through creation of new covariates, commonly in the form of a polynomial basis expansion of an original set of covariates, allowing one to approximate a wide variety of mean and median functions and leading to the linear model’s prominence in many areas of science. The computational convenience and analytical tractability of least squares estimation gave the edge to the mean-based formulation of the linear model.

In today’s world, the relevance of the linear model remains undiminished, and much of the practice of statistics involves developing, fitting, choosing among and averaging forecasts from linear models. Modern themes include automated screening of enormous numbers of models and adjustment to least squares through penalization of the likelihood and use of Bayesian methods. The latter are the focus of this dissertation.
Bayesian methods have made substantial contributions to linear model. The contributions of Bayesian methods include the coherence of the analysis which follows from the internal consistency of Bayesian methods, the seamless integration of model fitting and inference which follows from a decision theoretic specification of the problem, the ability to combine pre-experimental and post-experimental information which follows from Bayes theorem, and the ability to account for both model uncertainty and parameter uncertainty when making predictions which also follows from Bayes theorem.

In contrast to classical work on the linear model, a Bayesian analysis requires specification of the prior distribution. The prior distribution may be subjectively specified, as is commonly done when one has a specific model in mind, or a rule may be used to specify it, as is often the case when one wishes to screen hundreds, thousands or millions of models. The analysis then flows from the combination of prior distribution and likelihood into the posterior distribution, and from the posterior distribution into inference. The prior distribution is thus a key determinant of the success or failure of the analysis, and an essential part of the Bayesian analysis is an understanding of the consequences of selecting a prior distribution of one form or another.

The first focus of this dissertation is the investigation of properties of certain types of prior distributions for the linear model. These prior distributions include commonly used priors such as the $g$ prior and a variety of mixtures of $g$ priors. This investigation makes use of novel limiting arguments which lead to the description of the Essentially Least Squares phenomenon and uncover a new paradox which we call the Conditional Lindley’s Paradox. Current prior distributions perform poorly under
the phenomenon and paradox. This leads to the development of a new class of prior distributions, the block hyper-$g$ priors, which perform well on traditional criteria such as model selection consistency and predictive consistency, and equally well under our new evaluations.

The second focus of this dissertation is the practical application of the Bayesian linear model. Investigation into use of the model is twofold. First, an implementation is developed for the block hyper-$g$ prior distributions, with attention given to computational and modeling issues. Second, a set of science driven prior distributions are developed. These prior distributions make use of a small amount of subjective input, focusing on the implied distribution of an important summary of an experiment. The science driven prior distributions keep this implied distribution approximately constant across the set of models under consideration, in sharp contrast to currently popular prior distributions. These prior distributions are developed, a computational implementation is devised, and their benefits for data analysis are illustrated.

1.1 Prior Specification in the Linear Model

The basic regression problem can be described as explaining the behavior of the response vector $y = (y_1, y_2, \ldots, y_n)^T$ using a known set of $p$ potential predictor variables $x_1, x_2, \ldots, x_p$. If $X = (x_1, x_2, \ldots, x_p)$ denotes the $n \times p$ design matrix and $\beta$ denotes the vector of regression coefficients, then the linear model can be represented as:

$$y = 1 \alpha + X \beta + \epsilon$$

where $1$ is a $n \times 1$ vector of 1’s and $\alpha$ is the intercept. The vector of errors $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ is generally assumed to be distributed as a Gaussian random vector with
independent and identically distributed components, \( \epsilon \sim N(0, \sigma^2 I_n) \). The assumption of mean zero errors leads to the expression \( E(y) = \alpha 1 + X \beta \). The method adopted by Bayesians involves placing a prior distribution on the vector of unknown parameters \((\alpha, \beta, \sigma^2)\) and the choice of prior distributions on these parameters can be based on a variety of theoretical, computational and practical considerations.

Choice of the prior distribution is generally considered to be more important for model selection than for estimation. A classic example which advocates the significance of the prior distribution in model selection is the Lindley Paradox (also known as Lindley’s Paradox). It demonstrates how a diffuse prior distribution can sometimes lead to a major conflict between Bayesian and frequentist approaches to inference. Selection of meaningful prior distributions for model specific parameters is crucial as well, since model selection procedures are highly sensitive to prior specification (Berger and Pericchi, 2001). The prior distribution plays a strong role in determining the model-averaged predictions since these involve the relative weighting of multiple models. Predictions from Bayes procedures are generally averaged over a collection of models to account for model uncertainty, and the contribution of each model to the weighted sum is directly influenced by prior choice.

When working with the linear model, one fundamental task is the determination of which covariates are active (i.e., have a non-zero regression coefficient) and which are inactive (i.e., have a zero regression coefficient). In the Bayesian setting, this is accomplished by specifying a mixed distribution on the regression coefficients. For a linear model with a specified set of active variables, a probability distribution is placed directly on the active regression coefficients. Some probability \( 1 - w \in (0, 1) \) is assigned to each regression coefficient being inactive, so that the coefficient is
active with probability $w$ and the size of the non-zero active coefficient is distributed according to some absolutely continuous (w.r.t. Lebesgue measure) density function. When the continuous distribution is uniform, these prior distributions are called the “spike and slab” priors \[ \text{Mitchell and Beauchamp 1988]}. A uniform “slab” might not always be ideal, and often times it is appropriate to impose more flexible (and complex) prior distributions on the non-zero regression coefficients. The foremost choice among non-uniform “slabs” is the normal slab which generates the following prior on a single coefficient $\beta_i$

$$\beta_i \sim (1 - w) \delta_0 + w N(\beta_0, \sigma_{\beta_i}^2)$$

where $\delta_0$ represents a point mass at 0 and $\sigma_{\beta_i}^2$ is a variance parameter.

The choice of $\sigma_{\beta_i}$ also plays a major role in inference and over the years researchers have suggested distinctive forms of the variance parameter. Often times for convenience, the priors on $\beta_i, i = 1, \ldots, p$, are taken to be independent. But researchers have also come up with distributions to model structural dependence among the regression coefficients and a leading example of such a prior is the $g$ prior due to Zellner (Zellner 1986):

$$\beta \mid g, \sigma^2 \sim N(0, g\sigma^2(X^TX)^{-1})$$

The $g$ prior can also be specified as a mixed distribution when there is uncertainty regarding the “correct” model to be used. As in the case of the “spike and slab”-type priors, the $g$ prior may be imposed only on the active coefficients in a particular model and prior specification is then repeated over a big collection (generally all possible combinations) of sets of active variables. Each set of active variables determines a unique model.
An alternative is to design prior distributions without the complexity of determining the sets of active and inactive covariates and the prior weights associated with each combination. This variation of priors has gained importance in statistical literature as the class of shrinkage priors, where a prior distribution is specified focusing on the shrinkage properties of the estimators of regression coefficients. The idea is to impart a continuous prior to each regression coefficient with the prior density having a spike at zero and, preferably, heavy tails. This can lead to substantial computational benefits in large problems, but this approach precludes any predictor from being exactly zero. Unimportant covariates in the model admit near zero coefficient estimates, but those small values can never exactly equal zero.

It is a familiar tactic to use shrinkage priors that can be expressed as scale mixtures of normal distributions (West, 1987); currently there is an enormous literature and a profusion of priors that can be represented as scale mixtures of normals. While most of the normal scale mixture priors assume the regression coefficients to be independent, priors inducing dependence between the coefficients through the empirical design matrix (or some other covariance matrix) are also not uncommon. A prior distribution with such dependence structure is usually expressed as a scale mixture of $g$ priors (Cui and George, 2008; Liang et al., 2008; Maruyama and George, 2011; Bayarri et al., 2012), which is a special case of a scale mixture of normal distributions.

1.2 Problems with Mono-Shrinkage

Section 1.1 outlines the broad classification of popular variations of priors assigned to coefficients in the linear model. The repeated reference to the Gaussian prior and its many variants in the preceding discussion can be attributed to these probably
being the most regularly used prior distributions in the statistics literature. A general class of prior distribution on regression coefficients with widespread use in regression modeling is described as:

\[ \boldsymbol{\beta} \mid c, \sigma^2 \sim N(\mathbf{0}, c\sigma^2 \Sigma) \quad (1.1) \]

for some scale parameter \( c \) and prior covariance matrix \( \Sigma \). Notice that the popular \( g \) prior belongs to this class of priors if we consider \( c = g \) and \( \Sigma = (X^T X)^{-1} \). The scale parameter \( c \) greatly impacts the amount of shrinkage experienced by the regression coefficient estimates. Large values of \( c \) are associated with little or no shrinkage in the estimated coefficients while small values of \( c \) lead to strong shrinkage in the estimates.

In many practical regression problems with a sizable number of predictor variables, it is not unreasonable to expect that only a few of them will really explain the variation in the response while the other variables will not be useful. In such a situation one should expect a sensible Bayesian procedure to generate estimates that get shrunk toward zero for the small or moderate coefficients, and to leave the large coefficients unaltered at the same time. So it is often necessary to accommodate distinctive degrees of shrinkage in separate groups of predictor coefficients. Expecting such an ideal poly-shrinkage behavior is unrealistic when one implements a prior of the form \((1.1)\), since it is only possible to have identical shrinkage across all coefficients (mono-shrinkage) with a common scale \( c \).

In recent literature, priors like \((1.1)\) have been extended further by placing a prior distribution on the unknown parameter \( c \). No fixed choice of \( c \) can work well in all regression problems, and a distribution on \( c \) helps to generate a data adaptive robust solution that usually leads to better estimation and prediction. The prior distribution on \( c \) is suitably chosen to generate a thick-tailed marginal prior on \( \boldsymbol{\beta} \),
since thick tails have been linked to some essential consistency properties in the linear model. However, despite the data adaptiveness with a hierarchical prior, the original complication still persists since the prior causes adaptive shrinkage, but mono-shrinkage nonetheless.

Further pitfalls of mono-shrinkage are evident when one closely examines the properties of many standard priors. The motivation behind the development of most single scale priors mainly stems from model comparison and consistency properties of Bayesian procedures, and are occasionally driven by computational considerations. We demonstrate in Chapter 3 that priors associated with mono-shrinkage display undesirable inferential performance when there is an appreciable difference in coefficient sizes for predictor variables. The structure of shrinkage arising from specific priors of the form (1.1) acts as a catalyst for some paradoxical and perplexing behaviors, described in Chapter 3 as the Conditional Lindley’s Paradox (CLP) and Essentially Least Squares (ELS) estimation. The reason behind these aberrations is that the single scale parameter adapts with the data, and the presence of even a few huge coefficients forces most of the posterior probability of \( c \) toward big values. Dissemination of the probability mass of \( c \) away from small values makes the prior distribution (1.1) on the regression coefficients very diffuse. As a consequence, ELS and CLP behaviors emerge in the priors connected to mono-shrinkage, and deteriorate the quality of inference.

Due to all of these considerations, the need for multiple scale parameters in a prior, capable of inducing poly-shrinkage in the coefficients, becomes indispensable in many practical problems. The block hyper-\( g \) prior, which is an adaptation of the traditional \( g \) prior with a single scale parameter, is introduced in Chapter 3 as a
remedy to the aforementioned problems. The properties of the new multiscale prior are studied in detail and the benefits of poly-shrinkage resulting from the block $g$ priors are highlighted in Chapters 3 and 4.

1.3 Implications of the Prior on Other Regression Summaries

The most commonly used prior distributions for Bayesian regression models, some of which are briefly described in Section 1.1, typically assume that coefficients are \textit{a priori} independent or induce dependence via the empirical design matrix. While these standard priors (and recently refined versions of them) may exhibit desirable behavior with respect to targeted inferential goals, they do not distribute probability throughout the entire parameter space in a way that is consistent with all prior beliefs. For example, most standard priors lead to a peculiar modeling implication on the strength of regression measured by the \textit{population correlation coefficient} $\rho$. A serious aberration arises due to these common priors allocating unreasonably high probability to values of $\rho^2$ near 1 when the number of predictor variables increases.

To illustrate this effect, consider a regression model with a very basic prior distribution on the regression coefficients

$$\beta \sim N(0, g\Sigma)$$

where $g$ is a scale parameter and $\Sigma$ is a covariance matrix. In the simple case when the covariance matrix $\Sigma = I_p$ and $g = 1$, the left panel of Figure 1.1 shows how the prior on $\rho^2$ focuses most of its mass around 1 as $p$ grows. It is implausible to believe that merely incorporating more covariates, irrespective of their significance in the particular model, would explain nearly all of the variation in the response.
One possible way to control this peculiar behavior of the prior distribution on $\rho^2$ is to fix the scale parameter $g$ to be a function of the number of predictors $p$, but even in this case some problems persist as illustrated in the right panel of Figure 1.1. With a scale of $g = \frac{1}{p}$, one is able to force the unrealistic amount of prior mass away from 1, but now the prior concentrates around a neighborhood of $\frac{1}{2}$. In the general case when $g$ is chosen to be an arbitrary decreasing function of $p$, the prior on $\rho^2$ again moves toward a degenerate distribution and concentrates either around one or around some arbitrary value within the unit interval (depending on the particular choice of $g$), as the number of predictors grows.

Figure 1.1: The implied prior on $\rho^2$ as a function of model size $p$.

The correlation coefficient is not the only measure that suffers from strange modeling implications. Oftentimes direct prior specification on regression coefficients
severely impacts inference related to other regression parameters and summaries. Strong knowledge about a function of the model parameters can be appropriately modeled with a meaningful prior distribution on the function itself. As an example, in some designed experiments it would be completely impractical to expect the individual effect sizes as well as the sizes of some specific contrasts (or other linear combinations) to exceed certain threshold values. In situations like these, it makes better sense to elicit carefully designed priors jointly on the contrasts and particular effects, since common variable selection priors may concentrate excessive prior mass outside of the tolerable bounds. It is perfectly logical to impart information indirectly to the regression model in order to be compatible with prior beliefs that are more significant to the researcher.

This motivates the origin of a new class of priors, called science driven priors, to overcome shortcomings of traditional priors and to develop sensible Bayesian models with favorable inferential properties. A science driven prior can be developed based on a variety of model summaries, but our formulation in Chapter 5 is aimed specifically to deal with the undesirable implications on $\rho^2$. Following this new approach, prior information is incorporated into the model by directly eliciting a prior on the strength of linear regression relationship. Chapter 5 presents the success of one version of science driven prior called the $R$-prior, and elaborates on its statistical properties.
Chapter 2: Review of Popular Priors in Bayesian Regression
Problems

There have been two main streams of thought in analyzing the Bayesian estimation and variable selection problem for the linear model. The first stream casts uncertainty about predictor variables being active or not in the traditional framework of variable selection. A regression coefficient is zero if the corresponding predictor is inactive and non-zero if the predictor is active. Hence, the prior specification is carried out over a collection of different models, with each model determined by its set of active predictor variables. Given a set of $p$ explanatory variables, the collection of models arising from including or excluding specific predictors in the model is often taken to be the entire set of $2^p$ possible models. This approach is described in Section 2.1. The second stream of thought retains all candidate variables in the model. Unimportant predictor variables have regression coefficients that are shrunk toward zero. The relative importance of the predictors is driven by the amount of shrinkage, and the focus of research is on developing prior distributions that lead to an appropriate shrinkage pattern. This approach is described in Section 2.2.
2.1 Variable Selection Priors

The development of prior distributions in Bayesian regression problems is often motivated by theoretical and practical considerations related to model selection and estimation. Sometimes the primary goal of research is to discover the model (or collection of models) that best explains the data among the set of available models. When the model space consists of all possible combinations of predictors (any covariate can be in or out of the model), variable selection can be cast as a special case of model selection. Each group of active variables selected from the entire collection of candidate predictors determines a particular model. The set of prior distributions specified (by a rule) on the regression parameters from all such models forms the variable selection prior.

The variable selection prior is suitable to deal with model uncertainty when equipped with a prior on the model space. This generates a posterior distribution on the set of models which helps to quantify the amount of uncertainty associated with each model as well as each individual predictor. The marginal inclusion probability of a predictor variable manifests its importance to the regression model, while a high posterior weight for a model indicates that predictors in the given model are useful collectively. Even while implementing a variable selection prior, it is rare to use a single model like the highest posterior probability model or the median probability model (Barbieri and Berger, 2004) for estimation purposes. No model is believed to be true with certainty and the usual path adopted to account for the unpredictability in model selection is Bayesian model averaging. Estimation and prediction with these priors are carried out by averaging estimates over all possible models, where each estimate is weighted according to the relative importance of the model.
2.1.1 “Spike and Slab”-type Priors

The most basic version of a variable selection prior is the “spike and slab” prior proposed by Mitchell and Beauchamp (1988). In this prior, the inactive covariates are chosen with a prespecified prior probability and their coefficients are set equal to zero in the model. This constitutes the “spike” in the prior while the “slab” part comes from a uniform distribution on the set of active predictor variables, which includes all candidate variables excluding the selected inactive ones. The uniform slab is defined over a very wide interval on the real line, and so there is effectively no shrinkage in the estimates of the active coefficients with this prior.

Over the years there have been many extensions of the basic prior by Mitchell and Beauchamp (1988), including non-uniform “slabs”, continuous “spikes” and empirical Bayes as well as fully Bayesian versions of the model. George and McCulloch (1993) and Ishwaran and Rao (2005) modified the “spike and slab” idea to implement continuous “spikes” in place of the degenerate spike. They proposed an independent two component mixture of normals prior (centered at zero) for the regression coefficients, assigning a very small variance to one of the normal components in order to closely mimic a degenerate distribution at zero. Ishwaran and Rao (2005) improved on the prior by George and McCulloch (1993) with a “rescaled spike and slab” prior and added another level of distributions in the hierarchical prior specification to reduce reliance on prior hyperparameters. There is no clear distinction between active and inactive covariates in these priors since none of the coefficients can be exactly zero, but the normal component in the prior with near-zero variance forces the unimportant predictors to have very small coefficient estimates. Other notable extensions of Mitchell and Beauchamp (1988)’s prior include the work of Foster and George (1994),
Kass and Wasserman (1995) and Fernandez et al. (2001). They implemented the $g$ prior as a particular form of a normal (non-uniform) “slab,” and suggested different choices of the hyperparameter $g$ based on various theoretical motivations.

Some researchers also pursued an empirical Bayes (EB) approach to the linear regression problem. George and Foster (2000) worked with the familiar $g$ prior form and used EB estimates of $g$ to resolve the issue of specifying a predetermined value for the hyperparameter. Johnstone and Silverman (2004) analyzed the asymptotic properties of the familiar “spike and slab” procedure, but replaced the known mixing probability $w$ with a data driven EB choice of $w$.

The fully Bayesian approach to prior specification is popular for its robust behavior, as it enables the prior distribution on regression coefficients to adapt according to the data. In a fully Bayesian formulation, the distributions on all the parameters adjust with the sparsity pattern in the data, which ultimately guide the amount of shrinkage in the coefficient estimates. Liang et al. (2008), Maruyama and George (2011), Bayarri et al. (2012) and Maruyama and Strawderman (2010) resorted to a fully Bayesian analysis of “$g$-type” priors by placing a prior distribution on the hyperparameter $g$. The main benefit of this approach is that it obviates selection of hyperparameters for each individual problem.

Scott and Berger (2010) showed that there is an inherent difference between the empirical Bayes and the fully Bayes procedures and that inference from the two approaches might be quite different, and drastically so in certain situations. Most of these “variable selection” priors were developed subject to the design matrix $X$ having full column rank, but they have been extended to the large $p$, small $n$ setting as well (West (2003) Maruyama and George (2011)).
2.1.2 Stochastic Search Algorithms

The choice of the different “spike” and “slab” components over the years has been guided by both theoretical and computational considerations and it is impossible to unearth one absolute winner that is superior in performance to all others in any given problem. The main drawback for this kind of priors, however, is that the complexity of computation grows exponentially fast with the number of explanatory variables and it becomes impractical to perform a complete analysis involving the entire model space even in moderately large problems. The specification of the two component normal mixture prior was initially intended to simplify calculations and to obtain efficient MCMC routines to explore the posterior. An important consideration in selecting the prior on regression parameters in the earlier days had been to select conjugate priors that lead to simple analytic expressions of the marginal likelihood. The normal–inverse-gamma family (George and McCulloch, 1993, 1997; Raftery et al., 1997) gradually gained importance because the conjugate normal prior on the coefficients along with an inverse-gamma prior on the error variance $\sigma^2$ promoted wonderful computationally tractable algorithms aiding posterior inference.

Stochastic search routines have been devised in recent years as a solution to the computational limitation in large problems, and various algorithms with distinctive attributes have been suggested (George and McCulloch, 1993, 1997; Raftery et al., 1997; Berger and Molina, 2005; Nott and Kohn, 2005; Hans et al., 2007; Scott and Carvalho, 2008; Bottolo and Richardson, 2010; Clyde et al., 2011). Instead of exploring the whole model space, these efficient routines aim to detect models with higher posterior probabilities and provide approximate values for many posterior summaries of interest, which would have otherwise necessitated a full enumeration of the model
space. The Shotgun Stochastic Search (SSS) algorithm developed by Hans et al. (2007) is useful to rapidly discover regions of high posterior probability in the model space by defining appropriate neighborhoods of models which can be explored in parallel. Online estimation of marginal posterior inclusion probabilities drive the model search strategies of Berger and Molina (2005), Scott and Carvalho (2008) and Clyde et al. (2011), with a goal of identifying highly likely models containing the most relevant predictors. The collection of models can then be used to furnish reasonably accurate posterior summaries of the entire model space.

2.1.3 Priors on the Model Space

This approach to Bayesian linear regression entails specification of a prior distribution on the model space since there is uncertainty about which predictors are active. The prior distribution on the model space plays a direct role in determining the posterior weights of individual models and one has to be careful in assigning prior probabilities to the models. When the size of the model space is small, it is possible to assign a subjective probability to each model based on prior beliefs. However, subjective specification of these prior probabilities is unrealistic for even a moderate number of candidate predictors. With $p$ predictors, if all possible combinations of predictors are considered to be within the model space, there are $2^p$ models. For $p = 30$, there are over 1 billion models. For large problems, subjective specification of the prior distribution cannot be done, and so the prior distribution is specified by a rule. The most common rule leads to the Bernoulli variable inclusion prior. Under this prior distribution, a single parameter, $w$, is specified. The predictor variables
are independently declared active (with probability \( w \)) or inactive (with probability \( 1 - w \)).

The value of \( w \) determines whether larger or smaller models receive greater or lesser prior probability. A larger value of \( w \) favors bigger models while a smaller value of \( w \) favors smaller models. The “indifference prior” takes \( w = \frac{1}{2} \) under which all models receive the same prior probability. This prior, however, is strongly informative about the number of variables to be included in the model, and heavily favors half of the total number of predictors to receive the largest probability of inclusion \textit{a priori}. Unreasonable values of the hyperparameter \( w \) can have a negative impact on inference and so it is imperative to choose the Bernoulli variable inclusion prior judiciously.

Use of a fixed choice of \( w \) in the Bernoulli prior, independent of the total number of predictors in the model, is prone to problems. \cite{ScottBerger2010} show that such priors (with fixed \( w \)) do not account for multiplicity correction in model selection. Multiplicity correction is essential in a Bayesian procedure, and more so in high-dimensional problems where the prior can wash out the effect of the likelihood without the necessary multiplicity adjustments. For example, in a regression problem with \( p \) predictors, the full model (with all \( p \) predictors) is \((\frac{w}{1-w})^p\) times more likely in the prior than the null model (with no predictors). When \( p \) is large and \( w = \frac{1}{3} \) (say), \((\frac{w}{1-w})^p = 2^{-p}\) which makes the null model favored over the full model so strongly by the prior that it becomes difficult for the data to override the prior effect. The full model is unlikely to be preferred over the null model in the posterior distribution, even when it is clearly a much better fit.

Placing a prior distribution on \( w \) resolves this issue, as this fully Bayesian method has the desirable property of automatically adjusting for multiplicity. The usual
path adopted in the fully Bayesian approach is to assign a suitable Beta prior to \( w \), where the prior hyperparameters are chosen to match the prior belief regarding the expected number of predictor variables in the data generating model. The EB (Type-II maximum likelihood) estimate of \( w \) is sometimes used as a plug-in value for \( w \) in the analysis. This EB model prior can also successfully handle the multiplicity problem (Scott and Berger, 2010). Both the EB and fully Bayesian priors on the model space display admirable properties because of their ability to adapt with data.

2.2 Shrinkage Priors

The second stream of thought adopted in Bayesian regression problems is to work with the full model and specify a shrinkage prior on the regression coefficients. This is a computationally efficient alternative to the “spike and slab” priors. Most shrinkage priors can be expressed as different scale mixtures of normals. The Bayesian Lasso prior (Park and Casella, 2008; Hans, 2009), the orthant normal prior (Hans, 2011), the generalized double Pareto prior (Armagan et al., 2013), the Horseshoe prior (Carvalho et al., 2010) and the normal-exponential-gamma/normal-gamma prior (Griffin and Brown, 2005, 2010, 2011, 2012) all conform to this well-studied mixture representation. The properties of the mixture determine the shrinkage properties of the prior.

2.2.1 Scale Mixture Representation

Polson and Scott (2012a) represent common shrinkage priors as part of the class of global-local scale mixture of normals:

\[
\beta_i \mid \tau^2, \lambda_i^2 \sim N(0, \tau^2 \lambda_i^2)
\]
\[
\lambda_i^2 \sim \pi(\lambda_i^2) \\
(\tau^2, \sigma^2) \sim \pi(\tau^2, \sigma^2), \ i = 1, 2, \ldots, p
\]

where \(\lambda_i^2\) denotes a local variance component and \(\tau^2\) denotes the global variance component. It has been argued by many researchers (using different techniques to justify their claim) that the prior on \(\tau^2\) should focus a lot of mass around zero while the prior on \(\lambda_i^2\) should have fat tails. The reasoning behind such thought comes from the fact that the global scale \(\tau\) with its substantial mass around zero would help to shrink noise vigorously while the heavy tails of the local scales \(\lambda_i\) would nullify the shrinkage effect on large signals. Marginalizing over the local scale parameter \(\lambda_i\) results in a symmetric distribution for \(\beta_i\) with a mode at zero and both (left and right) tails decaying to zero away from the origin. The mode at zero helps to shrink noise toward zero while presence of sufficient prior mass in the tails is necessary to explain large signals. Tails that are heavier than exponential (e.g. Cauchy tails) are recommended for the marginal distribution of \(\beta_i\), otherwise the prior will always shrink coefficient estimates toward zero, even when they correspond to real signals (Carvalho et al., 2010). The degree of shrinkage in observations that are noise is controlled by the “spike” of the marginal prior density at the origin and a bigger spike is more effective in squelching noise to zero.

Following Polson and Scott (2011), some popular shrinkage priors that can be encompassed within the framework of global-local normal scale mixtures are:

- **Double-exponential**: \(\lambda_i^2\) has an exponential density resulting in the Bayesian lasso prior (Park and Casella, 2008; Hans, 2009). The Bayesian lasso due to
Hans (2009), however, has been characterized by an orthant normal prior (mixture of normal distributions restricted to different orthants of the Euclidean space) rather than by normal scale mixtures.

- **Student-t**: $\lambda_i^2$ has an inverse-gamma density resulting in a Student-t distribution for $\beta_i$ with $\nu$ (say) degrees of freedom. The relevance vector machine (Tipping, 2001) arises from the t distribution prior with the degrees of freedom parameter $\nu \rightarrow 0$.

- **Normal/Jeffreys**: $\lambda_i^2$ has an improper prior (Jeffreys’ prior) given by $\pi(\lambda_i^2) \propto \frac{1}{\lambda_i^2}$ that results in $\pi(\beta_i) \propto \frac{1}{|\beta_i|}$ (Figueiredo 2003; Bae and Mallick 2004).

- **Horseshoe**: $\lambda_i^2$ has an inverted-beta IB($a, b$) distribution with parameters $a = b = \frac{1}{2}$ (Carvalho et al. 2010). The horseshoe prior can be envisioned to be part of the broader class of hypergeometric-beta mixture priors for regression coefficients by choosing arbitrary values of $a$ and $b$.

- **Strawderman-Berger**: This prior is represented as $\beta_i \mid \kappa_i \sim N(0, \kappa_i^{-1} - 1)$ and $\kappa_i \sim \text{Beta}(\frac{1}{2}, 1)$ (Strawderman 1971; Berger 1980).

- **Elastic net**: The mixing parameter is represented in a slightly different form and the mixing density is inverse-gamma restricted to the (0,1) interval resulting in the Bayesian elastic net prior (Li and Lin 2010, Hans 2011). The Bayesian elastic net characterization in Hans (2011) is also based on the orthant normal prior.
• **Normal-gamma (NG)/Normal-exponential-gamma (NEG):** For the NG prior, $\lambda^2_i$ has a gamma distribution while for the NEG prior, $\lambda^2_i$ has an exponential distribution with a further gamma mixing distribution for the exponential rate parameter (Griffin and Brown, 2005, 2010, 2011). Another prior used in sparse regression modeling comes from a linear combination of independent NG priors inducing prior correlation between the regression coefficients and is called the correlated normal-gamma (CNG) prior (Griffin and Brown, 2012).

• **Generalized double Pareto:** This prior also originates from an exponential prior on $\lambda^2_i$, but instead of assigning a gamma prior on the exponential rate parameter as in the NEG case, the generalized double Pareto (Armagan et al., 2013) has the gamma prior on the square root of the rate parameter.

• **Dirichlet-Laplace (DL):** The DL prior (Bhattacharya et al., 2014) is of a slightly more general form than the global-local normal scale mixture representation described earlier. The induced prior on $\beta_i$ is implied by the hierarchical structure:

$$
\beta_i \mid \tau^2, \lambda^2_i, \phi^2_i \sim N(0, \tau^2 \lambda^2_i \phi^2_i)
$$

$$
\lambda^2_i \sim \text{Exp} \left(\frac{1}{2}\right) \quad \text{(with mean 2)}
$$

$$
\phi \sim \text{Dirichlet} \left(a, a, \ldots, a\right).
$$

These prior distributions on the local variance components have been specifically developed to produce heavy tails for $\beta_i$ and a spike around zero. The prior specification in the shrinkage prior setup is usually completed by assigning an improper (Jeffreys') prior to the error variance $\sigma^2$ and in most situations, an inverse-gamma
prior to the global scale parameter $\tau^2$. Gelman (2006) and Polson and Scott (2012b) argue against the routine use of the inverse-gamma prior and instead advise using a half-Cauchy distribution as the default prior for $\tau^2$. They show that posterior inference based on the inverse-gamma prior is highly unstable and sensitive to hyper-parameter choices, and that even “non-informative” or “vague” inverse-gamma prior choices, in truth, impart a lot of information to the prior distribution. The half-Cauchy prior leads to more stable inference without adding to the computational burden as it allows conditionally conjugate MCMC steps.

Other variations of shrinkage priors and connections to nonparametric procedures can be found in Polson and Scott (2011), which includes an elaborate discussion on a more general class of shrinkage priors. A common observation in this regard is the equivalence between the maximum a posteriori (MAP) estimates from these shrinkage priors and frequentist solutions to many penalized regression problems. The LASSO (Tibshirani 1996), bridge regression (Fu 1998) and the elastic net (Zou and Hastie 2005) are classic examples of the penalized regression problem where (penalized) coefficient estimates can be viewed as the mode of the posterior distribution of regression coefficients given the global variance parameter. The global variance $\tau^2$ is closely related to the penalty parameter in a penalized likelihood. MAP coefficient estimates from the conditional posteriors generated by these shrinkage priors lead to sparse solutions when certain required conditions are satisfied by the penalized likelihood function (Fan and Li 2001).
2.3 Classification Based on Prior Knowledge

2.3.1 Subjective Priors

Some researchers also classify prior specification into two broad categories: subjective prior elicitation and objective prior specification. Subjective prior elicitation quantifies expert opinion about the regression problem in the form of a prior distribution on model parameters. The shape of the prior distribution and its hyperparameters are chosen to build on the expert’s knowledge. Garthwaite and Dickey (1988, 1992) and Kadane and Wolfson (1998) stressed the importance of careful prior specification, focusing mainly on the predictive distribution of the response variable. Garthwaite and Dickey (1988, 1992) elicit a prior distribution by assessing quartiles of the response variable under an optimal choice of design points. These “optimal” points are determined by minimizing the interquartile range of the prior predictive distribution subject to some design constraints. Kadane and Wolfson (1998) developed different general and application-specific prior elicitation methods, also by assessing certain quantiles of the response at “optimal” design points, but used a different criterion for defining optimality.

Another example of subjective prior elicitation is that of a suitable prior being imposed on the strength of linear regression relationship, with the distribution chosen according to an expert’s beliefs. Chapter 5 of this dissertation shows how such a prior indirectly induces a distribution with nice properties on the regression coefficients and distributes probabilities consistently in the parameter space in accordance with the expert’s prior knowledge. Bedrick et al. (1996) suggested a new form of an informative prior distribution, called the conditional means prior, in the generalized linear model problem. The conditional means prior is built by assigning a prior distribution to
each of $p$ ($p$ is the number of covariates) chosen points on the regression surface. The subjective priors at these design locations collectively induce a prior on the regression coefficients which has close connections to a data augmentation prior.

### 2.3.2 Objective Priors

Objective prior specification has been much more popular in statistical literature than subjective prior elicitation primarily because of the discomfort many feel in specifying a subjective prior distribution, and also because objective prior distributions are more easily treated from a mathematical perspective. The recent interest in higher dimensional models has, of necessity, tilted prior specification in the direction of rule-based methods.

The prior distributions recommended for use in model selection differ from those used in estimation. The conventional choices of objective priors in estimation problems are typically improper, with the impropriety tied to the dimension of the model. This creates a new obstacle in model selection as the arbitrary multiplicative constants arising from these improper priors lead to indeterminate Bayes factors. This obstacle can be overcome by using an improper prior on parameters common to all of the candidate models, provided the corresponding parameters are orthogonal to the remaining parameters in the full model (Berger et al., 1998). The non-common parameters are given proper prior distributions to avoid the arbitrariness of the Bayes factor.

Bayarri et al. (2012) lay out seven important criteria that any objective prior should satisfy in order to conform to a coherent and rational statistical analysis. With
a view to construct an “uninformative” objective prior so that the likelihood domi-
nates posterior inference, the reference prior (Bernardo, 1979; Berger and Bernardo; 
Berger et al., 2009) was developed from an information theoretic view by maximizing 
an appropriate measure of distance between the prior and the likelihood.

In some other situations, researchers start with a noninformative improper prior 
(generally a reference prior) and then define a new kind of default Bayes factor, ex-
pressed as a combination of two distinct Bayes factors, one calculated with the entire 
sample and the other with a sample of a properly defined “minimal” size. The vari-
ous intrinsic Bayes factors (Berger and Pericchi, 1996b[a] and fractional Bayes factors 
(O’Hagan, 1995) are examples of this approach. Posterior probabilities for all the 
models can be directly calculated via the default Bayes factors and the prior distri-
bution on the model space. Although these default Bayes factors are not generated 
like real Bayes factors with respect to some inherent prior distribution on the re-
gression parameters, many intrinsic Bayes factors still asymptotically correspond to 
procedures developed from proper objective priors known as intrinsic priors (Berger 
and Pericchi, 2001).

2.4 Other Discussions on Bayesian Regression

This review of recent (and some not so recent) advances in Bayesian linear re-
gression outlines many attractive and popular methodologies widely practiced by 
statisticians all over the world, but the list is by no means close to being exhaustive. 
Clyde and George (2004) and Heaton and Scott (2010) assess the major innovations 
in Bayesian regression from a different perspective. Immense volumes of commendable 
research on the regression problem have already been completed within the Bayesian 
paradigm and there are many more new techniques emerging with time.
Chapter 3: Block $g$ Priors

Mono-shrinkage originating from a single latent scale parameter often has a detrimental effect on inference in many common priors (see Section 1.2). Use of a common scale parameter in the prior covariance matrix of the regression coefficients, as is the case in the current crop of “$g$-type” priors, is often not only questionable from a modeling perspective, but, as we show in this chapter, also produces peculiar paradoxical behavior in the resulting posterior distributions. The block $g$ priors and their mixtures are capable of fixing these troubling anomalies and they generate procedures that correspond to a sound Bayesian analysis.

3.1 Review of $g$ Priors

Suppose we have a set of responses $y = (y_1, y_2, \ldots, y_n)^T$ and a corresponding set of $p$ candidate predictors $x_1, x_2, \ldots, x_p$. We consider the traditional setting where $n > p$. Let $\gamma \in \Gamma = \{0, 1\}^p$ denote the index set of the subsets of the predictor variables to be included/excluded in a model so that under a particular model $M_\gamma$, the $i^{th}$ element of the vector $\gamma$ signifies inclusion of predictor $i$ (in model $M_\gamma$) if $\gamma_i = 1$ and exclusion if $\gamma_i = 0$. Thus, $\Gamma$ describes the collection of all the $2^p$ models possible with the $p$ predictor variables and each element $\gamma$ represents a unique model in $\Gamma$. Let $X_\gamma$ denote the $n \times p_\gamma$ design matrix and $\beta_\gamma$ denote the $p_\gamma$ length vector of
regression coefficients corresponding to model $\mathcal{M}_\gamma$. Then the linear model $\mathcal{M}_\gamma$ can be represented as:
\[
  y = 1\alpha + X_\gamma \beta_\gamma + \epsilon
\]
with $\epsilon \sim N(0, \sigma^2 I_n)$. Assume that a $g$ prior is assigned to the vector of the unknown parameters $(\alpha, \beta_\gamma, \sigma^2)$ corresponding to each model $\mathcal{M}_\gamma$ along with an additional prior on the model space $\Gamma$. To keep the implication of the intercept $\alpha$ the same across all models, without loss of generality, we center the columns of $X_\gamma$ and the response $y$ so that $1^T y = 0$ and $1^T X_\gamma = 0$. This transformation allows specification of a flat prior on the common regression parameters [Berger et al., 1998].

3.1.1 Zellner’s $g$ Prior

Zellner’s $g$ prior ([Zellner, 1986]) is specified as
\[
  \pi(\alpha, \sigma^2 \mid \mathcal{M}_\gamma) \propto \frac{1}{\sigma^2}
\]
\[
  \beta_\gamma \mid g, \sigma^2, \mathcal{M}_\gamma \sim N(0, g\sigma^2 (X_\gamma^T X_\gamma)^{-1})
\]
which results in simple closed form expressions for the marginal likelihoods and Bayes factors. It is straightforward to show that the Bayes factor for any model $\mathcal{M}_\gamma$ compared to the null model $\mathcal{M}_0$ is
\[
  BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \frac{p(y \mid \mathcal{M}_\gamma)}{p(y \mid \mathcal{M}_0)} = \frac{(1 + g)^{(n - p_\gamma - 1)/2}}{[1 + g(1 - R^2_\gamma)]^{(n-1)/2}}
\]
where $R^2_\gamma$ is the coefficient of determination for the model $\mathcal{M}_\gamma$. Under the standard squared error loss, the Bayes estimator of $\beta_\gamma$ is
\[
  \hat{\beta}_\gamma = \frac{g}{1 + g} \hat{\beta}_\gamma,LS,
\]
where $\hat{\beta}_\gamma,LS$ is the least squares estimator of $\beta_\gamma$.
The definition of the $g$ prior requires specification of the unknown parameter $g$. A variety of choices have been suggested based on different considerations. Some of the well known (fixed) $g$ priors are:

(i) **Unit information prior** (Kass and Wasserman [1995]): $g = n$.

(ii) **Risk inflation criterion** (Foster and George [1994]): $g = p^2$.

(iii) **Benchmark prior** (Fernandez et al. [2001]): $g = \max(n, p^2)$.

(iv) **Local empirical Bayes** (Hansen and Yu [2001]):

$$
\hat{g}^{EBL}_{\gamma} \text{ (for model } M_{\gamma}) = \max \left( \frac{R^2_{\gamma}/p_{\gamma}}{(1-R^2_{\gamma})/(n-p_{\gamma}-1)} - 1 , \ 0 \right).
$$

(v) **Global empirical Bayes** (George and Foster [2000]):

$$
\hat{g}^{EBG} = \max_{g > 0} \sum_{\gamma} p(M_{\gamma}) \frac{(1+g)^{(n-p_{\gamma}-1)/2}}{[1+g(1-R^2_{\gamma})]^{(n-1)/2}}.
$$

Liang et al. (2008) review the fixed $g$ priors and summarize the justifications behind the use of these specific values of $g$.

### 3.1.2 Mixtures of $g$ Priors

There are a variety of motivations for considering a “fully Bayes” approach where $g$ is modeled with a prior distribution, leading to a marginal prior for $\beta_{\gamma}$ which can be represented as a “mixture of $g$ priors”

$$
\pi(\beta_{\gamma} \mid \alpha, \sigma^2, M_{\gamma}) = \int_0^\infty N(\beta_{\gamma} \mid 0, g\sigma^2(X^T_{\gamma}X_{\gamma})^{-1}) \pi(g)dg.
$$

Careful choice of the mixing distribution can result in thick-tailed priors for $\beta_{\gamma}$ after marginalization of $g$. An early example is the Zellner–Siow prior (Zellner and Siow [1980]), which can be expressed as a mixture of $g$ priors with an inverse-gamma ($\frac{1}{2}, \frac{n}{2}$)
mixing density for $g$. Mixing over $g$ also endows the Bayes estimator of $\beta_\gamma$ with data-adaptive shrinkage of the least squares estimator:

$$\hat{\beta}_\gamma = E(\beta \mid y, M_\gamma) = E \left( \frac{g}{1+g} \mid y, M_\gamma \right) \hat{\beta}_{\gamma,LS}. \quad (3.4)$$

The quantity $E(\frac{g}{1+g} \mid y, M_\gamma)$ is often called the shrinkage factor.

A variety of prior distributions for $g$ have been considered in the literature. Zellner and Siow (1980), West (2003), Maruyama and Strawderman (2010), Maruyama and George (2011) and Bayarri et al. (2012) are notable examples.

**Hyper-$g$ Priors**

The hyper-$g$ prior was proposed by Liang et al. (2008) to overcome deficiencies of the fixed $g$ priors and to lead to a robust estimator that is not overly reliant on a particular value of $g$. The main idea is to place a proper prior on $g$ and let the data dictate the probable values that $g$ can take. The mixture $g$ prior suggested in the paper imposes the following prior on $g$:

$$\pi(g) = \frac{a - 2}{2} (1 + g)^{-a/2}, \quad g > 0.\quad a > 2$$

$a > 2$ is needed for the prior to be proper and the authors suggested anything in the range $2 < a \leq 4$ to be quite reasonable, with $a = 3$ being a default choice. The hyper-$g$ prior offers a solution to the Lindley and Information paradoxes and serves as a robust fully Bayesian solution, but it does not suppress the computational benefits of the original $g$ prior since this particular prior proposal preserves the closed form representations of the marginal likelihood and the Bayes factors. The Bayes factor under the hyper-$g$ prior can be expressed as
\[
BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \int_0^\infty (1 + g)^{(n-p_\gamma -1)/2}[1 + g(1 - R_\gamma^2)]^{-(n-1)/2}\pi(g)dg
= \frac{a - 2}{p_\gamma + a - 2} \, 2F_1\left(\frac{n - 1}{2}, 1; \frac{a + p_\gamma}{2}; R_\gamma^2\right)
\]
where \(2F_1(\cdot)\) is the Gaussian hypergeometric function.

Moreover, the Bayes estimator (posterior mean) for the hyper-\(g\) prior can be expressed as
\[
\hat{\beta}_\gamma = E(\beta \mid y, \mathcal{M}_\gamma) = E\left(\frac{g}{1 + g} \mid y, \mathcal{M}_\gamma\right) \hat{\beta}_{\gamma,LS}
\]
where \(\hat{\beta}_{\gamma,LS}\) is the least squares estimate of the parameter \(\beta\) under model \(\mathcal{M}_\gamma\).

The Bayes estimate is always a shrunken version of the least squares estimate. The quantity \(E(\frac{g}{1 + g} \mid y, \mathcal{M}_\gamma)\) is called the **shrinkage factor** since this is the fraction of the least squares estimate that the Bayes estimate shrinks to. There is a simple expression for the posterior expectation of the shrinkage factor for the hyper-\(g\) prior.

The hyper-\(g\) prior has a nice interpretation in terms of the shrinkage factor since a change of variables from \(g\) leads to \(\frac{g}{1 + g} \sim \text{Beta}(1, \frac{a}{2} - 1)\).

### 3.2 Asymptotic Evaluations: Paradoxes, Old and New

While the \(g\) priors described in Section 3.1.1 offer many conveniences, they are known to have several undesirable properties commonly referred to as “paradoxes.” In their purest form, the paradoxes associated with \(g\) priors are revealed when taking a limit. \cite{Liang et al. (2008)} describe two such paradoxes which arise from different limits. The first, Lindley’s paradox, relies on a limit which weakens the prior distribution. The second, the Information Paradox, relies on a limit where the signal in the data becomes stronger. Both limits hold the design \(X\) (and hence sample size) fixed. These two “old” paradoxes can be summarized as follows.
Bartlett’s Paradox/Lindley’s Paradox: Lindley’s Paradox is an anomaly associated with a fixed $g$ prior when the scale factor $g$ is intentionally chosen to be large in an attempt to make the prior weakly informative. Holding the data $(X_\gamma, y)$ fixed, the Bayes factor (3.2) comparing any arbitrary non-null model $M_\gamma$ to the null model $M_0$ approaches zero in the limit when $g \to \infty$, irrespective of the data. The full description of the paradox contrasts this undesirable behavior with the results of a classical hypothesis test (Lindley, 1957; Bartlett, 1957; Jeffreys, 1961; Liang et al., 2008).

Information Paradox: The Information paradox is associated with a strong signal in the data, as manifested by a high value of $R^2_\gamma$. Holding $(X_\gamma, \epsilon)$ fixed, let $||\beta_\gamma|| \to \infty$, so that $||\hat{\beta}_{\gamma, LS}|| \to \infty$ and $R^2_\gamma \to 1$. It follows from (3.2) that $BF(M_\gamma : M_0) \to (1 + g)^{(n-p_\gamma-1)/2}$, a finite constant. Thus the Bayes factor for $M_\gamma$ relative to $M_0$ is bounded even though the likelihood evidence in favor of $M_\gamma$ grows without bound (Zellner, 1986; Berger and Pericchi, 2001; Liang et al., 2008).

These undesirable properties can be avoided by using mixtures of $g$ priors with a careful choice of mixing distribution. Liang et al. (2008) provide sufficient conditions under which a prior $\pi(g)$ resolves the Information Paradox, and prove that the hyper-$g$ prior avoids both of the above paradoxes (as does the robust prior of Bayarri et al., 2012). While these “old” paradoxes have been studied extensively, the limits taken to produce them have further, less well known implications. The first is initially seen with the limit in Lindley’s Paradox. The second, a new paradox, follows from a modification to Liang et al. (2008)’s limit. Qualitative descriptions of these behaviors are as follows, with formal results provided in Section 3.2.2.
Essentially Least Squares Estimation (ELS): It is well-known that under the $g$ prior, the Bayes estimator of $\beta_\gamma$ in (3.3) tends to $\hat{\beta}_{\gamma,LS}$ as $g \to \infty$. Formally, we identify ELS behavior as $||\hat{\beta}_\gamma - \hat{\beta}_{\gamma,LS}|| / ||\hat{\beta}_{\gamma,LS}|| \to 0$ under some appropriate limit. In Sections 3.2.1 and 3.2.2, we consider limits which are driven by changes to the data rather than changes to the prior and show that several common mixtures of $g$ priors exhibit ELS behavior.

Conditional Lindley’s Paradox (CLP): The Conditional Lindley’s Paradox arises when comparing two models $M_{\gamma_1}$ and $M_{\gamma_2}$ with $M_{\gamma_1} \subset M_{\gamma_2}$, where $M_{\gamma_2}$ is the “correct” model. Specific asymptotics for the data (described explicitly in Section 3.2.1) yield $BF(M_{\gamma_2} : M_{\gamma_1}) \to 0$, compelling one to accept the smaller (incorrect) model.

Before connecting these behaviors to existing mixtures of $g$ priors, we describe the limits driving the phenomena.

### 3.2.1 A Conditional Information Asymptotic

We consider an asymptotic analysis of a sequence of problems, where each element in the sequence is related to the linear regression model $y = \alpha + X\beta + \epsilon$. The design matrix $X$ is an $n \times p$ matrix with full column rank, and the columns of $X$ and the response $y$ are centered. Specifically, we write the linear model as $y = \alpha + X_1\beta_1 + X_2\beta_2 + \epsilon$, where $X = (X_1, X_2)$, $X_1$ is an $n \times p_1$ matrix, $X_2$ is an $n \times p_2$ matrix and $\beta = (\beta_1^T, \beta_2^T)^T$. We construct a sequence $\{\Psi_N\}_{N=1}^\infty$ where each element $\Psi_N$ represents the linear model with

$$
\Psi_N = (X_{1(N)}, X_{2(N)}, \alpha_N, \beta_{1(N)}, \beta_{2(N)}, \epsilon_N) = (X_{1}, X_{2}, \alpha, \beta_{1(N)}, \beta_{2}, \epsilon),
$$

(3.5)
and $||\beta_1(N)|| \to \infty$ as $N \to \infty$ while $X_1, X_2, \alpha, \beta_2$ and $\epsilon$ are held fixed. This is a fixed-$n$, fixed-$p$, asymptotic, and represents a strengthening of the likelihood that is driven by one particular set of predictor variables.

We refer to this as a conditional information asymptotic, as it can be viewed as the limit that drives the Information Paradox of [Liang et al. (2008)](i.e., $||\beta|| \to \infty$) with the additional condition that a portion of $\beta$ remains fixed in the analysis. The consequences of the information limit considered in [Liang et al. (2008)] were driven by $R^2 \to 1$. The following lemma notes that the conditional information asymptotic produces the same behavior.

**Lemma 3.2.1.** Let $R^2(N)$ denote the coefficient of determination for element $N$ in the sequence $\{\Psi_N\}$ as defined in (3.5). Then $R^2(N) \to 1$ as $N \to \infty$.

The lemma follows immediately by noting that the error vector $\epsilon$ is fixed, and hence $\hat{\sigma}^2 = \frac{||y - \hat{\alpha}_{LS} - X\hat{\beta}_{LS}||^2}{n - p - 1}$ also remains unchanged. $R^2(N) = \frac{||X\hat{\beta}_{LS}||^2}{(||X\hat{\beta}_{LS}||^2 + (n - p - 1)\hat{\sigma}^2)}$ which tends to one as $||\beta_1|| \to \infty$.

### 3.2.2 A Conditional Lindley’s Paradox

In this section we investigate the behavior of several mixtures of $g$ priors under the conditional information asymptotic defined by (3.5). To streamline notation, we drop the subscript $\gamma$ from $M_\gamma$ and refer to $R^2(N)$ simply as $R^2$. The following results apply to an arbitrary model $M_\gamma$ unless otherwise mentioned. The first theorem reveals a behavior of the Bayes estimator ($\hat{\beta}$) under the hyper-$g$ prior.

**Theorem 3.2.1.** (ELS) Under the hyper-$g$ prior, for the sequence $\{\Psi_N\}$ defined in (3.5), $\frac{||\hat{\beta} - \hat{\beta}_{LS}||}{||\beta_{LS}||} \to 0$ as $N \to \infty$, provided $n \geq a + p - 1$. 
The proof of the theorem is given in Appendix A.1. The theorem indicates that when at least one of the coefficients in the model grows really large in size, the estimates from the hyper-\(g\) procedure are essentially identical to those from the ordinary least squares procedure. The failure of the hyper-\(g\) in such a situation is worrisome, and it runs counter to the conventional wisdom that, for a low-information prior, the small (near zero) coefficients should be shrunk substantially while larger coefficients should be left unchanged (Berger, 1985). For a single model, we often “know” that one or some of the regression coefficients are large, and the use of the complex procedure to generate a more sound and robust analysis than simple least squares would thus appear futile.

**Theorem 3.2.2.** (CLP) Consider the two models \(M_1\) and \(M_2\) such that

\[
M_1 : y = \alpha_1 + X_1\beta_1 + \epsilon \\
M_2 : y = \alpha_1 + X_1\beta_1 + X_2\beta_2 + \epsilon
\]  

(3.6)

where \(\beta_i\) is a vector of length \(p_i, (p_i > 0)\) for \(i = 1, 2\) and \(p_1 + p_2 = p\). Under the hyper-\(g\) prior, when \(||\beta_1|| \rightarrow \infty, (i.e., N \rightarrow \infty)\) in the sequence \(\{\Psi_N\}\) defined in (3.5) and \(n \geq a + p_1 - 1\), the Bayes factor \(BF(M_2 : M_1)\) comparing model \(M_2\) to model \(M_1\) goes to zero, irrespective of the data.

The proof of this theorem can be found in Appendix A.2. The import of the theorem is that when comparing a pair of nested models, if at least one of the regression coefficients common to both models is quite large compared to the additional coefficients in the bigger model, then the Bayes factor due to the hyper-\(g\) shows unwarranted bias toward choosing the smaller model. The theorem refers to the limiting case with the size of the common coefficients growing infinitely large, resulting in the
Bayes factor inappropriately choosing the smaller model with probability 1. This absurd behavior of the Bayes factors which we described as the CLP in Section 3.2 is particularly unsettling since no matter how significantly different from zero the additional coefficients in the bigger model are, the more important predictor(s) common to both the models will bias the Bayes factor towards inferring those additional coefficients are zero.

The behavior of the posterior of $\sigma^2$ in the sequence of problems we consider here is not the reason for the strange behavior of the Bayes factor, as some might envisage. The posterior distribution of $\sigma^2$ does not get botched up as $N$ grows, causing the likelihood to become diffuse. Rather, the posterior of $\sigma^2$ is well-behaved and converges to a proper distribution with finite center and spread.

**Corollary 3.2.1.** *Under the hyper-$g$ prior, the posterior distribution of $\sigma^2$ in the sequence of problems $\{\Psi_N\}$ defined in (3.5) converges to an $IG\left(\frac{n+1-a-p}{2}, \frac{2}{(n-p-1)\hat{\sigma}^2}\right)$ distribution when $n > a + p - 1$.*

The proof of this corollary is in Appendix A.3. The CLP coincides with Lindley’s Paradox for any fixed $g$ prior, and as such it is easily shown that fixed $g$ priors are also adversely affected by the CLP.

**3.2.3 The CLP in Other Mixtures of $g$ Priors**

The generalized $g$ prior of Maruyama and George (2011) was developed as a more general “$g$-type” prior which can be extended easily to the $p > n - 1$ case. They specify a prior on $\beta$ through a prior on the rotated vector $W^T\beta$, where $W$ is defined through the singular value decomposition $X = UDW^T$. We show that in the simple
situation where $X$ is a block orthogonal matrix, the generalized $g$ prior suffers from the CLP under the asymptotics described earlier.

**Theorem 3.2.3.** Consider the models in (3.6) with the assumption that $X_1 \perp \perp X_2$. Under the generalized $g$ prior, $BF(M_2 : M_1) \to 0$ as $N \to \infty$ in the sequence $\{\Psi_N\}$ defined in (3.5), irrespective of the data.

**Proof.** Using the expressions derived in Maruyama and George (2011) and defining the two models $M_1$ and $M_2$ as in Theorem 3.2.2, it follows that

$$BF(M_2 : M_1) = C \frac{(1 - R_1^2)^{(n-p_1-2)/2-a} (1 - Q_2^2)^{-p/2-a-1}}{(1 - R_2^2)^{(n-p-2)/2-a} (1 - Q_1^2)^{-p_1/2-a-1}},$$

where $C$ is a constant, $R_1^2 = \lambda_1^2 + \ldots + \lambda_{p_1}^2$, $R_2^2 = \tau_1^2 + \ldots + \tau_p^2$, $Q_1^2 = \sum_{i=1}^{p_1} (1 - \frac{1}{\nu_{i1}}) \lambda_i^2$ and $Q_2^2 = \sum_{i=1}^{p_1} (1 - \frac{1}{\nu_{i2}}) \tau_i^2$. Here $\lambda_i = \text{corr}(u_i, y)$, the correlation between the response $y$ and $u_i$, the $i^{th}$ principal component of $X_1$ in model $M_1$ and $\tau_i$ is the corresponding entity for model $M_2$ with $X = (X_1, X_2)$. $\nu_{i1}$ and $\nu_{i2}$ are arbitrary constants appearing in the prior covariance matrix with the restriction that $\nu_{ij} \geq \nu_{i(j+1)} \geq 1$ for all $i, j$.

For the asymptotic considered here and because of the block orthogonality assumption $(X_1 \perp \perp X_2)$, $R_1^2 \to 1$ and $R_2^2 - R_1^2 \to 0$ as in Theorem 3.2.2. Since $\nu_{ij} \geq 1$ $\forall i, j$, as $N \to \infty$, $Q_1^2 \to \eta_1$ and $Q_2^2 \to \eta_2$ for some $\eta_1, \eta_2$ satisfying $1 > \eta_1, \eta_2 \geq 0$.

$$\lim_{||\beta|| \to \infty} BF(M_2 : M_1) = \lim_{z \to 1} \lim_{q \to 0} C \frac{(1 - z)^{(n-p_1-2)/2-a} (1 - \eta_2)^{-p/2-a-1}}{(1 - z - q)^{(n-p-2)/2-a} (1 - \eta_1)^{-p_1/2-a-1}}$$

$$= C^* \lim_{z \to 1} \frac{(1 - z)^{(n-p_1-2)/2-a}}{(1 - z - q)^{(n-p-2)/2-a}}$$

$$= C^* \lim_{z \to 1} (1 - z)^{(p-p_1)/2} = 0.$$
The robust prior of Bayarri et al. (2012) also suffers from the CLP.

**Theorem 3.2.4.** Consider the models in (3.6) with \( n > p_1 + 2 \). Under the robust prior of Bayarri et al. (2012) with the recommended hyperparameters \( a = 1/2, b = 1 \) and \( \rho = \frac{1}{p+1} \), as \( N \to \infty \) in the sequence \( \{\Psi_N\} \) defined in (3.5), \( BF(M_2 : M_1) \to 0 \) irrespective of the data.

**Proof.** As before, consider the setup as in Theorem 3.2.2 and assume that \( X_1 \perp \perp X_2 \) without any loss of generality. If the design matrix is not block orthogonal, we can use the technique in the proof of Theorem 3.2.2 since the “robust” prior is also a specific mixture of a \( \varphi \) prior and transform the design to be block orthogonal. Using expressions derived in Bayarri et al. (2012)

\[
BF(M_2 : M_1) = C \left( \frac{Q_{10}}{Q_{20}} \right)^{(n-1)/2} \frac{2F_1 \left[ (p + 1)/2; (n - 1)/2; (p + 3)/2; \frac{(1-Q_{10}^{-1})(p+1)}{1+n} \right]}{2F_1 \left[ (p_1 + 1)/2; (n - 1)/2; (p_1 + 3)/2; \frac{(1-Q_{20}^{-1})(p_1+1)}{1+n} \right]}
\]

where \( Q_{10} = 1 - R_1^2 \) and \( Q_{20} = 1 - R_2^2 \).

Since \( 2F_1(a; b; c; z) = 2F_1(b; a; c; z) \),

\[
BF(M_2 : M_1) = C \left( \frac{Q_{10}}{Q_{20}} \right)^{(n-1)/2} \frac{2F_1 \left[ (n - 1)/2; (p + 1)/2; (p + 3)/2; \frac{(1-Q_{10}^{-1})(p+1)}{1+n} \right]}{2F_1 \left[ (n - 1)/2; (p_1 + 1)/2; (p_1 + 3)/2; \frac{(1-Q_{20}^{-1})(p_1+1)}{1+n} \right]}
\]

\[
= C \left( \frac{1 - z}{1 - z - q} \right)^{(n-1)/2} \frac{2F_1 \left[ (n - 1)/2; (p + 1)/2; (p + 3)/2; \frac{(z+q)(p+1)}{(1+n)(1-z-q)} \right]}{2F_1 \left[ (n - 1)/2; (p_1 + 1)/2; (p_1 + 3)/2; \frac{z(p_1+1)}{(1+n)(1-z)} \right]}
\]

\[
= C^* \left( \frac{1 - z}{1 - z - q} \right)^{(n-1)/2} \int_0^1 t^{(n-1)/2} \left( 1 + \frac{t(z+q)(p+1)}{(n+1)(1-z-q)} \right)^{-n/2} dt
\]

\[
= \int_0^1 t^{(p-1)/2} \left( 1 + \frac{t(z(p_1+1))}{(n+1)(1-z)} \right)^{-n/2} dt
\]

where \( z = R_2^2, q = R_2^2 - R_1^2 \), and \( C \) and \( C^* \) are constants. Here \( N \to \infty \) implies \( z \uparrow 1 \) and \( q \downarrow 0 \). This is because \( R_2^2 = R_1^2 + \frac{(x_2 \tilde{\theta}_2)^T x_2 \tilde{\theta}_2}{y^T y} \) due to the block orthogonal design.
Hence \( \lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) \)

\[
= \lim_{z \to 1} \int_{0}^{1} t^{(p-1)/2} \left(1 - z - q + \frac{t(z+q)(p+1)}{n+1}\right)^{-\frac{n-1}{2}} dt
\]

\[
= \lim_{z \to 1} C^* \int_{0}^{1} t^{(p-1)/2} \left(1 - z + \frac{tz(p+1)}{n+1}\right)^{-\frac{n-1}{2}} dt
\]

\[
= \lim_{z \to 1} K \int_{0}^{1} t^{(p-1)/2} (B + t)^{-\frac{n-1}{2}} dt = 0. \quad (3.7)
\]

where \( K \) is a constant, \( B = B(z) = \frac{(n+1)(1-z)}{zp+1} \) and \( B^* = B^*(z) = \frac{(n+1)(1-z)}{(p+1)z} \). Clearly both \( B \) and \( B^* \) go to zero as \( z \to 1 \) and \( \frac{B}{B^*} = \frac{p+1}{p+1} = 1 \) for all \( z \).

The limit in (3.7) is zero when \( n > p_1 + 2 \) (see Appendix A.4 for a proof). \( \square \)

The performance of these popular and widely used priors can thus be criticized on the basis of the CLP. The introduction of the block hyper-\( g \) priors in Section 3.3.1 is meant to rectify their flaws, hopefully leading to an improved regression analysis and to a better quantification of uncertainty for the parameters in the model.

### 3.2.4 Avoiding ELS and CLP Behaviors

The ELS and CLP behaviors described in Section 3.2.2 for mixtures of \( g \) priors arise as a result of the use of a single, latent scale parameter \( g \) that is common to each predictor variable. In order for the model to fit the data in the presence of one (or more) large coefficients, \( g \) must be large (with high probability). Because \( g \) affects estimation of all coefficients (3.4), this has the side-effect that small coefficients are not shrunk, producing ELS behavior. The CLP can be explained by an argument
similar to the one that explains Lindley’s Paradox: as the common parameter \( g \) is
driven to be larger and larger (by a portion of the data, in our case) the diminishing
 corresponding prior mass in the neighborhood near zero containing any small, nonzero
 coefficients effectively rules out these predictors.

As we show in the next section, these behaviors can be avoided through the use
of multiple latent mixing parameters in place of a single, common \( g \). This approach
has a connection to the concept of “local shrinkage,” which has a rich history in the
study of the related normal means problem, e.g. Strawderman (1971) and Berger
(1980). Recent research in this area includes Scott and Berger (2006), Carvalho et al.
(2010), Scott and Berger (2010), Polson and Scott (2012b) and Bhattacharya et al.
(2014). The use of multiple latent scale parameters in regression settings has typically
focused on ridge-regression like settings where regression coefficients are conditionally
independent \textit{a priori} (e.g., Polson and Scott 2011, Armagan et al. 2013). Polson
and Scott (2012a) consider local shrinkage in regression where the local shrinkage
parameters are attached to linear combinations of the regression coefficients. A similar
setting is considered by West (2003).

### 3.3 Block \( g \) Priors

Our approach is to endow collections of regression coefficients with their own,
independent, mixture of \( g \) priors. Having latent scale parameters \( g_i \) that are local to
collections of coefficients result in models that avoid \textit{ELS} and \textit{CLP} behavior under the
conditions described in Section 3.4. The extreme case where each predictor variable
is associated with its own \( g_i \) was described, but not pursued, by Liang et al. (2008)
as representing “scale mixtures of independent \( g \) priors.” The approach we propose
emerges as a more general version of this idea with added theoretical underpinning related to ELS and the CLP.

Assume that the design matrix $X$ of dimension $n \times p$ is of full column rank. We build a block $g$ prior distribution by partitioning the predictors into $k$ blocks, $X = (X_1, X_2, \ldots, X_k)$, where $X_i$ is a submatrix of dimension $n \times p_i$, $i = 1, 2, \cdots, k$. The subscript $\gamma$ is suppressed here to simplify notation. The regression setup for the block $g$ prior is:

$$y | \alpha, \beta, \sigma^2 \sim N(\alpha 1 + X\beta, \sigma^2 I)$$

$$\beta | g, \sigma^2 \sim N(0, A\sigma^2)$$

and $\pi(\alpha, \sigma^2) \propto \frac{1}{\sigma^2}$ (3.8)

where $A$ is a block diagonal matrix defined as

$$A = \begin{pmatrix}
  g_1(X_1^T X_1)^{-1} & 0 & \cdots & 0 \\
  0 & g_2(X_2^T X_2)^{-1} & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & g_k(X_k^T X_k)^{-1}
\end{pmatrix}.$$

The $k$ distinct groups of predictors are chosen to be independent a priori and the separate scales $g_i$ are meant to allow differential shrinkage on distinct blocks with the amount of shrinkage on a block governed almost entirely by the characteristics of the concerned group. When the design matrix $X$ is block orthogonal with $k$ orthogonal blocks $X_1, X_2, \ldots, X_k$ and $g = (g_1, g_2, \ldots, g_k)^T = g_1 \mathbf{1}_k$, the block $g$ prior reduces to the ordinary $g$ prior.

The block $g$ prior is invariant to a linear reparameterization of the predictors within a particular block (see Appendix A.5 for a proof of the invariance property).
This ensures that coding the predictors differently within an arbitrary block does not result in a new prior, as one would expect in any coherent prior specification.

### 3.3.1 The Block Hyper-$g$ Prior

The block hyper-$g$ prior arises as a specific mixture of the block $g$ priors with the mixing distribution on each component of the vector $g$ resembling the mixing distribution of $g$ in an ordinary hyper-$g$ prior. The mixing distributions on the $g_i$’s are also chosen to be a priori independent so that the data adaptive robustness of $g_i$ has minimal effect on any block $j \neq i$. The prior on $g$ is

$$g \sim \prod_{i=1}^{k} \pi_i(g_i)$$

and

$$\pi_i(g_i) = \frac{a-2}{2} (1 + g_i)^{-a/2}, \quad g_i > 0. \quad (3.9)$$

We follow the recommendations in Liang et al. (2008) regarding the choice of the hyperparameter $a$ and prescribe $a$ to be bounded by 2 and 4 with $a = 3$ the default choice in absence of any prior information. It is possible to allow for different hyperparameters $a_i, i = 1, 2, \ldots, k$ for each of the individual blocks, but this would only complicate the analysis and choosing a default value of $a_1 = a_2 = \ldots = a_k = a$ with $a = 3$ or $a = 4$ seems to achieve good results in applications.

### 3.3.2 Sampling from the Block Hyper-$g$ Prior

An important factor working in the favor of the ordinary hyper-$g$ prior is the availability of closed form expressions for marginal likelihoods and Bayes factors. Our formulation of the block hyper-$g$ prior does not allow such simple analytic solutions for the marginal likelihood under a general design, but we suggest three efficient
ways of drawing samples from the posterior to aid inference. First, a Gibbs sampler for the block hyper-$g$ is simple since most of the full conditionals are standard distributions admitting efficient sampling techniques. The second approach relies on marginalization of all parameters other than $g$. Conditional on $g$, the model (3.8) is of completely conjugate form. This allows one to draw samples from the posterior distribution of $g$ using a Metropolis–Hastings algorithm. Third, when $k$ is small, direct Monte Carlo sampling of $g$ is effective. Estimation of features of the model such as the marginal distribution of $\beta$ is improved with use of distributions conditional on $g$, whether viewed as Rao–Blackwellization or mere use of conditional distributions.

Gibbs Sampling Steps

The full conditionals in the Gibbs sampling steps are

I. **Full conditional for $\beta$, defined on $\mathbb{R}^p$:**

$$
\pi(\beta \mid \alpha, \sigma^2, g, y) = N(\beta \mid (X^T X + A^{-1})^{-1}X^T(y - \alpha 1), (X^T X + A^{-1})^{-1}\sigma^2).
$$

II. **Full conditional for $\sigma^2$, defined on $(0, \infty)$:**

$$
\pi(\sigma^2 \mid \alpha, \beta, g, y) = IG(\sigma^2 \mid \frac{n + p}{2}, \frac{2}{(y - \alpha 1 - X\beta)^T(y - \alpha 1 - X\beta) + \beta^T A^{-1} \beta}).
$$

III. **Full conditional for $\alpha$, defined on $\mathbb{R}$:**

$$
\pi(\alpha \mid \beta, \sigma^2, g, y) = N(\alpha \mid \frac{1}{n}1^T(y - X\beta), \frac{\sigma^2}{n}).
$$

IV. **Full conditional for $g$, defined on $(0, \infty)^k$:**

$$
\pi(g \mid \alpha, \beta, \sigma^2, y) \propto \prod_{i=1}^k \left[ \exp \left( \frac{-1}{2g_i\sigma^2} \beta_i^T(X_i^T X_i)\beta_i \right) g_i^{-p_i/2}(1 + g_i)^{-\alpha/2} \right].
$$

The full conditional distribution for $g$ is not a standard distribution, unlike the other full conditionals, and one has to resort to Metropolis–Hastings methods to sample.
from it. A random-walk Metropolis algorithm with a Gaussian proposal distribution performs well and allows good mixing in the Gibbs sampling steps.

Instead of generating $\alpha$ and $\beta$ samples separately, one can perform a joint update from the full conditional of $(\alpha, \beta)$. In that case steps (I) and (III) will be replaced by a single step (V).

V. **Full conditional for $(\alpha, \beta)$, defined on $\mathbb{R}^{p+1}$:**

$$
\pi(\alpha, \beta \mid \sigma^2, g, y) = N \left( \alpha, \beta \right) \left| \begin{pmatrix} 1^T X \\ X^T 1 \\ X^T X + A^{-1} \end{pmatrix} \right|^{-1} \left( \begin{pmatrix} 1^T X \\ X^T 1 \\ X^T X + A^{-1} \end{pmatrix} \right) y \cdot \left( \begin{pmatrix} 1^T X \\ X^T 1 \\ X^T X + A^{-1} \end{pmatrix} \right)^{-1} \sigma^2.
$$

We recommend using steps (II), (IV) and (V) in the Gibbs sampler since blocking of parameters usually leads to better and faster convergence. Sampling from the full conditionals (I), (II), (III) and (V) is quite simple and efficient, only (IV) demands a more involved sampling technique. However, this can be handled with little effort by a random-walk Metropolis method.

**Marginalized Posterior Distributions**

It is not feasible with a block hyper-$g$ prior to obtain a simple analytical expression for the marginal likelihood given a particular model. But partial marginalization is possible and given the value of $g$, we can integrate out all other parameters in the model. The joint density involving all the parameters in the model is given by:

$$
\pi(\alpha, \beta, \sigma^2, g \mid y) \propto \frac{\prod_{i=1}^{k} \pi_i(g_i)}{\sigma^{n+p+2}} \exp \left[ - \frac{1}{2\sigma^2} (y - \alpha 1 - X\beta)^T (y - \alpha 1 - X\beta) \right. \
\left. - \frac{1}{2\sigma^2} \beta^T A^{-1} \beta \right].
$$

The posterior for $g$ and the posteriors of the other parameters conditional on $g$ are obtained as:
I. Partially Marginalized Posterior Density of $\beta$

$$\pi(\beta \mid g, y) \propto \frac{1}{\left[1 + \frac{1}{n-1}(\beta - \mu_\beta)^T R^{-1}(\beta - \mu_\beta)\right]^{(n+p-1)/2}}, \beta \in \mathbb{R}^p.$$ 

This is the density for a $p$-variate $t$-distribution with degrees of freedom ($\nu = n - 1$), location parameter vector ($\mu = \mu_\beta = (X^T(I - \frac{1}{n}J_n)X + A^{-1})^{-1}X^Ty$) and scale matrix ($\Sigma = R = y^T(I - F)\Sigma_\beta$). Here $\Sigma_\beta = (X^T(I - \frac{1}{n}J_n)X + A^{-1})^{-1}$ and $F = X\Sigma_\beta X^T$.

II. Partially Marginalized Posterior Density of $\alpha$

$$\pi(\alpha \mid g, y) \propto \frac{1}{\left[1 + \frac{1}{n-1}\left(\frac{(n-1)A_0^2}{D^2}\right)\right]^{n/2}}, \alpha \in \mathbb{R}.$$ 

This is the density for a one dimensional $t$-distribution with degrees of freedom ($\nu = n - 1$), location parameter ($\mu = \tau_\alpha = \frac{1}{T(I - F_\alpha)^T} y$) and scale parameter ($\sigma = \frac{D^2}{(n-1)A_0^2}$). Here $F_\alpha = X(X^TX + A^{-1})^{-1}X^T$, $D = y^T(I - F_\alpha)y$, $D^2 = y^T(I - F_\alpha)y - \frac{|1^T(I - F_\alpha)y|^2}{1^T(I - F_\alpha)y}$ and $A_0^2 = 1^T(I - F_\alpha)y$.

III. Partially Marginalized Posterior Density of $\sigma^2$

$$\pi(\sigma^2 \mid g, y) \propto \frac{1}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2}y^T(I - F)y\right], \sigma^2 > 0.$$ 

This is the density of an inverse-gamma distribution with shape parameter ($\alpha = \frac{n - 1}{2}$) and scale parameter ($\beta = \frac{2}{y^T(I - F)y}$). $F = X(X^TX + A^{-1})^{-1}X^T$ as defined in (I).

IV. Posterior Density of $g$

$$\pi(g \mid y) \propto \frac{\prod_{i=1}^k g_i^{-p_i/2}(1 + g_i)^{-a/{2}}}{\left|y^T(I - F)y\right|^{(n-1)/2}}|\Sigma_\beta|^{1/2}, g \in (0, \infty)^k.$$ 

As before, $\Sigma_\beta = (X^T(I - \frac{1}{n}J_n)X + A^{-1})^{-1}$ and $F = X\Sigma_\beta X^T$.

The usefulness of these partly marginalized posterior distributions is evident by examining the structure of the conditional densities (I), (II) and (III), all of which are familiar distributions with known moment expressions.
3.3.3 Expansion with respect to Least Squares Estimates under the Block $g$ prior

The Bayes estimate of $\beta$ under the ordinary $g$ prior is $E(\beta \mid y) = \frac{g}{1+g} \hat{\beta}_{LS}$, which can never be greater than the least squares estimate $\hat{\beta}_{LS}$ in magnitude. For the block $g$ prior, the Bayes estimate of $\beta$ is given by

$$\hat{\beta} = E(\beta \mid y) = (X^T X + A^{-1})^{-1} X^T y$$

with $A$ defined as in (3.8).

The posterior mean can be rewritten in terms of the least squares estimate using the Sherman–Morrison–Woodbury matrix identity (Golub and Van Loan 1996)

$$\hat{\beta} = \{ (X^T X)^{-1} - (X^T X)^{-1} [A + (X^T X)^{-1}]^{-1} (X^T X)^{-1} \} X^T y$$

$$= \hat{\beta}_{LS} - [A(X^T X) + I]^{-1} \hat{\beta}_{LS}$$

$$= \hat{\beta}_{LS} - \begin{pmatrix} (g_1 + 1) I_{p_1} & g_1 T_{12} & \cdots & g_1 T_{1k} \\ g_2 T_{21} & (g_2 + 1) I_{p_2} & \cdots & g_2 T_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ g_k T_{k1} & g_k T_{k2} & \cdots & (g_k + 1) I_{p_k} \end{pmatrix}^{-1} \hat{\beta}_{LS}$$

where $T_{ij} = (X_i^T X_i)^{-1} X_i^T X_j$.

While the $g$ prior always leads to contraction of the Bayes estimator, the block $g$ prior can result in either contraction or expansion depending on the design matrix $X$, the response $y$, and the vector $g = (g_1, g_2, \cdots, g_k)^T$. For a block orthogonal design where each $X_i \perp \perp X_j$, however, the Bayes estimator $\hat{\beta}$ will always be shrunk with respect to the least squares estimate. In this case $T_{ij} = 0$, $\forall i, j$ and hence

$$\hat{\beta}^T = \left( \frac{g_1}{1+g_1} \hat{\beta}_{1,LS}^T, \ldots, \frac{g_k}{1+g_k} \hat{\beta}_{k,LS}^T \right),$$

so that each block of the Bayes estimate contracts relative to the least squares estimate. The amount of shrinkage varies from block to block, in contrast to the behavior under the $g$ prior where all coefficients experience
the same degree of contraction. An illustration of expansion of the estimate for a simple case is provided below.

**Expansion in the Two Groups (Blocks) case**

Consider the simplest two blocks case with the regression model defined as

$$
y = \alpha 1 + x_1 \beta_1 + x_2 \beta_2 + \epsilon$$

where $y, x_1,$ and $x_2$ are $n \times 1$ vectors. Further assume that the predictors have been standardized so that $x_i^T x_i = 1, i = 1, 2$ and define $x_1^T x_2 = a$. Here, $a$ is the correlation between the two predictors and hence $0 \leq |a| \leq 1$. The regression model is given in (3.8). The prior on $\beta = (\beta_1, \beta_2)^T$ is $\beta \sim N(0, A\sigma^2)$ with $A = \begin{pmatrix} g_1 & 0 \\ 0 & g_2 \end{pmatrix}$ since $x_i^T x_i = 1$.

In this situation, the Bayes estimator of the regression coefficients is

$$\hat{\beta} = \hat{\beta}_{LS} - \frac{1}{g_1 g_2 + g_1 + g_2 + 1 - g_1 g_2 a^2} \begin{pmatrix} g_2 + 1 & -g_1 a \\ -g_2 a & g_1 + 1 \end{pmatrix} \hat{\beta}_{LS}.$$ 

The conditions for expansion of the individual coordinates ($\hat{\beta}_1$ and $\hat{\beta}_2$) of the Bayes estimate under squared error loss can be derived explicitly for this simple situation in terms of the coordinates ($\hat{\beta}_{1,LS}$ and $\hat{\beta}_{2,LS}$) of the least squares estimate.

**Conditions leading to expansion in $\hat{\beta}_1$:**

1. $a \hat{\beta}_{2,LS} < 0$ and $\frac{g_1 a \hat{\beta}_{2,LS}}{g_2 + 1} < \hat{\beta}_{1,LS} < -\frac{g_1 a \hat{\beta}_{2,LS}}{2g_1 g_2 (1 - a^2) + 2g_1 + g_2 + 1}$
2. $a \hat{\beta}_{2,LS} > 0$ and $\frac{g_1 a \hat{\beta}_{2,LS}}{g_2 + 1} > \hat{\beta}_{1,LS} > -\frac{g_1 a \hat{\beta}_{2,LS}}{2g_1 g_2 (1 - a^2) + 2g_1 + g_2 + 1}$

**Conditions leading to expansion in $\hat{\beta}_2$:**

1. $a \hat{\beta}_{1,LS} < 0$ and $\frac{g_2 a \hat{\beta}_{1,LS}}{g_1 + 1} < \hat{\beta}_{2,LS} < -\frac{g_2 a \hat{\beta}_{1,LS}}{2g_1 g_2 (1 - a^2) + 2g_2 + g_1 + 1}$
2. $a \hat{\beta}_{1,LS} > 0$ and $\frac{g_2 a \hat{\beta}_{1,LS}}{g_1 + 1} > \hat{\beta}_{2,LS} > -\frac{g_2 a \hat{\beta}_{1,LS}}{2g_1 g_2 (1 - a^2) + 2g_2 + g_1 + 1}$
The regions of expansion for the first and second coordinates of \( \hat{\beta} \) are disjoint, implying that there can never be expansion in both coordinates simultaneously. This restriction generalizes in the \( k \) blocks scenario to guarantee that not all \( k \) components can simultaneously expand relative to the least squares estimator. Note that as \( a \) becomes smaller, the regions of expansion begin to shrink and for \( a = 0 \) which corresponds to the case of blockwise orthogonality, the regions of expansion cease to exist. To understand the effects of \( g \) and \( a \), we consider specific values of \( a, g_1 \) and \( g_2 \) and observe how the regions of coefficient expansion change with these parameters.

![Graphs showing regions of expansion](image)

Figure 3.1: Regions of expansion of the Bayes estimator: Expansion of \( \hat{\beta}_1 \) occurs in areas shaded with vertical lines and expansion of \( \hat{\beta}_2 \) occurs in areas shaded with horizontal lines.

- **Effect of \( g \)**

As is evident from Figures 3.1 and 3.2, the regions of coefficient expansion are determined mainly by the ratio of the \( g_i \) values, and not so much by the individual values of \( g_1 \) and \( g_2 \). For a constant value of \( a \), the regions of expansion grow larger as
the discrepancy between $g_1$ and $g_2$ grows larger, and the wider region corresponds to
the coordinate of $g$ with the larger value. When $g_1$ and $g_2$ are of similar magnitude,
the regions of expansion become narrow but the size of each region is quite strongly
dependent on $a$.

Figure 3.2: Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs
in areas shaded with vertical lines and expansion of $\hat{\beta}_2$ occurs in areas shaded with
horizontal lines.

- **Effect of $a$**

  Figures 3.3 and 3.4 show that for any fixed set of values of $g_1$ and $g_2$, the regions of
  expansion grow larger as $|a|$ increases. When there is a big difference in the $g$ values
  (or equivalently their ratio is quite high or quite low), the region of expansion for the
  larger coordinate of $g$ dominates the other region, and this region further grows in
  size with an increment in $|a|$. When the magnitudes of $g_1$ and $g_2$ are comparable,
  increasing the strength of the correlation has the effect of widening the regions of
  expansion for both coordinates. The sign of $a$ does not influence the size of a region.
Figure 3.3: Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines and expansion of $\hat{\beta}_2$ occurs in areas shaded with horizontal lines.

of coefficient expansion, but has an effect on the relative positions of these regions on the $\mathbb{R}^2$ plane. Changing the sign of $a$ while keeping its magnitude fixed flips the position of each region from one quadrant to its adjacent one and the resulting plot looks like a mirror image of the one with the opposite sign on $a$.

- **Effect of one $g_i$ becoming unbounded**

Suppose that one of the components of $g$, say $g_1 \to \infty$ while the other component $g_2$ is fixed at a finite value. In this case the formula for the Bayes estimate becomes $\hat{\beta}_1 = \hat{\beta}_{1,LS} + \frac{a\hat{\beta}_{2,LS}}{g_2(1-a^2)+1}$ and $\hat{\beta}_2 = \hat{\beta}_{2,LS}\left(\frac{g_2(1-a^2)}{g_2(1-a^2)+1}\right)$. The second coordinate of the Bayes estimate can never be larger than the corresponding least squares estimate in absolute value. The first coordinate on the other hand expands over a huge region covering the entire first and third quadrants and portions of the other two quadrants depending on the $g_2$ and $a$ values.
Figure 3.4: Regions of expansion of the Bayes estimator: Expansion of $\hat{\beta}_1$ occurs in areas shaded with vertical lines. $\hat{\beta}_2$ does not expand since the region of expansion is negligible.

Conditions leading to expansion in $\hat{\beta}_1$:

1. $a\hat{\beta}_{2,LS} < 0$ and $-\infty < \hat{\beta}_{1,LS} < -\frac{a\hat{\beta}_{2,LS}}{2g_2(1 - a^2) + 2}$

2. $a\hat{\beta}_{2,LS} > 0$ and $\infty > \hat{\beta}_{1,LS} > -\frac{a\hat{\beta}_{2,LS}}{2g_2(1 - a^2) + 2}$

As mentioned above, $\hat{\beta}_2$ cannot expand.

Significance of Coefficient Expansion

The $g$ prior is closely associated with the notion of shrinkage and the hyper-$g$ prior mixes over a collection of such $g$ priors. The posterior mean of $\beta$ lies between its prior mean and the maximum likelihood estimate $\hat{\beta}_{LS}$. For prediction of the response $y$ given a covariate vector $x$, the posterior predictive mean lies between the prior predictive mean and the predictive mean based on the plug-in maximum likelihood estimate $x^T\hat{\beta}_{LS}$. Whether we consider $\beta$ or prediction of $y$, each estimate would shrink toward some point, and so the hyper-$g$ shows shrinkage behavior.
The results in this subsection show that the block $g$ and block hyper-$g$ priors show shrinkage when coupled with a block orthogonal design. However, the block $g$ prior (and as a consequence any related mixture prior) does not necessarily lead to shrinkage of coefficient estimates or predictions in any general design. This is, of course, not a bad property of the priors. Rather, it illustrates the richness of Bayesian analysis which goes far and beyond stylized descriptions of effects such as shrinkage. The previous example shows that the block $g$ prior does not cause expansion in the small or moderate coefficient estimates when the scale parameters $g_i$ are carefully assigned, which is a desirable feature of any reasonable prior.

### 3.4 Asymptotic Evaluations of the Block Hyper-$g$ Prior

In this section an asymptotic analysis of the block hyper-$g$ prior is carried out, under a situation identical to the *conditional information asymptotic* described in Section 3.2.1, to present the advantages of using a prior with multiple latent scale parameters. It shows that careful selection of blocks in the block hyper-$g$ prior can resolve the ELS and CLP behaviors described in Section 3.2 associated with ordinary mixtures of $g$ priors.

Consider the same model as in (3.8) and (3.9), with $y = \alpha + X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \epsilon$. As in Section 3.2.1, define the sequence of problems $\Psi_N = (X_{1(N)}, \ldots, X_{k(N)}, \alpha_N, \beta_{1(N)}, \ldots, \beta_{k(N)}, \epsilon_N)$ so that the only quantity that changes in the sequence is the group of regression parameters $\beta_{1(N)}$. Thus,

$$\Psi_N = (X_1, \ldots, X_k, \alpha, \beta_{1(N)}, \ldots, \beta_k, \epsilon)$$

with $||\beta_{1(N)}|| \to \infty$ as $N \to \infty$. We make use of the following assumption in many of the forthcoming theoretical results.
Condition 3.4.1. The predictors and the response are centered and the design matrix is block orthogonal:

\[ \mathbf{1} \perp \mathbf{y}, X_1, X_2, \ldots, X_k \text{ and } X_i \perp X_j, \text{ where } i \neq j. \]

The assumption of block orthogonality greatly facilitates the asymptotic analysis by providing simpler expressions for many posterior summaries. The non-block orthogonal design case is considered in Chapter 4. The Bayes estimator (under squared error loss) for the regression coefficient \( \mathbf{\beta} \) given \( \mathbf{g} \) under the model assumption Condition 3.4.1 becomes:

\[
E(\mathbf{\beta} | \mathbf{g}, \mathbf{y}) = (X^T X + A^{-1})^{-1} X^T \mathbf{y} = \left( \frac{g_1}{1 + g_1} \mathbf{\hat{\beta}}_{1,LS}^T, \ldots, \frac{g_k}{1 + g_k} \mathbf{\hat{\beta}}_{k,LS}^T \right)^T
\]

where \( \mathbf{\hat{\beta}}_{i,LS} \) denotes the \( i \)th component of the least squares estimator \( \mathbf{\hat{\beta}}_{LS} \).

The posterior density of \( \mathbf{g} \) for the general regression model defined by (3.8), (3.9) simplifies under Condition (3.4.1) to

\[
\pi(\mathbf{g} | \mathbf{y}) \propto \frac{\prod_{i=1}^{k} (1 + g_i)^{-\frac{a + p_i}{2}}}{||\mathbf{y}||^{n-1} \left[ 1 - \sum_{i=1}^{k} \frac{g_i}{1 + g_i} R_i^2 \right]^{(n-1)/2}}
\]

where \( R_i^2 = \frac{\mathbf{y}^T P_{X_i} \mathbf{y}}{\mathbf{y}^T \mathbf{y}}, \ i = 1, 2, \ldots, k \) and \( P_{X_i} \) is the projection matrix for the column space of \( X_i \). Define \( t_i = \frac{g_i}{1 + g_i} \) for \( i = 1, \ldots, k \). In the block orthogonal design, each \( t_i \) represents the shrinkage factor for the \( i \)th block under a block fixed \( \mathbf{g} \) prior. Then

\[
\pi(\mathbf{t} | \mathbf{y}) \propto \prod_{i=1}^{k} (1 - t_i)^{-\frac{a + p_i}{2}} (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{a-1}{2}}.
\]

These simplified forms for the density expressions make the asymptotic analysis for the block hyper-\( \mathbf{g} \) prior more tractable. The following lemma, similar to Lemma 3.2.1, is the building block for the main results in this section.
Lemma 3.4.1. Assume that the model in (3.8) and (3.9) holds and Condition (3.4.1) is true. Then $R^2_1 \to 1$ and $R^2_i \to 0$, $\forall \ i \neq 1$ as $N \to \infty$.

The proof of this lemma is identical to that of Lemma 3.2.1. The next lemma acts as a crucial precursor to the proofs of theorems in this section and the next one.

Lemma 3.4.2. If $f_1(t_m)$ and $f_2(t_m)$ denote properly normalized pdfs on $(0, 1)$ with

$$f_1(t_m) \propto \int_{(0,1)} \prod_{i=1}^{k} (1 - t_i) \frac{a + p_i}{2} - 2 \left( 1 - \sum_{i=1}^{k} t_i R^2_i \right)^{-\frac{n-1}{2}} dt_m$$

and $f_2(t_m) \propto \int_{(0,1)} \prod_{i=1}^{k} (1 - t_i) \frac{a + p_i}{2} - 2 \left( 1 - t_j R^2_j \right)^{-\frac{n-1}{2}} dt_m$

for some $m, j \in \{1, 2, ..., k\}$, where $j$ may or may not equal $m$ and $t_{-m} = \{t_i : i \neq m\}$, then $E_{f_1}(t_m) \geq E_{f_2}(t_m)$. Strict inequality holds when $R^2_m > 0$ and $R^2_i > 0$, for at least one $i \neq m$.

The proof of the lemma is in Appendix A.6.

The next result describes the shrinkage behavior of the posterior mean of the regression coefficients under the block hyper-$g$ prior.

Theorem 3.4.1. For the regression model described by (3.8) and (3.9) and satisfying Condition (3.4.1),

$$E(\beta | y) = \begin{pmatrix} E\left( \frac{g_1}{1+g_1} \mid y \right) \tilde{\beta}_{1,LS} \\ \vdots \\ E\left( \frac{g_k}{1+g_k} \mid y \right) \tilde{\beta}_{k,LS} \end{pmatrix}.$$ 

Further assume that $n \geq a + p_1 - 1$. Then, as $N \to \infty$ in the sequence $\{\Psi_N\}$ defined in (3.10), $E\left( \frac{g_1}{1+g_1} \mid y \right) \to 1$ and, for $i \neq 1$, $E\left( \frac{g_i}{1+g_i} \mid y \right) \to \Delta_i$ with $\frac{2}{a+p_i} \leq \Delta_i < 1$.

The proof of this theorem can be found in Appendix A.7. The result shows that the amount of shrinkage is specific to the block. Under the block hyper-$g$ prior, the
mere presence of one or a few very large regression coefficients does not prevent one from shrinking moderate and small coefficients. The amount of shrinkage is driven largely by the ratio of $y^TP_Xy$ and $\hat{\sigma}^2$. The lower bound $\frac{2}{\alpha+p_i}$ occurs when $R_i^2 = 0$ in which case $\hat{\beta}_{i,LS} = 0$.

The next result shows that the block hyper-g prior with a block orthogonal design does not suffer from the CLP.

**Theorem 3.4.2.** Consider the two models $M_1$ and $M_2$ such that

\[ M_1 : y = \alpha 1 + X_1\beta_1 + \epsilon \]
\[ M_2 : y = \alpha 1 + X_1\beta_1 + X_2\beta_2 + \epsilon \]

where $\beta_i$ is a vector of length $p_i$, ($p_i > 0$) for $i = 1, 2$ and $p_1 + p_2 = p$. Assume a block hyper-g prior on $\beta$ as in (3.8) and (3.9) (with $k = 2$) and that blocks $X_1$ and $X_2$ satisfy Condition (3.4.1) on the design. When $||\beta_i|| \rightarrow \infty$ ($N \rightarrow \infty$) in the sequence $\{\Psi_N\}$ defined in (3.10), the Bayes factor $BF(M_2 : M_1)$ comparing model $M_2$ to model $M_1$ is bounded away from zero.

**Proof.** For the block hyper-g prior, the expressions for the Bayes Factors $BF(M_i : M_0)$ comparing the models $M_i$, $i = 1, 2$ to the null (intercept only) model are

\[ BF(M_2 : M_0) = \left( \frac{a-2}{2} \right)^2 \int_0^1 \int_0^1 \prod_{i=1}^{2} \left(1-t_i\right)^{a+p_i-2} \left(1-\sum_{i=1}^{2} t_i R_i^2\right)^{-\frac{n-1}{2}} dt_1dt_2, \]

and \[ BF(M_1 : M_0) = \frac{a-2}{2} \int_0^1 \left(1-t_1\right)^{a+p_1-2} \left(1-t_1 R_1^2\right)^{-\frac{n-1}{2}} dt_1. \]

Thus,

\[
BF(M_2 : M_1) = \frac{BF(M_2 : M_0)}{BF(M_1 : M_0)} \]
\[
= \frac{a-2}{2} \frac{\int_0^1 \prod_{i=1}^{2} \left(1-t_i\right)^{a+p_i-2} \left(1-\sum_{i=1}^{2} t_i R_i^2\right)^{-\frac{n-1}{2}} dt_1dt_2}{\int_0^1 \left(1-t_1\right)^{a+p_1-2} \left(1-t_1 R_1^2\right)^{-\frac{n-1}{2}} dt_1}.
\]
\[
\geq \frac{a - 2}{2} \int_0^1 \int_0^1 \prod_{i=1}^2 \left[ (1 - t_i)^{\frac{a + p_i}{2} - 2}(1 - t_i R_i^2)^{-\frac{n - 1}{2}} \right] dt_1 dt_2
\]
\[
= \frac{a - 2}{2} \int_0^1 (1 - t_2)^{\frac{a + p_2}{2} - 2}(1 - t_2 R_2^2)^{-\frac{n - 1}{2}} dt_2.
\]

The above inequality follows from the following result which holds for any \( k \in \mathbb{N} \)
\[
1 - \sum_{i=1}^k x_i \leq \prod_{i=1}^k (1 - x_i), \text{ when } 0 \leq x_i \leq 1 \forall i. \tag{3.12}
\]

As \( ||\beta|| \to \infty \), \( R_1^2 \to 1 \) and \( R_2^2 \to 0 \), and so
\[
\lim_{||\beta|| \to \infty} BF(M_2 : M_1) \geq \frac{a - 2}{2} \int_0^1 (1 - t_2)^{\frac{a + p_2}{2} - 2} dt_2 = \frac{a - 2}{a + p_2 - 2}.
\]

Thus the limiting Bayes Factor is bounded away from zero. The lower bound is decreasing in \( p_2 \), the number of additional predictors in the superset model. \( \square \)

The posterior distribution of \( \sigma^2 \) in the sequence \( \{\Psi_N\} \) under the block hyper-\( g \) prior is also well-behaved. The posterior for \( \sigma^2 \) does not converge to a standard distribution as \( N \to \infty \), but Corollaries 3.4.1 and 3.4.2 show that the center of the limit distribution is finite and has an upper bound that can be easily calculated.

**Corollary 3.4.1.** Consider a regression model of the form \( (3.8) \) and \( (3.9) \) which satisfies Condition \( (3.4.1) \) and let \( p = \sum_{i=1}^k p_i \). When \( n > k(a - 2) + p + 1 \), as \( N \to \infty \) in the sequence \( \{\Psi_N\} \) defined in \( (3.10) \), the sequence of posteriors of \( \sigma^2 \) converges to the distribution \( F(\cdot) \) with density
\[
f(\sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{n - p - 1}{2}}} \exp \left[ -\frac{(n - p - 1)\tilde{\sigma}^2}{2\sigma^2} \right] \prod_{i=2}^k \gamma \left( \frac{a + p_i}{2} - 1, \frac{(X_i\tilde{\beta}_i)^T (X_i\tilde{\beta}_i)}{2\sigma^2} \right)
\]
where \( \gamma(s,x) = \int_0^x ts^{-1}e^{-t}dt \) is the lower incomplete gamma function.

The proof can be found in Appendix A.8
Corollary 3.4.2. Consider a regression model of the form (3.8) and (3.9) which satisfies Condition (3.4.1). Then in the sequence \( \{\Psi_N\} \) defined in (3.10), \( \lim_{N \to \infty} E(\sigma^2 \mid y) \leq \frac{1}{n-1-a-p_1} [(n-p-1)\hat{\sigma}^2 + \sum_{i=2}^{k} (X_i\hat{\beta}_i)^T(X_i\hat{\beta}_i)] \) when \( n > a + p_1 + 1 \).

The proof of the corollary is in Appendix A.9. We show that the posterior expectation of \( \sigma^2 \) as \( N \to \infty \) can be bounded from above by a constant involving least squares estimates of the fixed regression parameters in the sequence. The bound on the expectation is \( \frac{1}{n-1-a-p_1} [(n-p-1)\hat{\sigma}^2 + \sum_{i=2}^{k} (X_i\hat{\beta}_i)^T(X_i\hat{\beta}_i)] \). This bound is finite and as a result, \( E(\sigma^2 \mid y) \) is finite for all the problems in the sequence \( \{\Psi_N\} \).

The upper bound for the expectation is achieved by the ordinary hyper-\( g \) prior (a block hyper-\( g \) prior with \( k = 1 \)). In such a case

\[
\lim_{N \to \infty} E(\sigma^2 \mid y) = \frac{(n-p-1)\hat{\sigma}^2}{n-p-a-1}
\]

which is consistent with Corollary 3.2.1.

3.5 Consistency of the Block Hyper-\( g \) Prior

In this section, we analyze the block hyper-\( g \) prior with respect to three existing notions of consistency: information consistency, model selection consistency and prediction consistency. All three are considered by Liang et al. (2008) with respect to the hyper-\( g \) prior, and the first two are two of the seven “criteria for Bayesian model choice” posited by Bayarri et al. (2012). The hyper-\( g \) prior was shown to be consistent, performing favorably on all three criteria, barring one trivial situation.

The authors proposed the hyper-\( g/n \) prior to fix the procedure under that special situation.
3.5.1 Information Consistency

This form of consistency is directly related to the Information Paradox described in Section 3.2. Liang et al. (2008) define a Bayesian normal linear regression model under a particular prior to be information consistent if, under an appropriate limit on the data vector \( y \) for a fixed sample size \( n \), \( R^2_\gamma \to 1 \) for model \( M_\gamma \) implies \( BF(M_\gamma : M_0) \to \infty \), where \( M_0 \) is the null model. Bayarri et al. (2012) provide a formal definition of information consistency that applies to models other than the normal linear model. The ordinary hyper-g prior is information consistent when the condition \( n \geq a + p_\gamma - 1 \) is satisfied. The following theorem confirms that the block hyper-g prior is as good as the ordinary hyper-g in regard to this form of consistency.

**Theorem 3.5.1.** Consider a regression model of the form (3.8) and (3.9) which satisfies Condition (3.4.1). The block hyper-g prior is “information consistent” when either of two sufficient conditions hold:

1. \( R^2_\gamma \to 1 \) and \( n > k_\gamma(a - 2) + p_\gamma + 1 \), where \( k_\gamma \) is the total number of blocks, \( p_\gamma = \sum_{j=1}^{k_\gamma} p_{j,\gamma} \) and \( p_{j,\gamma} \) is the size of block \( X_{j,\gamma} \) in model \( M_\gamma \), or

2. For some \( i = 1, \ldots, k_\gamma \), \( R^2_{i,\gamma} \to 1 \) and \( n \geq a + p_{i,\gamma} - 1 \), where \( R^2_{i,\gamma} \) is the component of \( R^2_\gamma \) corresponding to the \( i^{th} \) orthogonal block.

The proof of the theorem is in Appendix A.10. Note that Condition (1) provides a general form of the theorem where \( R^2_\gamma \) approaches 1 with explosion in size of any arbitrary set of coefficients. Condition (2) is a special case when the coefficients blowing up in size all belong to a single block. Condition (2) is restrictive in the sense that one needs to have all expanding predictors in a single group and this forces additonal constraints on the selection of blocks, but Condition (1) states that

\[ \]
the block hyper-$g$ prior is *information consistent* irrespective of the choice of groups. Condition (1) requires a larger sample size than does Condition (2).

### 3.5.2 Conditions and Assumptions

For the remaining two consistency results we revert to the traditional asymptotic setting where parameters are held fixed and the sample size increases. Before proceeding, we first need to fix the notion of a “true” model from which the data $y$ are assumed to have been generated. Assume that $B_T \subseteq \{1, 2, \ldots, k\}$ denotes the indices of the blocks included in the true model $M_T$, where each block has at least one non-zero coefficient. Then $M_T : y = \alpha_T 1 + \sum_{i \in B_T} X_i \beta_{i,T} + \epsilon$ denotes the true data generating process. Under the model $M_T$ there are $|B_T| = k_T$ different blocks with separate and independent hyper-$g$ priors on each block. The following basic model assumptions and conditions are used in the results below.

**Condition 3.5.1.** The $n \times p$ design matrix $X$ grows in size with the restriction that

$$\lim_{n \to \infty} \frac{1}{n} X^T X = D,$$

for some $p \times p$ positive definite matrix $D$.

The following condition is a direct consequence of Condition 3.5.1 (Maruyama and George, 2011).

**Condition 3.5.2.** The design allows the true model $M_T$ and any arbitrary model $M_\gamma \nsubseteq M_T$ to be asymptotically distinguishable:

$$\lim_{n \to \infty} \frac{1}{n} \beta_T^T X_T^T (I - P_{X_\gamma}) X_T \beta_T = V_\gamma > 0$$

where $P_{X_\gamma} = X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T$. 

59
These conditions are standard assumptions used to establish consistency of Bayesian procedures. Fernandez et al. (2001), Liang et al. (2008), Maruyama and George (2011) and Bayarri et al. (2012) also use these conditions (or slight variations) to demonstrate posterior model selection consistency and prediction consistency of their priors.

3.5.3 Model Selection Consistency

The second form of consistency we are interested in is posterior model selection consistency which describes the ability of a Bayesian procedure to recover the true model in the presence of a large number of data points. The ordinary hyper-\(g\) prior has been shown to be model selection consistent when the true model is any model but the null model. According to Fernandez et al. (2001), the definition of posterior consistency for model selection can be expressed as

\[
\pi(M_T | y) \xrightarrow{P} 1 \text{ as } n \to \infty, \text{ assuming } M_T \text{ is the true model.}
\]

It follows naturally from the relation between posterior probabilities and Bayes factors that the above consistency criterion is equivalent to:

\[
BF(M_\gamma : M_T) \xrightarrow{P} 0 \text{ as } n \to \infty, \text{ for any model } M_\gamma \neq M_T. \tag{3.13}
\]

In both the relations above, convergence in probability is with respect to the sample size growing unbounded, i.e., when \(n \to \infty\). It is simpler to check condition (3.13) in showing model selection consistency. The following theorem checks whether or not (3.13) holds for different choices of the “true” model under the block hyper-\(g\) prior.

**Theorem 3.5.2.** Consider a regression model of the form (3.8) and (3.9) which satisfies Conditions 3.4.1, 3.5.1 and 3.5.2. For any model \(M_\gamma\) such that \(M_\gamma \supset M_T\)
and all predictors in $\mathcal{M}_\gamma$ that are not in $\mathcal{M}_T$ are in blocks not in $\mathcal{M}_T$, $BF(M_\gamma : M_T) \overset{d}{\to} W_\gamma$ as $n \to \infty$ for some non-degenerate random variable $W_\gamma$. For all other models $M_\gamma$, $BF(M_\gamma : M_T) \overset{P}{\to} 0$.

The block hyper-$g$ prior is not model selection consistent, as the Bayes factor in (3.13) does not converge to zero in all situations. We refer the readers to Appendices A.11 and A.16 for a complete proof. Convergence of the Bayes factor to zero depends on the grouping of predictors in the true model as well as the arbitrary model $M_\gamma$. This is also the case for the hyper-$g$ prior of Liang et al. (2008). The defect in these priors is that, as $n \to \infty$, the priors do not stabilize. This issue for the block hyper-$g$ prior is addressed in Section 3.6.

3.5.4 Prediction Consistency

Prediction consistency is an important property for any satisfactory Bayesian procedure. Given a set of $n$ observed data points and predictor values corresponding to them, the task is to predict the unknown response $y^*$ for a new vector of predictors $x^* \in \mathbb{R}^p$. When the true model is unknown (as is usually the case in practice), the optimal prediction with respect to squared-error loss is the Bayesian model averaged prediction defined as

$$\hat{y}_n^* = E(\alpha \mid y) + \sum_\gamma \pi(M_\gamma \mid y)x^{*T}E(\beta \mid y, M_\gamma)$$

and consistency is achieved when $\hat{y}_n^* \overset{P}{\to} E(y^*) = \alpha_T + x^{*T}\beta_T$ as $n \to \infty$. The following lemma and its extension are used in the main result on prediction consistency of the block hyper-$g$ prior.
Lemma 3.5.1. Consider a regression model of the form (3.8) and (3.9) which satisfies Condition 3.4.1. When $\mathcal{M}_T$ is the true model, for any $i \in B_T$

$$\lim_{n \to \infty} \int_{(0,1)^{k_T}} \frac{g_i}{1 + g_i} \pi(g \mid \mathcal{M}_T, y) dg = 1.$$ 

The proof of the lemma is in Appendix A.12. Following the proof of the lemma, it is quite easy to derive the following result:

Corollary 3.5.1. Consider a regression model of the form (3.8) and (3.9) which satisfies Condition 3.4.1. For any model $\mathcal{M}_\gamma$ containing the true model, i.e., for any $\mathcal{M}_\gamma \supseteq \mathcal{M}_T$,

$$\lim_{n \to \infty} E\left( \frac{g_i}{1 + g_i} \mid \mathcal{M}_\gamma, y \right) = 1, \text{ if } i \in B_T.$$ 

The generalization of Lemma 3.5.1 in the form of the preceding corollary is possible because the only required condition for this extension is $0 < \lim_{n \to \infty} R^2_{i, \gamma} < 1$, which is true for any superset model $\mathcal{M}_\gamma \supseteq \mathcal{M}_T$ as long as $i \in B_T$ (see Lemma A.11.1 in Appendix A.11). The rest of the proof is identical to that of Lemma 3.5.1.

The next theorem states the main conclusion of prediction consistency for the block hyper-$g$. Even though the block hyper-$g$ posterior is not consistent in model selection, the procedure is consistent in prediction under Bayesian model averaging (BMA).

Theorem 3.5.3. Consider a regression model of the form (3.8) and (3.9) which satisfies Conditions 3.4.1, 3.5.1 and 3.5.2. The predictions under BMA are consistent for this model.
Proof. Using the same notation as in Theorem 3.5.2, we observe that block orthogonality of the design gives

\[ E(\beta \mid y, M_\gamma) = E \left[ E(\beta \mid y, g, M_\gamma) \right] \]

\[ = \int \left( \begin{array}{c} \frac{g_{i_1}}{1+g_{i_1}} \tilde{\beta}_{i_1, \gamma, \text{LS}} \\ \vdots \\ \frac{g_{i_k}}{1+g_{i_k}} \tilde{\beta}_{i_k, \gamma, \text{LS}} \end{array} \right) \pi(g \mid y, M_\gamma) dg \]

and \( E(\alpha \mid y) = \tilde{\alpha}_{\text{LS}} \)

assuming \( B_\gamma = \{i_1, i_2, \ldots, i_k\} \).

When \( M_T = M_0 \), \( \tilde{\beta}_{\gamma, \text{LS}} \to 0 \) and \( \tilde{\alpha}_{\text{LS}} \to \alpha_T \) for every \( M_\gamma \) since least squares estimators are consistent. Thus the model averaged prediction \( \hat{y}_n^* \) converges to \( E(y^*) = \alpha_T \).

Denote the set of all models belonging to Case 2C of Theorem 3.5.2 together with \( M_T \) by \( \Omega \). We have shown that, as \( n \to \infty \), \( \pi(M_\gamma \mid y) \to 0 \) for any model \( M_\gamma \not\in \Omega \). Thus, \( \lim_{n \to \infty} \sum_{\gamma : M_\gamma \in \Omega} \pi(M_\gamma \mid y) = 1 \).

When \( M_T \neq M_0 \), the least squares estimates are consistent for \( M_\gamma \in \Omega \), so that \( \tilde{\beta}_{i, \gamma, \text{LS}} \to \beta_{i,T} \) for \( i \in B_T \) and \( \tilde{\beta}_{i, \gamma, \text{LS}} \to 0 \) for \( i \not\in B_T \). Hence,

\[ \lim_{n \to \infty} \hat{y}_n^* = \alpha_T + \sum_{\gamma : M_\gamma \in \Omega} \lim_{n \to \infty} \pi(M_\gamma \mid y)x^{*T} \times \]

\[ \left( \begin{array}{c} \beta_{i_1,T} \lim_{n \to \infty} \int \frac{g_{i_1}}{1+g_{i_1}} \pi(g \mid y, M_\gamma) dg \\ \vdots \\ \beta_{i_k,T} \lim_{n \to \infty} \int \frac{g_{i_k}}{1+g_{i_k}} \pi(g \mid y, M_\gamma) dg \end{array} \right) \]

Use Lemma 3.5.1 and Corollary 3.5.1 to get \( \lim_{n \to \infty} E \left( \frac{g_{i}}{1+g_{i}} \mid M_\gamma, y \right) = 1 \forall i \in B_T \) when \( M_\gamma \in \Omega \), while \( 0 \leq E \left( \frac{g_{i}}{1+g_{i}} \mid M_\gamma, y \right) \leq 1 \) for any other \( i \). So,

\[ \lim_{n \to \infty} \hat{y}_n^* = \alpha_T + x^{*T} \beta_T \lim_{n \to \infty} \sum_{\gamma : M_\gamma \in \Omega} \pi(M_\gamma \mid y) = E(y^*) \]

indicating that the block hyper-\( g \) prior is prediction consistent under BMA. \( \square \)
3.6 The Block Hyper-$g/n$ Prior

The block hyper-$g$ prior, despite its desirable properties, can still be criticized on the basis of its asymptotic performance in model selection. While prediction consistency and information consistency are critical and desirable aspects of a sensible Bayesian procedure, it is troubling that for some choices of the “true” model the block hyper-$g$ prior cannot guarantee correct identification of the model. A similar problem of posterior model inconsistency plagues the ordinary hyper-$g$ prior when the null model is true. The hyper-$g/n$ prior was suggested by Liang et al. (2008) to eliminate the irregularity. In this section we introduce the block hyper-$g/n$ prior which is similar to the block hyper-$g$ prior, but the prior on $g$ is now scaled by the sample size. The good properties of the block hyper-$g$ are retained by the block hyper-$g/n$ prior. In addition, the block hyper-$g/n$ prior stabilizes as $n$ grows and is asymptotically consistent (in all three aspects described in Section 3.5) for any choice of the true model. All of these properties point toward the block hyper-$g/n$ prior being the preferred mixture of block $g$ priors.

Suppose we have the same design as in (3.8) with the only difference in the regression setup appearing in the form of the prior on $g$. Instead of the usual prior (3.9) on the individual components of $g$, a similar prior is placed on each component after scaling it by the sample size $n$. The prior on $g_i$ is

$$
\pi_i(g_i) = \frac{a - 2}{2n} \left(1 + \frac{g_i}{n}\right)^{-a/2}, \ i = 1, 2, ..., k.
$$

(3.14)

The $g_i$ are independent in the prior as before. We refer to the design setup (3.8) along with the prior (3.14) as the block hyper-$g/n$ prior. Under the additional Condition (3.4.1), the posterior distribution on $g$ can be expressed as
\[
\pi(g \mid y) \propto \prod_{i=1}^{k} \left\{ (1 + g_i)^{-p_i/2} \left( 1 + \frac{g_i}{n} \right)^{-a_i/2} \right\} \left[ 1 - \sum_{i=1}^{k} \frac{g_i R_i^2}{1 + g_i} \right]^{-(n-1)/2}.
\]

Define \( t_i = \frac{g_i}{1 + g_i} \), \( i = 1, 2, \ldots, k \). The posterior distribution on \( t \) is then
\[
\pi(t \mid y) \propto \prod_{i=1}^{k} \left\{ (1 - t_i)^{a_i+2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a_i}{2}} \right\} \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}}.
\]

3.6.1 The Conditional Information Asymptotic in the Block Hyper-\( g/n \) Prior

Consider the sequence of problems \( \{ \Psi_N \} \) defined in (3.10) in Section 3.4. The results in this section display that in the setting of a conditional information asymptotic, the desirable properties of the block hyper-\( g \) prior are also shared by the block hyper-\( g/n \) prior.

Lemma 3.6.1. If \( f_1(t_m) \) and \( f_2(t_m) \) denote properly normalized pdfs on \( (0, 1) \) with
\[
f_1(t_m) \propto \int_{(0,1)^{k-1}} \left[ \prod_{i=1}^{k} (1 - t_i)^{a_i+2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a_i}{2}} \right] \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt_m
\]
\[
f_2(t_m) \propto \int_{(0,1)^{k-1}} \left[ \prod_{i=1}^{k} (1 - t_i)^{a_i+2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a_i}{2}} \right] \left( 1 - t_j R_j^2 \right)^{-\frac{n-1}{2}} dt_m
\]
for some \( m, j \in \{1, 2, \ldots, k\} \), where \( j \) may or may not equal \( m \) and \( t_m = \{t_i : i \neq m\} \). Then \( E_{f_1}(t_m) \geq E_{f_2}(t_m) \) and strict inequality holds when \( R_m^2 > 0 \) and at least one \( R_i^2 > 0 \), \( i \neq m \).

Proof. The proof is identical to the proof of Lemma 3.4.2. \( \square \)

Lemma 3.6.1 is needed in the proof of the following theorem which demonstrates that the block hyper-\( g/n \) prior does not result in the Bayes estimator emulating the least squares estimator in presence of a few large predictor coefficients.
Theorem 3.6.1. For the regression model described by (3.8) and (3.14) and satisfying Condition (3.4.1),

\[ E(\beta \mid y) = \begin{pmatrix} E \left( \frac{g_1}{1+g_1} \mid y \right) \hat{\beta}_{1,LS} \\ \vdots \\ E \left( \frac{g_k}{1+g_k} \mid y \right) \hat{\beta}_{k,LS} \end{pmatrix} \]

Further assume that \( n \geq a + p_1 - 1 \). Then, as \( N \to \infty \) in the sequence \( \{\Psi_N\} \) defined in (3.10), \( E \left( \frac{g_1}{1+g_1} \mid y \right) \to 1 \) and, for \( i \neq 1 \), \( E \left( \frac{g_i}{1+g_i} \mid y \right) \to \Delta_i \) with \( \frac{2}{a+p_i} \leq \Delta_i < 1 \).

Proof. After establishing the result below, the proof should be identical to that of Theorem 3.4.1.

\[ P = \frac{\int_{(0,1)^k} \prod_{j=1}^k \left[ (1-t_j)^{a+p_j}/2 - 2 \left( \frac{1-n}{n} t_j \right)^{-2} \right] t_1 (1-t_1 R_1^2 - \frac{n-1}{2} dt \right]}{\int_{(0,1)^k} \prod_{j=1}^k \left[ (1-t_j)^{a+p_j}/2 - 2 \left( \frac{1-n}{n} t_j \right)^{-2} \right] (1-t_1 R_1^2 - \frac{n-1}{2} dt \right]} = \frac{\int_0^1 t_1 (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1}{\int_0^1 (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1} \to 1 \]

when \( N \to \infty \) or equivalently when \( R_1^2 \to 1 \).

Define \( h_1(t_1) = \frac{1}{C_1} (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} \) and \( h_2(t_1) = \frac{1}{C_2} (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} \) where \( C_1 \) and \( C_2 \) are the respective normalizing constants for the pdfs \( h_1(\cdot) \) and \( h_2(\cdot) \). For any \( n \geq 1 \), it is clear that the ratio \( \frac{h_1(t_1)}{h_2(t_1)} \) is non-decreasing in \( t_1 \) which according to the Useful Result (UR) in Appendix A.6 implies that \( h_1 \) is stochastically larger than \( h_2 \) and hence \( E_{h_1}(t_1) \geq E_{h_2}(t_1) \).

\[ \lim_{N \to \infty} P = \lim_{N \to \infty} \frac{\int_0^1 t_1 (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1}{\int_0^1 (1-t_1)^{a+p_1}/2 - 2 \left( \frac{1-n}{n} t_1 \right)^{-2} \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1} \geq \frac{\int_0^1 t_1 (1-t_1)^{a+p_1}/2 - 2 \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1}{\int_0^1 (1-t_1)^{a+p_1}/2 - 2 \left( 1-t_1 R_1^2 \right)^{-\frac{n-1}{2}} dt_1} = 1. \] (3.15)

Imitating the proof of Theorem 3.4.1 and using Lemma 3.6.1 it is easy to show that \( E \left( \frac{g_1}{1+g_1} \mid y \right) \geq P \) so that \( \lim_{N \to \infty} E \left( \frac{g_1}{1+g_1} \mid y \right) = 1 \) as stated in the theorem. The
proof for the second part of the theorem is similar to the proof of the corresponding statement in Theorem 3.4.1.

Similar to the block hyper-g prior, the block hyper-g/n prior also causes block-specific shrinkage in regression coefficients. In particular, a block of huge coefficients will not force inference to be affected by ELS behavior. The lower bound on the shrinkage factor \( \frac{2}{\sigma^2 + p_i} \) is not very tight and our investigations indicate that there are specific situations when the bound cannot be achieved at all.

**Remark.** For the regression model defined by (3.8) and (3.14) with the assumption of Condition (3.4.1), the limiting posterior distribution of \( \sigma^2 \) in the sequence \( \{\Psi_N\} \) defined by (3.10) does not have a closed form representation as it does for the ordinary hyper-g and block hyper-g priors. Nevertheless, it can be shown that \( \lim_{N \to \infty} E(\sigma^2 \mid y) \) is finite, since the bound in Corollary 3.4.2 acts as a conservative upper bound for the limit of the expectation. To prove this, we use Lemma 3.6.1 and relation (3.15), and then follow the steps in the proof of Corollary 3.4.2.

The next theorem shows that the block hyper-g/n prior does not suffer from the CLP.

**Theorem 3.6.2.** Consider the two models \( M_1 \) and \( M_2 \) such that

\[
M_1 : \ y = \alpha 1 + X_1 \beta_1 + \epsilon \\
M_2 : \ y = \alpha 1 + X_1 \beta_1 + X_2 \beta_2 + \epsilon
\]

where \( \beta_i \) is a vector of length \( p_i, (p_i > 0) \) for \( i = 1, 2 \) and \( p_1 + p_2 = p \). Assume a block hyper-g prior on \( \beta \) as in (3.8) and (3.14) (with \( k = 2 \)) and that blocks \( X_1 \) and \( X_2 \) satisfy Condition (3.4.1) on the design. When \( \|\beta_1\| \to \infty (N \to \infty) \) in the sequence \( \{\Psi_N\} \) defined in (3.10), the Bayes factor \( BF(M_2 : M_1) \) comparing model \( M_2 \) to model \( M_1 \) is bounded away from zero.
Proof. For the block hyper-\( g/n \) prior, the expressions for the Bayes factors \( BF(M_i : M_0) \) comparing the models \( M_i, i = 1, 2 \) to the null (intercept only) model are given by

\[
BF(M_2 : M_0) = \left( \frac{a - 2}{2n} \right)^2 \int_0^1 \int_0^1 \prod_{i=1}^2 \left( 1 - t_i \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_i \right)^{-\frac{a}{2}} \\
\times (1 - \sum_{i=1}^2 t_i R_i^2)^{-\frac{n-1}{2}} dt_1 dt_2
\]

and

\[
BF(M_1 : M_0) = \frac{a - 2}{2n} \int_0^1 \left( 1 - t_1 \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_1 \right)^{-\frac{a}{2}} (1 - t_1 R_1^2)^{-\frac{n-1}{2}} dt_1.
\]

Thus,

\[
BF(M_2 : M_1) = \frac{a - 2}{2n} \int_0^1 \int_0^1 \prod_{i=1}^2 \left( 1 - t_i \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_i \right)^{-\frac{a}{2}} (1 - \sum_{i=1}^2 t_i R_i^2)^{-\frac{n-1}{2}} dt_1 dt_2 \]

\[
\geq \frac{a - 2}{2n} \int_0^1 \left( 1 - t_1 \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_1 \right)^{-\frac{a}{2}} (1 - t_1 R_1^2)^{-\frac{n-1}{2}} dt_1 dt_2 \]

\[
= \frac{a - 2}{2n} \int_0^1 \left( 1 - t_2 \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_2 \right)^{-\frac{a}{2}} (1 - t_2 R_2^2)^{-\frac{n-1}{2}} dt_2.
\]

The above inequality follows from the fact that when \( 0 \leq x_i \leq 1 \forall i \),

\[
1 - \sum_{i=1}^m x_i \leq \prod_{i=1}^m (1 - x_i) \text{ for any } m \in \mathbb{N}.
\]

As \( ||\beta_1|| \to \infty, R_1^2 \to 1 \) and \( R_2^2 \to 0 \), due to which

\[
\lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) \geq \frac{a - 2}{2n} \int_0^1 \left( 1 - t_2 \right)^{\frac{a+pi}{2} - 2} \left( 1 - n - \frac{1}{n} t_2 \right)^{-\frac{a}{2}} dt_2.
\]

This lower bound keeps the Bayes factor away from zero in the limit. \(\Box\)
The lower bound for the limit of the Bayes factor is decreasing in both \( p_2 \) and \( n \). As the sample size \( n \to \infty \), the bound decreases to 0 as one would hope, for the bound must tend to 0 if we are to establish model selection consistency. A closed form, conservative lower bound for the above limit is

\[
\lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) \geq \frac{a - 2}{2n} \int_0^1 (1 - t_2)^{a + p_2 - 2} \left( 1 - \frac{n - 1}{n} t_2 \right)^{-\frac{a}{2}} dt_2 \\
> \frac{a - 2}{2n} \int_0^1 (1 - t_2)^{a + p_2 - 2} dt_2 = \frac{a - 2}{n(a + p_2 - 2)}.
\]

### 3.6.2 Consistency of the Block Hyper-\( g/n \) Prior

We suggest the block hyper-\( g/n \) prior as a remedy to the posterior model selection inconsistency of the block hyper-\( g \) prior. However, we do not wish to sacrifice the other consistency properties. Briefly, this scaled version of the block hyper-\( g \) prior inherits the information consistent and prediction consistent behavior from the block hyper-\( g \) while the dependence on \( n \) in the prior provides posterior model selection consistency. We need two extra assumptions, namely Conditions [3.5.1] and [3.5.2] from Section [3.5.2] to validate asymptotic consistency of the block hyper-\( g/n \) prior.

**Information Consistency**

The sufficient conditions needed for information consistency of the block hyper-\( g \) prior are adequate to achieve information consistency in the block hyper-\( g/n \) prior.

**Theorem 3.6.3.** Consider a regression model of the form (3.8) and (3.14) which satisfies Condition (3.4.1). The block hyper-\( g/n \) prior is “information consistent” when either of two sufficient conditions hold:

1. \( R^2_\gamma \to 1 \) and \( n > k_\gamma(a - 2) + p_\gamma + 1 \), where \( k_\gamma \) is the total number of blocks, \( p_\gamma = \sum_{j=1}^{k_\gamma} p_{j,\gamma} \) and \( p_{j,\gamma} \) is the size of block \( X_{j,\gamma} \) in model \( M_\gamma \), or
(2) For some $i = 1, \ldots, k_{\gamma}$, $R_{i,\gamma}^2 \rightarrow 1$ and $n \geq a + p_{i,\gamma} - 1$, where $R_{i,\gamma}^2$ is the component of $R_\gamma^2$ corresponding to the $i^{th}$ orthogonal block.

The proof is similar to the proof of Theorem 3.5.1 and is provided in Appendix A.13.

Model Selection Consistency

_Model selection consistency_ holds for the block hyper-$g/n$ prior for any choice of the true model. This means that as $n \rightarrow \infty$, $BF(M_\gamma : M_T) \rightarrow 0$ for any arbitrary model $M_\gamma (\neq M_T)$ and under any assumed true model $M_T$.

**Theorem 3.6.4.** Consider a regression model of the form (3.8) and (3.14) which satisfies Conditions 3.4.1, 3.5.1 and 3.5.2. The posterior distribution is consistent in model selection.

The proof of this theorem can be found in Appendix A.14. This result highlights the success of the block hyper-$g/n$ prior in the only consistency criterion where the block hyper-$g$ prior fails.

Prediction Consistency

Predictions from the block hyper-$g$ prior have been shown to be consistent under BMA. Bayesian model averaged predictions are consistent in the sense outlined in Section 3.5.4 under the block hyper-$g/n$ prior as well.

**Theorem 3.6.5.** Consider a regression model of the form (3.8) and (3.14) which satisfies Conditions 3.4.1, 3.5.1 and 3.5.2. The predictions under BMA are consistent for this model.
The proof is given in Appendix A.15. The asymptotic consistency of the block hyper-$g/n$ posterior in prediction further encourages selection of the block hyper-$g/n$ prior as a default prior choice in many Bayesian variable selection problems.

### 3.7 Application to Real Data

#### 3.7.1 US Crime Data

In this subsection, the performance of the block hyper-$g/n$ prior is examined on Ehrlich’s US crime data from 1960 which has 15 predictor variables and 47 cases resulting in a total of $2^{15} = 32,768$ possible models in the model space. The implementation of the block hyper-$g/n$ prior hinges on the choice of the group or block structure in the design matrix. In the presence of subjective information, we recommend grouping similar predictors together since all predictors in a single block share the same scale (say $g_i$) and the common scale causes their shrinkage pattern to be identical. In this application a specific group structure based on theoretical constructs specified in Ehrlich (1973) has been used to define the block $g$ prior. According to Ehrlich’s economic theory, the collection of potential predictors can be broadly classified into 3 groups (blocks 1, 2 and 3), and each of the remaining two (possibly irrelevant) predictors is assigned a separate group. The five groups according to the theoretical constructs and the predictors included in each block are:

- **Block 1: Pool of potential criminals**
  
  Number of young males, education level, labor force participation rate, unemployment rate in age group 14-24, unemployment rate in age group 35-39, sex ratio, percentage of non-whites.
• **Block 2: Rewards of crime**
  
  Income inequality, wealth of family.

• **Block 3: Consequences of crime**
  
  Probability of imprisonment, average time served in prisons, police expenditure in 1959, police expenditure in 1960.

• **Block 4: Indicator variable**
  
  Indicator for southern states.

• **Block 5: Population size**
  
  1960 state population.

Following [Liang et al. (2008)](#) and earlier analyses of the US crime data, the response and the predictor variables (excluding the one binary predictor) are first log transformed and the block hyper-$g/n$ prior is then imparted on the transformed data as described in (3.8) and (3.14) with the number of blocks $k = 5$ and $a$ is set to either of the standard choices $a = 3$ or $a = 4$. We add a further Bernoulli variable inclusion prior on the model space so that each predictor enters a model with prior probability $w$. Two different choices of $w$ are considered in this example: $w = \frac{1}{2}$ resulting in a uniform prior on the model space and $w = \frac{1}{3}$ leading to smaller models being favored in the prior. All the posterior summaries are calculated using the marginalized posterior distributions presented in Section 3.4.2, modified accordingly for the block hyper-$g/n$ prior.

Figure 3.5 shows how the two procedures—the ordinary hyper-$g$ and the block hyper-$g/n$—allocate posterior probabilities differently to the model space. It is evident that models having low posterior probabilities under the hyper-$g$ prior get severely weighted down by the block hyper-$g/n$ prior. This results in a more peaked
Figure 3.5: Comparison of \( \log(P(\mathcal{M}_\gamma \mid y)) \) under the two priors. The figures on the upper panel correspond to prior probability of inclusion of a predictor equal to \( \frac{1}{2} \) and the ones on the lower panel correspond to prior probability of inclusion equal to \( \frac{1}{3} \). The red line denotes the \( y = x \) line.

posterior on the model space under the block hyper-\( g/n \). Zooming in on the models with high posterior probabilities under both priors, one can see that a majority of these models fall on the region where the block hyper-\( g/n \) weights are larger than the ordinary hyper-\( g \) weights. The magnitude of these model probabilities on the log scale indicate that these are the models accounting for most of the posterior mass.
Investigations show that the posterior distribution over models under the hyper-$g$ prior is much flatter than it is under the block hyper-$g/n$ prior. In other words, the block hyper-$g/n$ prior escalates the posterior weights of the set of models comprising the important predictors from each separate block. Numerical evidence of this unflattening phenomenon can be acquired through entropy calculations. For a discrete distribution $P = \{p_i; i = 1, \ldots, m\}$, the entropy is calculated as $E = -\sum_{i=1}^{m} p_i \log(p_i)$ while entropy of distribution $P = \{p_i; i = 1, \ldots, m\}$ relative to distribution $Q = \{q_i; i = 1, \ldots, m\}$ is described by the Kullback-Leibler divergence $KL = \sum_{i=1}^{m} p_i \log(\frac{p_i}{q_i})$. Table 3.1 below shows (1) the entropy of the posterior distribution over the model space and (2) the entropy of the model posterior relative to the model prior distribution for different procedures. The definition of relative entropy suggests that small values imply less disparity in the distributions $P$ and $Q$ while small values of entropy imply less diffuseness (uniformity) in the distribution values. Looking at Table 3.1 it is clear that the entropy for the block hyper-$g/n$ is lower implying that the posterior distribution under the new prior is more peaked. The entropy of the model posterior relative to the model prior distribution is also larger for the block hyper-$g/n$ signifying that the block hyper-$g/n$ specification imparts more information to the model space from the likelihood.

Further entropy calculations (shown in Table 3.2) for the posterior distributions of individual variable inclusion and model size reveal the same picture. There are 15 entropy values for posterior inclusion probabilities of the $p = 15$ distinct predictor variables and Table 3.2(B) summarizes the information by presenting the average entropy/relative entropy value across all the 15 variables. The relative entropy (or average relative entropy) values for the posterior distributions of model size and marginal
Criteria | Hg \((w = \frac{1}{2})\) | Hg/n \((w = \frac{1}{2})\) | BHg/n \((w = \frac{1}{2})\) | Hg \((w = \frac{1}{3})\) | BHg/n \((w = \frac{1}{3})\) 
--- | --- | --- | --- | --- | --- 
Entropy | 7.274 | 7.155 | 6.879 | 6.794 | 6.516 

Table 3.1: Table showing entropy values for the model posterior under different priors. Hg and Hg/n stand for hyper-\(g\) and hyper-\(g/n\) respectively, BHg/n denotes the block hyper-\(g/n\) prior.

inclusion relative to their corresponding priors are higher for the block hyper-\(g/n\), again indicating that information from the data is more effectively extracted by the block hyper-\(g/n\) model. The simple entropy values under the block hyper-\(g/n\) are not always the lowest; the block hyper-\(g/n\) posteriors for certain summaries are sometimes more diffuse than the ordinary hyper-\(g\) ones.

(A) | Hg \((w = \frac{1}{2})\) | Hg/n \((w = \frac{1}{2})\) | BHg/n \((w = \frac{1}{2})\) | Hg \((w = \frac{1}{3})\) | BHg/n \((w = \frac{1}{3})\) 
--- | --- | --- | --- | --- | --- 
Entropy | 1.868 | 1.863 | 1.871 | 1.846 | 1.883 
Relative Entropy | 0.121 | 0.110 | 0.155 | 0.422 | 0.454 

(B) | Hg \((w = \frac{1}{2})\) | Hg/n \((w = \frac{1}{2})\) | BHg/n \((w = \frac{1}{2})\) | Hg \((w = \frac{1}{3})\) | BHg/n \((w = \frac{1}{3})\) 
--- | --- | --- | --- | --- | --- 
Entropy | 0.523 | 0.516 | 0.502 | 0.506 | 0.492 
Relative Entropy | 0.170 | 0.177 | 0.191 | 0.592 | 0.607 

Table 3.2: Table showing (A) entropy values for posterior distribution of model size and (B) average entropy values for posterior distributions of marginal inclusion under different priors. Hg, Hg/n and BHg/n stand for hyper-\(g\), hyper-\(g/n\) and block hyper-\(g/n\) respectively.

The Bayes estimator of the regression coefficients under squared error loss, \(E(\beta \mid y, M_\gamma)\), for a particular model \(M_\gamma\) has a quite contrasting nature under the different
Bayesian models. Defining a measure of distance \( d = 1 - \frac{||\hat{\beta} - \hat{\beta}_{LS}||}{||\hat{\beta}_{LS}||} \) plotted against log posterior probability. Prior inclusion probability \( w = \frac{1}{2} \) in the upper panel and \( w = \frac{1}{3} \) in the lower panel.

Figure 3.6: Distance \( d = 1 - \frac{||\hat{\beta} - \hat{\beta}_{LS}||}{||\hat{\beta}_{LS}||} \) plotted against log posterior probability. Prior inclusion probability \( w = \frac{1}{2} \) in the upper panel and \( w = \frac{1}{3} \) in the lower panel.
$d$ is plotted against the logarithm of model posterior probabilities. The hyper-$g$ distance shows an explicit pattern for both choices of $w$, the distinctive curves visible in the plot indicate identical model size, meaning any two models of the same size will always fall on the same curve. It is difficult to comprehend any pattern in the block hyper-$g/n$ plots other than the fact that the high posterior probability models have $d$ values close to 1, which is true for both models. This is reasonable, because there should be little difference between any of the Bayes estimators and the simple least squares estimator in the cluster of highly likely models. The haphazard clouds of points in each of the block hyper-$g/n$ plots indicate greater disparity between the block hyper-$g/n$ and the least squares estimates, caused by the different levels of shrinkage occurring within separate blocks.

The entropy calculations earlier demonstrate how the block hyper-$g/n$ prior induces a more peaked posterior distribution than the ordinary hyper-$g$. It is expected that a less diffuse posterior would diminish the effect of irrelevant variables and intensify the effect of essential ones. Table 3.3 illustrates how the block hyper-$g/n$ generates posterior inclusion probabilities with characteristics distinct from the mixtures of ordinary $g$ priors. The median probability model (MPM) proposed by [Barbieri and Berger (2004)] is defined as the model consisting of predictors, each of which has posterior inclusion probability greater than $\frac{1}{2}$. The predictors belonging to the MPM under a specific prior have been marked in the table, representing a set of highly relevant explanatory variables for predicting crime rate. The MPM is identical under the hyper-$g$ and hyper-$g/n$ priors, but for each value of $w$, the MPM under the block hyper-$g/n$ is of a bigger size. Notice that some of the insignificant predictors seem to be weighted down compared to the ordinary hyper-$g$ hinting at the proficiency
<table>
<thead>
<tr>
<th>Block Number</th>
<th>Predictors</th>
<th>Hg $(w = \frac{1}{2})$</th>
<th>Hg/n $(w = \frac{1}{2})$</th>
<th>BHg/n $(w = \frac{1}{3})$</th>
<th>Hg $(w = \frac{1}{3})$</th>
<th>BHg/n $(w = \frac{1}{3})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>log(M)</td>
<td>0.8430*</td>
<td>0.8468*</td>
<td>0.9298*</td>
<td>0.7086*</td>
<td>0.8207*</td>
</tr>
<tr>
<td>1</td>
<td>log(Ed)</td>
<td>0.9670*</td>
<td>0.9704*</td>
<td>0.9789*</td>
<td>0.9066*</td>
<td>0.9179*</td>
</tr>
<tr>
<td>1</td>
<td>log(LF)</td>
<td>0.2261</td>
<td>0.2134</td>
<td>0.2534</td>
<td>0.1296</td>
<td>0.1443</td>
</tr>
<tr>
<td>1</td>
<td>log(NW)</td>
<td>0.2279</td>
<td>0.2157</td>
<td>0.2558</td>
<td>0.1416</td>
<td>0.1488</td>
</tr>
<tr>
<td>1</td>
<td>log(U1)</td>
<td>0.2725</td>
<td>0.2615</td>
<td>0.3083</td>
<td>0.1533</td>
<td>0.1747</td>
</tr>
<tr>
<td>1</td>
<td>log(U2)</td>
<td>0.6075*</td>
<td>0.6089*</td>
<td>0.6920*</td>
<td>0.4133</td>
<td>0.4931</td>
</tr>
<tr>
<td>2</td>
<td>log(GDP)</td>
<td>0.3770</td>
<td>0.3670</td>
<td>0.5099*</td>
<td>0.2265</td>
<td>0.3242</td>
</tr>
<tr>
<td>2</td>
<td>log(Ineq)</td>
<td>0.9946*</td>
<td>0.9956*</td>
<td>0.9939*</td>
<td>0.9873*</td>
<td>0.9811*</td>
</tr>
<tr>
<td>3</td>
<td>log(Po1)</td>
<td>0.6625*</td>
<td>0.6634*</td>
<td>0.6477*</td>
<td>0.6382*</td>
<td>0.6501*</td>
</tr>
<tr>
<td>3</td>
<td>log(Po2)</td>
<td>0.4655</td>
<td>0.4572</td>
<td>0.4182</td>
<td>0.4277</td>
<td>0.3810</td>
</tr>
<tr>
<td>3</td>
<td>log(Prob)</td>
<td>0.8889*</td>
<td>0.8926*</td>
<td>0.7847*</td>
<td>0.7416*</td>
<td>0.5615*</td>
</tr>
<tr>
<td>3</td>
<td>log(Time)</td>
<td>0.3815</td>
<td>0.3747</td>
<td>0.2967</td>
<td>0.2093</td>
<td>0.1366</td>
</tr>
<tr>
<td>4</td>
<td>So</td>
<td>0.2953</td>
<td>0.2838</td>
<td>0.2260</td>
<td>0.1914</td>
<td>0.1249</td>
</tr>
<tr>
<td>5</td>
<td>log(Pop)</td>
<td>0.3848</td>
<td>0.3761</td>
<td>0.3008</td>
<td>0.2590</td>
<td>0.1688</td>
</tr>
</tbody>
</table>

Table 3.3: Table showing marginal inclusion probabilities of variables under different posteriors. Predictors marked with asterisks belong to the corresponding median probability model.

of the block hyper-$g/n$ prior to single out extraneous variables. Actually the escalation/diminishment of the marginal inclusion probabilities of predictors is greatly influenced by the choice of blocks in the design. Groups of significant (insignificant) predictors identified and placed in separate blocks have the most conspicuous increment (decline) in inclusion probabilities.

The predictive performance under the block hyper-$g/n$ prior specification on regression coefficients also looks promising. The 47 ($= n$) cases are randomly split into seven groups of size 6 and one of size 5 and an eight-fold cross-validation is performed to compare the cross-validated predictive mean squared errors (MSE) under different priors. Tables 3.4 and 3.5 display the MSE values under Bayesian model averaging.

78
Table 3.4: Table displaying the cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{2}$.

(BMA) with two different Bernoulli variable inclusion prior choices on the model space. Both the tables show that the block hyper-$g/n$ prior (with $a = 3$ or $a = 4$) does better than the ordinary hyper-$g$ and hyper-$g/n$ priors, and a lot better than the popular varieties of $g$ priors apart from the full based Zellner–Siow prior. In both situations $w = \frac{1}{2}$ and $w = \frac{1}{3}$, the full based Zellner–Siow prior has the minimum MSE and the block hyper-$g/n$ priors with $a = 4$ and $a = 3$ follow right after.

Table 3.5: Table displaying the cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{3}$.

While the block hyper-$g/n$ prior does not seem to do better than ZS-full in prediction, one should note that the block hyper-$g/n$ prior has been formulated according to
the null-based Bayes factor approach and the ZS-full prior according to the full-based approach (see Liang et al. (2008) for a discussion of the null-based and full-based Bayes factor approaches), which is internally incoherent with respect to overall prior specification. Among all the coherent null-based priors, the block hyper-\(g/n\) clearly has the best out of sample predictive performance. The choice of \(a\) seems to have an effect on the ordinary hyper-\(g\) predictions with \(a = 4\) outperforming the standard and default choice of \(a = 3\). The choice of \(a = 4\) also lowers the MSE in both situations for the block hyper-\(g/n\) predictions, but the amount of improvement is not particularly eye-catching.

3.7.2 Ozone Data

The ozone data (Breiman and Friedman, 1985) is a data set which has been used by multiple researchers over the years to test performances of different model selection techniques, both in Bayesian and frequentist frameworks. The data set consists of daily measurements in 1976 of ozone concentration and eight meteorological variables over a span of 330 days. Thus, the ozone data comprises 330 observations on 10 variables (including day as a predictor variable) collected with the specific goal of predicting future ozone concentrations. A brief description of the data set is provided in Appendix A.17. Yu et al. (2011) analyzed this data and compared the predictive performances of a host of automated methods against those arising from human intervened procedures and their novel Bayesian Synthesis method. Many analyses of the ozone data in earlier research papers have been carried out with log(\(ozone\)) as the response variable. Scatterplots of the response variable against the predictors also point towards a log transformation being ideal and we pursue the same strategy in
this application. First, from a classical regression analysis of the data, we unearth a set of fairly relevant and significant explanatory variables useful to predict log of ozone concentration. Then we compare the predictive performances of a variety of Bayesian methods, including the block hyper-\(g\) priors, using the discovered set of essential predictor variables. Our findings are contrasted with the predictive sum of squared error (SSE) values from the table appearing in Yu et al. (2011). Following the strategy undertaken by Yu et al. (2011), the initial analysis to fit the regression model is executed with a random split of two-thirds (220 observations) of the data and the predictive performance of each method is tested on the remaining one-third part (110 observations).

**Classical Regression Analysis**

The scatterplot matrices of the log(\(ozone\)) response variable plotted against the other nine predictor variables show strong patterns, indicating high importance of many of the predictors. The plots suggest a sinusoidal relation between the \(day\) variable and the response. So a linear relation between log(\(ozone\)) and sine of \(day\) scaled appropriately to ensure a period of \(\pi\) over the 366 days of the year seems appropriate. The variable \(ibht\) appears to be truncated at the value of 5000, so an indicator variable \(ind.ibht\) is included to capture any disparate behavior of the response at the truncation point. Following the addition of the indicator variable and the sine transformation of the \(day\) variable (named \(sin.day\)), a backward stepwise regression results in exclusion of the less significant variables like \(ibtp, hmdt, wdsp\) and \(vdht\). Scatterplots and best subsets regression with the adjusted \(R^2\) and Mallow’s \(C_p\) criteria indicate that the same set of variables are important. Indicators for \(dgpg\) and \(vsty\) corresponding to observations at the truncated value of \(ibht\) also seem to be useful.
predictors. We name the two indicator variables as $\text{ind.ibht.dgpg}$ (obtained as the interaction between $\text{ind.ibht}$ and $\text{dgpg}$) and $\text{ind.ibht.vsty}$ (interaction between $\text{ind.ibht}$ and $\text{vsty}$). When we allow for higher level terms, variables $\text{ibht}$, $\text{vsty}$, $\text{dgpg}$ and $\text{sin.day}$ are found to admit some squared terms and interactions as significant predictors. Among all possible combinations of squared (and subsequently cubic) terms of the main effects and interactions between variables, none appeared to improve on the earlier model apart from the interaction term between $\text{sin.day}$ and $\text{sbtp}$.

The Durbin–Watson test statistic is calculated to check for autocorrelation in the log($\text{ozone}$) residuals, but the value of the statistic is never beyond the universally accepted critical values in any of the few final model choices. Classical regression and ANOVA summaries for the final model (listed below) further confirm the importance of each of the selected predictors, all of which turn out to be highly significant in the corresponding $t$ and $F$ tests. Since the $\text{ind.ibht.dgpg}$ and $\text{ind.ibht.vsty}$ variables are generated like interaction terms, one might be inclined to add the main effect $\text{ind.ibht}$ first, but it always failed to be meaningful in presence of the other important $\text{ibht}$ terms. Any Bayesian analysis performed with the model including the $\text{ind.ibht}$ term also invariably led to uniformly worse prediction errors, displaying the deterrent effect of this variable. However, the two indicator variables $\text{ind.ibht.vsty}$ and $\text{ind.ibht.dgpg}$ can be also thought of as independent predictors intended to capture unexplained variability at the truncation point of $\text{ibht}$ ($\text{ibht}=5000$). From this perspective there is no apparent need to include the $\text{ind.ibht}$ term in the model if the other two are also present. The final regression model from the classical analysis then has the following 11 predictors: $\text{sbtp}$, $\text{ibht}$, $\text{ibht}^2$, $\text{vsty}$, $\text{vsty}^2$, $\text{dgpg}$, $\text{dgpg}^2$, $\text{sin.day}$, $\text{ind.ibht.dgpg}$, $\text{ind.ibht.vsty}$ and $\text{sin.day}^*\text{sbtp}$.
Particular tests for influential observations reveal that some observations may be influential, but none of the observations display significantly large Cook’s distance values. However, the 69th observation still appears to be moderately influential and we include additional results later on removing that observation while fitting the model.

**Comparison of Different Methods in Prediction**

Yu et al. (2011) fit different Bayesian and frequentist regression models, arising from human intervention as well as from automated procedures, with a randomly chosen two-thirds of the data set and used the fitted models to predict the log(ozone) concentration at specific covariate values from the remaining one-third of the data. The predictive sum of squared errors (SSE) for the two analysts who tested predictions on Data Set 2 (the unused one-third of the data) along with those from the novel Bayesian Synthesis method and the top six automated procedures for this random split of the data have been listed in Table 3.6. The result for the Convex Synthesis prediction error has not been included in Table 3.6 since it is identical to the Bayesian Synthesis error when implementing the “once and for all” method of Yu et al. (2011).

<table>
<thead>
<tr>
<th>Procedure</th>
<th>SSE</th>
<th>Procedure</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analyst 1</td>
<td>12.31</td>
<td>Bagged CART</td>
<td>14.91</td>
</tr>
<tr>
<td>Analyst 3</td>
<td>14.21</td>
<td>LASSO</td>
<td>16.76</td>
</tr>
<tr>
<td>Mean Human Prediction Error</td>
<td>13.26</td>
<td>AIC</td>
<td>16.91</td>
</tr>
<tr>
<td>Bayesian Synthesis</td>
<td>11.93</td>
<td>Forward Stagewise</td>
<td>17.20</td>
</tr>
<tr>
<td>BART</td>
<td>11.40</td>
<td>Smoothing Splines</td>
<td>17.21</td>
</tr>
</tbody>
</table>

Table 3.6: Table from Yu et al. (2011) displaying the cross-validated predictive SSE under different methods.
We used the same split of the ozone data (*Data Set 2* being the test data) and the set of 11 candidate predictors selected in the classical regression analysis as the full model for some traditional Bayesian and block hyper-$g/n$ prior implementations. For the Bayesian analyses with 11 predictor variables, there is a total of $2^{11} = 2,048$ different models under consideration, all of which are assumed to be equally likely *a priori*. Table 3.7 presents the observed prediction errors from a few popular Bayesian methods when the entire training data of 220 observations is used to fit the model. The result for the simple least squares method is also included, displaying how the more complex procedures improve the log(*ozone*) predictions from the simplest possible linear model fit. But specification of a block hyper-$g/n$ prior is not complete without mentioning a group structure among the predictors. We present three subjectively chosen block structures for the block hyper-$g/n$ prior based on slightly different rationale.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>SSE</th>
<th>Procedure</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$-prior ($g = n$)</td>
<td>13.32</td>
<td>Hyper-$g$ ($a = 3$)</td>
<td>13.46</td>
</tr>
<tr>
<td>$g$-prior ($g = p^2$)</td>
<td>13.40</td>
<td>Hyper-$g$ ($a = 4$)</td>
<td>13.48</td>
</tr>
<tr>
<td>EB-Local</td>
<td>13.44</td>
<td>BH$g/n$ ($a = 3$) Form 1</td>
<td>13.14</td>
</tr>
<tr>
<td>EB-Global</td>
<td>13.44</td>
<td>BH$g/n$ ($a = 3$) Form 2</td>
<td>13.23</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>13.41</td>
<td>BH$g/n$ ($a = 3$) Form 3</td>
<td>13.14</td>
</tr>
<tr>
<td>ZS-Full</td>
<td>13.33</td>
<td>Classical linear model</td>
<td>13.79</td>
</tr>
</tbody>
</table>

Table 3.7: Table displaying the cross-validated predictive SSE (under BMA) for different procedures using the entire training data.

- **Group structure in the block $g$ prior:**

  **Form 1:** Block 1 - $sbtp$, $ibht$, $vsty$, $dgpg$, $sin.day$; Block 2 - $ibht^2$, $vsty^2$, $dgpg^2$;

  Block 3 - $ind.ibht.dgpg$, $ind.ibht.vsty$, $sin.day*sbtp$.  

84
**Form 2**: Block 1 - sbtp, ibht, vsty, dgpg; Block 2 - \(ibht^2, vsty^2, dgpg^2\);
Block 3 - \(ind.ibht.dgpg, ind.ibht.vsty;\) Block 4 - \(sin.day, sin.day*sbtp.\)

**Form 3**: Block 1 - sbtp, ibht, vsty, dgpg, sin.day; Block 2 - \(ibht^2, vsty^2, dgpg^2;\)
Block 3 - \(ind.ibht.dgpg, ind.ibht.vsty;\) Block 4 - \(sin.day*sbtp.\)

In Form 1 the predictors have been grouped with all the linear terms (main effects) falling in one block, the quadratic terms in a second block and the predictors resembling interaction terms in another separate block. Form 2 separates the sine terms and places both of them in a single block due to their similar behavior, while the indicator terms belong to a distinct group of their own. As before, there are two more blocks for the remaining linear and quadratic terms. Form 3 is almost identical to Form 2, the only difference comes from separating the sine terms and including the variable \(sin.day\) in the first block of main effects. Table 3.7 displays the predictive SSEs from these three forms of block hyper-\(g/n\) prior together with the SSEs from other standard Bayesian priors.

All the models exhibit an enhanced predictive performance when the apparently influential 69th observation is removed from consideration while determining the initial model fit. The predictive SSEs for the same methods in the absence of the influential observation are listed in Table 3.8 and the numbers clearly show some improvement over the corresponding values from Table 3.7.

The SSE values corresponding to all three forms of the block hyper-\(g/n\) prior are the lowest in either situation (with and without the influential case) among all the standard Bayesian methods. But the block hyper-\(g/n\) priors fail to do better than Analyst 1, Bayesian Synthesis and BART in this example. High marginal posterior
<table>
<thead>
<tr>
<th>Procedure</th>
<th>SSE</th>
<th>Procedure</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g)-prior ((g = n))</td>
<td>13.10</td>
<td>Hyper-(g) ((a = 3))</td>
<td>13.23</td>
</tr>
<tr>
<td>(g)-prior ((g = p^2))</td>
<td>13.17</td>
<td>Hyper-(g) ((a = 4))</td>
<td>13.25</td>
</tr>
<tr>
<td>EB-Local</td>
<td>13.22</td>
<td>BHg/n ((a = 3)) Form 1</td>
<td>12.93</td>
</tr>
<tr>
<td>EB-Global</td>
<td>13.22</td>
<td>BHg/n ((a = 3)) Form 2</td>
<td>13.03</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>13.19</td>
<td>BHg/n ((a = 3)) Form 3</td>
<td>12.94</td>
</tr>
<tr>
<td>ZS-Full</td>
<td>13.10</td>
<td>Classical linear model</td>
<td>13.53</td>
</tr>
</tbody>
</table>

Table 3.8: Table displaying the cross-validated predictive SSE (under BMA) for different procedures after discarding the influential observation from the training data.

inclusion probabilities of predictors under the hyper-\(g\) prior as well as the block hyper-\(g/n\) (Form 1) prior imply that all of them are quite important and for a given model, the Bayes estimates are not much different from the least squares estimate. As a result, differences between model posterior probabilities for the two Bayesian models are not that contrasting. Figure 3.7 illustrates how similar the two sets of posterior probabilities are, but some marginal inclusion probabilities still tend to be upweighted or downweighted because of the group structure.

<table>
<thead>
<tr>
<th>Hyper-(g) BHg/n</th>
<th>sbtp</th>
<th>ibht</th>
<th>(ibht^2)</th>
<th>vsty</th>
<th>(vsty^2)</th>
<th>dgpg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hyper-(g) BHg/n</td>
<td>0.996</td>
<td>0.623</td>
<td>0.986</td>
<td>0.971</td>
<td>0.352</td>
<td>0.497</td>
</tr>
<tr>
<td>(dpgg^2)</td>
<td>0.988</td>
<td>0.682</td>
<td>0.994</td>
<td>0.980</td>
<td>0.496</td>
<td>0.367</td>
</tr>
<tr>
<td>sin.day</td>
<td>ind.ibht.dgpg</td>
<td>ind.ibht.vsty</td>
<td>(sin.day^*sbtp)</td>
<td>0.835</td>
<td>0.421</td>
<td></td>
</tr>
<tr>
<td>Hyper-(g) BHg/n</td>
<td>1.000</td>
<td>0.998</td>
<td>0.937</td>
<td>0.835</td>
<td>0.421</td>
<td>0.412</td>
</tr>
<tr>
<td>(dpgg^2)</td>
<td>1.000</td>
<td>0.997</td>
<td>0.927</td>
<td>0.800</td>
<td>0.412</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.9: Marginal inclusion probabilities under the block hyper-\(g/n\) (Form 1) and the hyper-\(g\) posteriors.
Figure 3.7: Comparison of $\log(P(M_\gamma | y))$ under the hyper-$g$ and block hyper-$g/n$ (Form 1) priors. The red line denotes the $y = x$ line.

Most of the low probability models seem to have diminished importance under the block hyper-$g/n$ prior, but the difference is not very noticeable. Observe that highly relevant explanatory variables increase the posterior weights of other predictors when they appear in the same group of the block hyper-$g/n$ prior. The marginal inclusion probabilities in Table 3.9 verify the occurrence of this phenomenon in the blockwise prior implementation, which is particularly evident when looking at the strikingly fluctuating inclusion probabilities of variables $vst y^2$ and $dp g$ under the two priors.

3.8 Summary

In this chapter, we identify two novel behaviors for Bayesian regression models, Essentially Least Squares estimation and the Conditional Lindley’s Paradox, that are exhibited by many common mixtures of $g$ priors in Bayesian regression. Both
behaviors stem from the use of a single latent scale parameter that is common to all regression coefficients. We argue that ELS behavior is, in general, undesirable, as it precludes shrinkage of small coefficients in the presence of large ones. Similarly, we argue that priors exhibiting the CLP should be avoided as, asymptotically, they can provide infinite evidence in support of a false hypothesis. Our analyses are driven by a new, *conditional information asymptotic* that sheds light on a Bayesian linear model’s behavior under a strengthening of the likelihood due to one component of the model. This style of asymptotic is important in practice, as it is often the case that models are comprised of covariates with coefficients of differing magnitude. We develop the block hyper-$g$ prior and provide conditions under which the prior does not suffer from either ELS or the CLP. The block hyper-$g/n$ prior follows suit, maintaining the desirable properties of the block hyper-$g$ prior along with the philosophical advantage of converging to a fixed limit (for the prior) as $n \to \infty$. The block hyper-$g/n$ is our recommended prior as it yields consistent model selection, an essential criterion for sensible prior choice, which is lacking in the block hyper-$g$ prior.
Chapter 4: Block $g$ Priors in Non-Orthogonal Designs

4.1 Introduction

In Chapter 3, we described new paradoxes which highlight undesirable behaviors of many popular “$g$-type” priors. The principal cause for these troubling behaviors is the mono-shrinkage generated from the single latent scale parameter appearing in the prior. The block $g$ priors and their mixtures, proposed in Chapter 3, allow for poly-shrinkage and are capable of avoiding these new paradoxes. However, the theoretical benefits of the block $g$ priors are mitigated by a few practical concerns.

The primary concerns leading to the development of the block $g$ priors are theoretical, with the desire to avoid Essentially Least Squares behavior and the Conditional Lindley’s Paradox. These issues are most easily investigated when the regression design is block orthogonal. However, in a typical applied problem, apart from experiments specifically designed to ensure block orthgonality, the regression design is not block orthogonal. This suggests a need to extend the theory for the block hyper-$g$ and block hyper-$g/n$ priors.

The secondary concern with a non-orthogonal design is computational. As mentioned in Section 3.3.2, the block hyper-$g$ or block hyper-$g/n$ prior does not lead to a simple closed form expression for the marginal likelihood unlike some well-known
mixtures of ordinary $g$ priors. This results in increased computational time for posterior inference. It has been observed throughout Chapter 3 that a block $g$ prior has simpler expressions for posterior summaries when the groups or blocks of regression parameters defined in the prior have orthogonal components in the design matrix. It is advantageous to work with the simplified posterior from a block orthogonal design since it allows for faster and simpler computational algorithms for inference.

In this chapter, we introduce a modified version of the block $g$ prior where the blockwise prior is imposed on the transformed regression coefficients after reparameterizing the model to admit orthogonal blocks. This approach is similar to the one adopted by Maruyama and George (2011) in their specification of the generalized $g$ prior. The generalized $g$ prior of Maruyama and George (2011) is a particular mixture of a “$g$-type” prior applied to regression coefficients in a rotated orthogonal design. The new variation of the block $g$ prior discussed in this chapter is the usual prior (described in (3.8)) applied to the regression coefficients after transforming the design by a block Gram–Schmidt orthogonalization process. The dependence of the prior on the regression design via the covariance matrix is realized through the transformed design matrix, and not directly through the original design. This allows us to tap into the benefits of orthogonality for both theory and computation. Permutations of the blocks leading to different orthogonalizations are also considered.

4.2 Block $g$ Prior on a Transformed Design

Consider the problem of modeling the response $y$ as a function of $p$ predictor variables $x_1, x_2, \ldots, x_p$ using a linear regression model. Henceforth, we suppress the
subscript $\gamma$ when working with model $M_\gamma$ whenever there is no scope for confusion. The following display captures the relationship between the response and the predictors in the model:

$$
y = \alpha 1 + X \beta + \epsilon
$$

$$
y = \alpha 1 + X_1 \beta_1 + X_2 \beta_2 + \ldots + X_k \beta_k + \epsilon
$$

(4.1)

where $\beta_{p \times 1}$ is a vector of regression coefficients, $\alpha$ is an intercept and $\epsilon_{n \times 1}$ is a vector of mutually independent errors, each of which is assumed to have a Gaussian distribution. Here $X_{n \times p} = (X_1, X_2, \ldots, X_k)$ is an arbitrary design matrix of full column rank, i.e., $\text{rank}(X) = p$ and each $X_i$ is a $n \times p_i$ submatrix formed by grouping the $p (= \sum_{i=1}^k p_i)$ predictors in $k$ distinct blocks. For the time being, take the predictors as grouped, and assume that the groups are ordered based on relative importance (the motivation behind this ordering is discussed later), so that the collection of predictors in $X_1$ are believed to be most useful while those in $X_k$ are believed to be least useful in explaining $y$. Without loss of generality, we assume that the columns of the design matrix $X$ and response $y$ have been centered.

4.2.1 Transformation to a Block Orthogonal Design

We implement a block Gram–Schmidt process (with submatrices as individual elements instead of the usual vector elements) to orthogonalize the design matrix $X$ and generate a block orthogonal matrix $Q$ with $k$ orthogonal blocks. The block Gram–Schmidt process results in the following matrix $Q$:

$$
Q_1 = X_1
$$

$$
Q_2 = (I - P_{Q_1})X_2
$$
\[
Q_k = (I - P_{Q_1} - P_{Q_2} - \ldots - P_{Q_{k-1}}) X_k
\]  

(4.2)

where \(P_{Q_i} = Q_i (Q_i^T Q_i)^{-1} Q_i^T\) for all \(i\). Clearly \(Q_i \perp Q_j \forall i \neq j\).

It is not difficult to show that \(X\) can now be represented as \(X = QU\) where

\[
U = \begin{pmatrix}
I & T_{12} & T_{13} & \cdots & T_{1k} \\
0 & I & T_{23} & \cdots & T_{2k} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & I & T_{k-1,k} \\
0 & 0 & 0 & 0 & I
\end{pmatrix}
\]

is a block upper triangular matrix with \(T_{ij} = (Q_i^T Q_i)^{-1} Q_i^T X_j\).

The regression model (4.1) can be expressed as

\[
y = \alpha 1 + X \beta + \epsilon = \alpha 1 + QU \beta + \epsilon = \alpha 1 + QT + \epsilon
\]

such that \(\tau = U \beta\) and the assumptions on \(\epsilon\) are unchanged. This implies that

\[
\tau_i = \beta_i + \sum_{j=i+1}^{k} T_{ij} \beta_j = \beta_i + \sum_{j=i+1}^{k} (Q_i^T Q_i)^{-1} Q_i^T X_j \beta_j.
\]  

(4.3)

A direct consequence of the relation \(X = QU\) is that the block matrices \(\{X_i\}_{i=1}^{k}\) can also be expressed in terms of the orthogonal blocks \(\{Q_i\}_{i=1}^{k}\) as follows:

\[
X_1 = Q_1
\]

\[
X_2 = Q_2 + Q_1 T_{12}
\]

\[
\vdots \quad \vdots \quad \vdots
\]

\[
X_k = Q_k + Q_{k-1} T_{k-1,k} + \ldots + Q_1 T_{1k}.
\]
The order in which the blocks $X_i$ are orthogonalized in the block Gram–Schmidt process plays an important role in determining $Q$ and hence $\tau$. This cascades into inference, which is demonstrated in Section 4.3.2. As stated earlier, we will only consider the situation where the blocks are arranged according to importance and are successively orthogonalized in the same order with the most important block coming first. For our purposes, the importance of a block is equated to its contribution to the coefficient of determination ($R^2$), given previous blocks in the model.

4.2.2 The Block Orthogonalized Block $g$ Prior

The block orthogonalized block $g$ prior (BOB $g$ prior) is a block $g$ prior on the reparameterized regression model $y = \alpha 1 + Q \tau + \epsilon$ and is defined by:

$$
y \mid \alpha, \tau, \sigma^2 \sim N(\alpha 1 + Q \tau, \sigma^2 I)$$
$$
\tau \mid g, \sigma^2 \sim N(0, A_{\tau} \sigma^2)$$

(4.4)

and

$$
\pi(\alpha, \sigma^2) \propto \frac{1}{\sigma^2}
$$

where the matrix $A_{\tau}$ is defined as:

$$
A_{\tau} = \begin{pmatrix}
g_1(Q_1^T Q_1)^{-1} & 0 & \cdots & 0 \\
0 & g_2(Q_2^T Q_2)^{-1} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & g_k(Q_k^T Q_k)^{-1}
\end{pmatrix}.
$$

The block orthogonalized block hyper-$g$ (BOB hyper-$g$) and the block orthogonalized block hyper-$g/n$ (BOB hyper-$g/n$) priors are defined in the natural way by assigning prior distributions in the form of (3.9) and (3.14) respectively to the vector of scale parameters $g$. The BOB $g$ prior does not directly specify a distribution on the original regression parameters $\beta$, but the prior on $\tau$ induces a prior on $\beta = U^{-1} \tau$. The generalized $g$ prior of Maruyama and George (2011) is defined in a similar way.
with the prior on $\beta$ defined through a distribution on $W^T \beta$ instead.

$$W^T \beta \mid \sigma^2, g \sim N(0, \sigma^2 \Psi(g, \nu))$$

for a particular covariance matrix $\Psi(g, \nu)$, and $W$ is an orthogonal matrix that diagonalizes $X^T X$, i.e., $W^T (X^T X) W = D^2$ for some diagonal matrix $D$. The generalized $g$ prior is also not a direct prior on $\beta$, but it is straightforward to derive the induced prior on the original regression coefficients from the prior on $W^T \beta$.

The benefit of the prior form (4.4) is that the reparameterized regression model admits orthogonal blocks and this facilitates posterior calculations immensely. After integrating out all the remaining parameters, the posterior density of $g$ under the BOB hyper-$g$ prior is simply

$$\pi(g \mid y) \propto \prod_{i=1}^k (1 + g_i)^{-\frac{a + p_i}{2}} \left[ 1 - \sum_{i=1}^k \frac{g_i}{1 + g_i} R_i^2 \right]^{-(n-1)/2}$$

where $R_i^2 = \frac{y^T P_i \bar{y}}{y^T y}$, $i = 1, 2, \ldots, k$. It should be pointed out here that the coefficient of determination ($R^2$) for the model is split into $k$ additive components based on the transformed block orthogonal design $Q$, and not the original non-orthogonal design $X$. Hence, $R^2 = \frac{y^T P \bar{y}}{y^T y} = \sum_{i=1}^k \frac{y^T P_i \bar{y}}{y^T y} = \sum_{i=1}^k R_i^2$, where $R_i^2$ denotes the $i^{th}$ component of $R^2$ due to the covariates in the $i^{th}$ orthogonal block $Q_i$.

As before, defining $t_i = \frac{g_i}{1 + g_i}$ for $i = 1, \ldots, k$, we get the posterior

$$\pi(t \mid y) \propto \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \left( 1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}}.$$  

The posterior for $t$, rather than the posterior for $g$, is used in all the posterior computations, because it is more convenient to work with the parameter $t$. Similarly, the posterior density of $t$ under the BOB hyper-$g/n$ prior is

$$\pi(t \mid y) \propto \prod_{i=1}^k \left[ (1 - t_i)^{-\frac{a + p_i}{2} - 2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{2}{2}} \right] \left( 1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}}.$$  

94
This leads to the following expressions for Bayes factors:

\[
BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \left(\frac{a-2}{2}\right)^k \int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{a/p_i - \frac{1}{2}} (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt
\]

for the BOB hyper-\(g\) prior and

\[
BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \left(\frac{a-2}{2}\right)^k \int_{(0,1)^k} \prod_{i=1}^{k} \left[(1 - t_i)^{a/p_i - \frac{1}{2}} (1 - \frac{n-1}{n} t_i)^{-\frac{n}{2}}\right] \\
\times (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt
\]

for the BOB hyper-\(g/n\) prior, where \(\mathcal{M}_0\) denotes the null (intercept only) model.

An important observation here is that a BOB \(g\) prior or a mixture of BOB \(g\) priors is not invariant to the order of orthogonalization in the block Gram–Schmidt process. As a result, in a general non-orthogonal design, the posterior distribution for regression parameters will vary if the positions of two or more blocks \(X_i\) from the original design are switched in the orthogonalization procedure. However, a BOB \(g\) prior is invariant to model reparameterizations by affine transformation within blocks, a characteristic which is shared by the block \(g\) prior as well. This follows immediately upon noting that the column space of \(Q_i\) is unchanged by this type of transformation.

### 4.3 Properties of the BOB Hyper-\(g\) Prior

The motivation behind the development of the block hyper-\(g\) and block hyper-\(g/n\) priors is to fix the unappealing behavior of the single scale \(g\) priors in presence of regression coefficients of greatly differing sizes. The two new paradoxes presented in Section 3.2, the CLP and ELS, reveal the shortcomings of the usual “\(g\)-type” priors. The block hyper-\(g/n\) is our recommended prior, as it enjoys a variety of sound properties and conforms to the prior selection criteria suggested by Bayarri et al.
under the assumption of a block orthogonal design. We introduce the BOB hyper-$g$ and BOB hyper-$g/n$ priors in this chapter to alleviate the computational burden of using mixtures of block $g$ priors and to relinquish the restriction of a block orthogonal design for producing a few crucial theoretical results. But it would be a cause for concern if the benefits of the block hyper-$g/n$ prior are lost in exchange for the computational advantage of the BOB $g$ priors. In this section we investigate the effectiveness of the BOB hyper-$g$ and BOB hyper-$g/n$ priors, with a particular emphasis on their ability to cope with ELS and the CLP, the two main defects that inspired the development of the block $g$ priors.

4.3.1 Effect of the CLP and ELS

Consider a sequence of regression problems \( \{ \Psi_N \} = \{ X_1(N), \ldots, X_k(N), \alpha_N, \beta_1(N), \ldots, \beta_s(N), \epsilon_N \} \) defined in an identical way as the sequences in Sections 3.2.1 and 3.4. As before, the only quantity that changes in the sequence \( \{ \Psi_N \} \) is \( \beta_s \) for some \( s \in \{ 1, 2, \ldots, k \} \) with \( ||\beta_s(N)|| \to \infty \) as \( N \to \infty \). Hence,

\[ \Psi_N = \{ X_1, \ldots, X_k, \alpha, \beta_1, \ldots, \beta_s(N), \ldots, \beta_k, \epsilon \} \]

(4.5)

with \( \lim_{N \to \infty} ||\beta_s(N)|| = \infty \).

We make the following assumption to ensure convergence of all the additive components of \( R^2 \) (the \( R^2_i \) terms) in the sequence \( \{ \Psi_N \} \) in (4.5) to fixed limits.

**Condition 4.3.1.** For every \( i = 1, 2, \ldots, k \), \( ||\tau_i||^2 = O(M_i^2) \), where \( M_i = \max(||\beta_i||, ||\beta_{i+1}||, \ldots, ||\beta_k||) \).

The relationship between \( \tau \) and \( \beta \) displayed in (4.3) indicates that the above condition might not hold if \( X_i \perp \perp X_j \) for some \( i \neq j \). Condition 4.3.1 calls for
a stronger form of blockwise non-orthogonality that requires columns of the block matrices $X_i$ not to be orthogonal to particular (sometimes all) columns of other blocks $X_j, j \neq i$. This assumption is automatically satisfied by most general non-orthogonal designs (without any kind of design structure) observed in practical problems, but does not hold for all designs. For example, when the block matrices are all orthogonal to each other, i.e., $X_i \perp X_j \forall i \neq j$, Condition 4.3.1 does not hold. But in such a case, the BOB $g$ prior reduces to the simple block $g$ prior and blockwise orthogonalization of $X$ becomes unnecessary. We make the above assumption to explore the behavior of mixtures of BOB $g$ priors in the most general designs without any sort of orthogonality pattern.

The following lemma furnishes a useful result which is applied in the proofs of some of the theoretical results from this chapter.

**Lemma 4.3.1.** If each problem from the sequence $\{\Psi_N\}$ defined in (4.5) is orthogonalized according to the order mentioned in (4.2), then under Condition 4.3.1, $\sum_{j=1}^{s} R_j^2 \rightarrow 1$ and $R_i^2 \rightarrow 0 \forall i = s + 1, \ldots, k$ as $N \rightarrow \infty$.

**Proof.** For any $i \in \{1, 2, \ldots, k\}$, $R_i^2 = \frac{y^T P_i y}{y^T y} = \frac{(Q_i \hat{\tau}_{i,LS})^T (Q_i \hat{\tau}_{i,LS})}{y^T y}$. When $||\beta_s|| \rightarrow \infty$, (4.3) and Condition 4.3.1 implies that $||\tau_j|| \rightarrow \infty \forall j = 1, 2, \ldots, s$ and for $j = s + 1, \ldots, k$, $\tau_j$ (and hence $||\tau_j||$) is unaffected in the sequence of problems. On the other hand, $||y|| \rightarrow \infty$ in the sequence.

Since the design matrix $Q$ is block orthogonal, this means that $||\hat{\tau}_{j,LS}|| \rightarrow \infty, j = 1, \ldots, s$ and all other $\hat{\tau}_{j,LS}$ are left unaltered as $N \rightarrow \infty$. But $\sigma^2 = \frac{||y||^2}{n-p-1}(1 - R^2)$ remains fixed in the sequence of problems implying that $R^2 \rightarrow 1$. The result follows immediately from the relation $R^2 = \sum_{i=1}^{k} R_i^2 = \sum_{i=1}^{k} \frac{(Q_i \hat{\tau}_{i,LS})^T (Q_i \hat{\tau}_{i,LS})}{y^T y}$, since $y$ grows unbounded and only the first $s$ components of $R^2$ grow in size simultaneously. \qed
The next theorem, a modified version of Theorem 3.2.2 from Chapter 3, displays how the ordinary hyper-g prior suffers from the CLP.

**Theorem 4.3.1.** Consider two models $M_1$ and $M_2$

\[
M_1 : \quad y = \alpha + X_1 \beta_1 + \ldots + X_l \beta_l + \epsilon \\
M_2 : \quad y = \alpha + X_1 \beta_1 + \ldots + X_l \beta_l + X_{l+1} \beta_{l+1} + \epsilon
\]

where each $\beta_{i_j}$ is a vector of length $p_{i_j}$ and $i_j$ is a unique index from the set $\{1, 2, \ldots, k\}$ such that the index $i_{l'} = s$ appears in both models. Under the hyper-g prior, when $||\beta_s|| \to \infty$ (i.e., $N \to \infty$) in the sequence $\{\Psi_N\}$ defined in (4.5) and $n \geq a + \sum_{j=1}^{l} p_{i_j} - 1$, the Bayes factor $BF(M_2 : M_1)$ goes to zero, irrespective of the data.

**Proof.** This is essentially the same result as in Theorem 3.2.2 for a more general form of the design matrix $X$. Since the hyper-g prior is invariant to affine transformations, this general design can be readily transformed to the setting appearing in Theorem 3.2.2 and the result follows. □

**Remark.** The robust prior of Bayarri et al. (2012) also falls prey to the CLP under any arbitrary regression design (which may or may not be orthogonal) as established in Theorem 3.2.4.

We will show that the BOB hyper-g prior defined through the transformed design also avoids the CLP and ELS when the block orthogonalization of $X$ is done in the “correct” order. Recall that the block Gram–Schmidt process is carried out in a specific order with the blocks successively orthogonalized according to importance. For the sequence of regression problems $\{\Psi_N\}$ in (4.5), the most important block is obviously the group of predictors in block $X_s$, since $X_s \beta_s$ accounts for almost all of
the variation in the response. So the block orthogonalization of the design matrix $X$ is carried out in the same manner as in (4.2), but with $X_s$ recognized as the most significant block, i.e., $Q_1 = X_s$. Theorem 4.3.2 confirms the potential of the BOB hyper-$g$ prior to handle the CLP when the most important block is correctly identified.

The proofs of many of the results from this section will depend on the following two lemmas. We shall use the notation $C$ or $C_i, i \in \mathbb{N}$ to denote generic finite constants which are fixed throughout the sequences unless mentioned otherwise.

**Lemma 4.3.2.** Let $\eta_1 + \eta_2 + \ldots + \eta_k = M$ such that $\lim_{N \to \infty} \eta_i > 0$ for each $i \in \{1, 2, \ldots, k \geq 2\}$. Consider the ratio of integrals:

$$I = \frac{\int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{a + p_i} \prod_{i=1}^{k} (1 - t_1 \eta_1 - t_2 \eta_2 - \ldots - t_k \eta_k)^{-m} dt}{\int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{a + p_j} \prod_{i=1}^{k} (1 - t_j L)^{-m} dt}$$

where each $p_i \in \mathbb{Z}^+$, $m > \frac{a + p_j}{2} + k - 2$, $a \geq 3$ and $j \in \{1, 2, \ldots, k\}$. Assume that as $N \to \infty$, $M \to 1$, $L \to 1$ and $\frac{1 - k}{1 - M} \to D$, a finite non-zero constant. If $a, m$ and all $p_i$’s are fixed for all $N$, then $\lim_{N \to \infty} I = 0$.

The proof for the lemma is in Appendix B.1

**Lemma 4.3.3.** Consider the sequence of problems $\{\Psi_N\}$ from (4.5) with the block Gram–Schmidt orthogonalization of $X$ carried out in a prespecified order. Suppose that $X_s$ is identified as the $l$th most important block and orthogonalized with respect to the first $l - 1$ blocks, where $l \in \{1, 2, \ldots, k \geq 2\}$. If a BOB hyper-$g$ prior is imposed on the model as in (4.4), then under the assumption of Condition 4.3.1

(i) $\lim_{N \to \infty} R_i^2 = 0 \ \forall \ i = l + 1, l + 2, \ldots, k$.

(ii) $\lim_{N \to \infty} R_i^2 > 0 \ \forall \ i = 1, 2, \ldots, l$. 

99
(iii) \( \lim_{N \to \infty} \sum_{i=1}^{l} R_i^2 = 1. \)

(iv) When \( n > 2k + 1 \) and \( a \geq 3 \),
\[
\lim_{N \to \infty} \int \prod_{i=1}^{k} (1 - t_i)^{-\frac{a+p_i}{2}} \left(1 - \sum_{i=1}^{k} t_i R_i^2\right)^{-\frac{n-1}{2}} \, dt
= C \times \lim_{N \to \infty} \int \prod_{i=1}^{k} (1 - t_i)^{-\frac{a+p_i}{2}} \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} \, dt, \text{ for some finite } C \geq 1.
\]

The proof of this result can be found in Appendix B.2.

With the help of these two valuable lemmas we can now establish a theoretical result on the ability of the BOB hyper-g prior to avoid the CLP.

**Theorem 4.3.2.** Consider two models \( M_1 \) and \( M_2 \)

\[
M_1 : \quad y = \alpha 1 + X_{i_1} \beta_{i_1} + \ldots + X_{i_l} \beta_{i_l} + \epsilon
\]

\[
M_2 : \quad y = \alpha 1 + X_{i_1} \beta_{i_1} + \ldots + X_{i_{l+1}} \beta_{i_{l+1}} + \epsilon
\]

where each \( \beta_{ij} \) is a vector of length \( p_{ij} \) and \( i_j \) is a unique index from the set \( \{1, 2, \ldots, k\} \) such that the index \( i_{iv} = s \) appears in both models. Assume that a BOB hyper-g prior is imposed on the reparameterized model with \( Q_1 = X_s \) in both models. Then for the sequence \( \{\Psi_N\} \) in (4.5), when \( n > 2l + 3 \) and \( a \geq 3 \), the Bayes factor \( BF(M_2 : M_1) \) is bounded away from zero as \( N \to \infty \).

**Proof.** We can rewrite the two models as

\[
M_1 : \quad y = \alpha 1 + Q_1 \tau_1 + Q_2 \tau_2 + \ldots + Q_l \tau_l + \epsilon
\]

\[
M_2 : \quad y = \alpha 1 + Q_1^* \tau_1^* + Q_2^* \tau_2^* + \ldots + Q_{l+1}^* \tau_{l+1}^* + \epsilon
\]

where the first block in the block Gram–Schmidt process is \( X_s \) for both \( M_1 \) and \( M_2 \) \( (Q_1 = Q_1^* = X_s) \). The successive blocks might be different in \( M_1 \) and \( M_2 \) depending on the placement of the block \( X_{i_{l+1}} \) in the order of orthogonalization.
It is not difficult to show that

\[
BF(M_2 : M_1) = \frac{a - 2\int_{(0,1)}^{1} \left[ \prod_{j=1}^{l+1} (1 - t_j)^{\alpha} \frac{\pi_j}{2} - 2 \right] \left( 1 - \sum_{j=1}^{l+1} t_j (R_j^{(2)})^2 \right) - \frac{n-1}{2} dt}{2 \int_{(0,1)}^{1} \left[ \prod_{j=1}^{l} (1 - t_j)^{\alpha} \frac{\pi_j}{2} - 2 \right] \left( 1 - \sum_{j=1}^{l} t_j (R_j^{(1)})^2 \right) - \frac{n-1}{2} dt}
\]

where \((R_j^{(c)})^2\) is the component of \(R^2\) due to the \(j^{th}\) orthogonal block \(Q_j\) or \(Q_j^*\) under model \(M_c, c = 1, 2\).

Due to Lemma 4.3.1, \([R_1^{(2)}]^2 \to 1\) and \([R_1^{(1)}]^2 \to 1\) while all other \([R_j^{(c)}]^2 \to 0\) as \(N \to \infty\). Hence,

\[
\lim_{N \to \infty} BF(M_2 : M_1) = C_1 \frac{\int_{(0,1)}^{1} \left[ \prod_{j=1}^{l+1} (1 - t_j)^{\alpha} \frac{\pi_j}{2} - 2 \right] \left( 1 - t_1[R_1^{(2)}]^2 \right) - \frac{n-1}{2} dt}{\int_{(0,1)}^{1} \left[ \prod_{j=1}^{l} (1 - t_j)^{\alpha} \frac{\pi_j}{2} - 2 \right] \left( 1 - t_1[R_1^{(1)}]^2 \right) - \frac{n-1}{2} dt}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1[R_1^{(2)}]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1[R_1^{(1)}]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 \lim_{N \to \infty} \frac{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(2)} \right]^2 \right) - \frac{n-1}{2} dt_1}{\int_{0}^{1} \left[ (1 - t_1)^{\alpha} \frac{\pi_1}{2} - 2 \right] \left( 1 - t_1 \left[ \lim_{N \to \infty} R_1^{(1)} \right]^2 \right) - \frac{n-1}{2} dt_1}
\]

\[= C_2 > 0.
\]
The equality (4.6) follows from Lemma 4.3.3 since \( n > 2l + 3 \). (4.8) is true due to the relation
\[
\left[ \frac{R^{(2)}_{1}}{R^{(1)}_{1}} \right]^{2} = \frac{y^{T}P_{Q_{1}}y}{y^{T}P_{Q_{1}}y} = \frac{y^{T}P_{X_{1}}y}{y^{T}P_{X_{1}}y} = 1 \text{ for all } N. \] (4.7) and (4.9) can be justified by the Monotone Convergence Theorem since the function \((1 - t_{j})^{\frac{a + p_{j} - 2}{2}}(1 - \frac{n - 1}{n}t_{i})^{-\frac{a}{2}}\) is increasing in \( r^{2} \).

The theorem shows that the BOB hyper-\( g \) prior avoids the CLP if the block Gram–Schmidt orthogonalization is carried out in the right order. An identical result can be shown to be true for the BOB hyper-\( g/n \) prior as well.

**Theorem 4.3.3.** Consider two models \( M_{1} \) and \( M_{2} \)
\[
M_{1} : \quad y = \alpha 1 + X_{i_{1}}\beta_{i_{1}} + \ldots + X_{i_{l}}\beta_{i_{l}} + \epsilon
\]
\[
M_{2} : \quad y = \alpha 1 + X_{i_{1}}\beta_{i_{1}} + \ldots + X_{i_{l}}\beta_{i_{l}} + X_{i_{l+1}}\beta_{i_{l+1}} + \epsilon
\]
where each \( \beta_{i_{j}} \) is a vector of length \( p_{i_{j}} \) and \( i_{j} \) is a unique index from the set \{1, 2, \ldots, k\} such that the index \( i_{\nu} = s \) appears in both models. Assume that a BOB hyper-\( g/n \) prior is imposed on the reparameterized model with \( Q_{1} = X_{s} \) in both models. Then for the sequence \( \{\Psi_{N}\} \) in (4.5), when \( n > 2l + 3 \) and \( a \geq 3 \), the Bayes factor \( BF(M_{2} : M_{1}) \) is bounded away from zero as \( N \to \infty \).

**Proof.** The proof is very similar to the one for Theorem 4.3.2, we just need one new result to bound the Bayes factor for the block hyper-\( g/n \) prior

\[
BF(M_{2} : M_{1}) = \frac{a - 2 \int_{(0,1)^{l+1}} \left[ \prod_{j=1}^{l+1} (1 - t_{j})^{\frac{a + p_{j}}{2} - 2} \left( 1 - \frac{n - 1}{n}t_{i} \right)^{-\frac{a}{2}} \right] \left( 1 - \sum_{j=1}^{l+1} t_{j} (R^{(2)}_{j})^{2} - \frac{n - 1}{2} \right) dt}{\int_{(0,1)^{l}} \left[ \prod_{j=1}^{l} (1 - t_{j})^{\frac{a + p_{j}}{2} - 2} \left( 1 - \frac{n - 1}{n}t_{i} \right)^{-\frac{a}{2}} \right] \left( 1 - \sum_{j=1}^{l} t_{j} (R^{(1)}_{j})^{2} - \frac{n - 1}{2} \right) dt}.
\]

If we are able to prove a modification of Lemma 4.3.3 (iv) for the BOB hyper-\( g/n \) prior, then the proof of the theorem can be completed with the same steps used in...
proving Theorem 4.3.2. We will show that under the same setup as in Lemma 4.3.3, the following result holds:

When \( n > 2k + 1 \) and \( a \geq 3 \),

\[
\lim_{N \to \infty} \int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a}{2}} \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt
\]

\[
= C \times \lim_{N \to \infty} \int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a}{2}} \left( 1 - l \sum_{i=1}^{l} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt
\]

for some finite \( C \geq 1 \).

Using the same logic as before, we know that the constant \( C \geq 1 \), but we still have to prove that \( C < \infty \). Since \( 1 \leq \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a}{2}} \leq n^{\frac{a}{2}} \), we have

\[
\frac{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a}{2}} \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \frac{n - 1}{n} t_i \right)^{-\frac{a}{2}} \left( 1 - \sum_{i=1}^{l} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}
\]

\[
\leq \frac{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} n^\frac{a}{2} \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \sum_{i=1}^{l} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}
\]

\[
= n^\frac{ak}{2} \frac{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \sum_{i=1}^{k} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}{\int \prod_{i=1}^{k} \left( 1 - t_i \right)^{\frac{a + p_i}{2} - 2} \left( 1 - \sum_{i=1}^{l} t_i R_i^2 \right)^{-\frac{n-1}{2}} dt}
\]

It has been shown in Appendix B.2 that the ratio of integrals in the last display is bounded from above as \( N \to \infty \). This is a sufficient condition to show that (4.10) is true. \( \square \)

Both the BOB hyper-\( g \) and the BOB hyper-\( g/n \) priors are unaffected by the ELS behavior when the most important block \( X_s \) is correctly identified as the first block in
the transformed model, i.e., when $Q_1 = X_s$. We demonstrate how the BOB hyper-g prior avoids ELS in the following theorem.

**Theorem 4.3.4.** For the regression model described by (4.4), (3.9) and (4.2) with $Q_1 = X_s$,

$$E(\tau \mid y) = \begin{pmatrix} E \left( \frac{q_1}{1+g_1} \mid y \right) \hat{\tau}_{1,LS} \\ \vdots \\ E \left( \frac{q_k}{1+g_k} \mid y \right) \hat{\tau}_{k,LS} \end{pmatrix}.$$  \[\text{Further assume that } n \geq a + p_1 - 1. \text{ As } N \to \infty \text{ in the sequence } \{\Psi_N\} \text{ defined in (4.5), } E \left( \frac{q_1}{1+g_1} \mid y \right) \to 1 \text{ and for } i \neq 1, E \left( \frac{q_i}{1+g_i} \mid y \right) \to \Delta_i \text{ with } \frac{2}{a+p_i} \leq \Delta_i < 1.\]

**Proof.** The result follows directly from the proof of Theorem 3.4.1 in Appendix A.7 once we establish that $\lim_{N \to \infty} R^2_1 = 1$ and $\lim_{N \to \infty} R^2_i = 0, \forall i \neq 1$, in the block orthogonal design $Q$. This is obviously true by Lemma 4.3.1 because $Q_1 = X_s$. \[\square\]

Since $\tau = U\beta$ for a block upper triangular matrix $U$, clearly $\hat{\tau}_{LS} = U\hat{\beta}_{LS}$ due to the invariance of least squares estimates to linear transformations. Also the relation $E(\beta \mid y) = U^{-1}E(\tau \mid y)$ implies that

$$\frac{\|\hat{\beta}_B - \hat{\beta}_{LS}\|}{\|\hat{\beta}_{LS}\|} = \frac{\|U^{-1}\tau_B - U^{-1}\hat{\tau}_{LS}\|}{\|U^{-1}\hat{\tau}_{LS}\|} = \frac{\|U^{-1}(\tau_B - \hat{\tau}_{LS})\|}{\|U^{-1}\hat{\tau}_{LS}\|}$$

where the notation $\hat{\theta}_B$ stands for the BOB hyper-g (Bayes) estimate and $\hat{\theta}_{LS}$ stands for the least squares estimate of the parameter $\theta$. This means that

$$\frac{\lambda_p \|\hat{\tau}_B - \hat{\tau}_{LS}\|}{\lambda_1 \|\hat{\tau}_{LS}\|} \leq \frac{\|\hat{\beta}_B - \hat{\beta}_{LS}\|}{\|\hat{\beta}_{LS}\|} \leq \frac{\lambda_1 \|\hat{\tau}_B - \hat{\tau}_{LS}\|}{\lambda_p \|\hat{\tau}_{LS}\|}$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p (> 0)$ are the eigenvalues of $(U^{-1})^TU^{-1}$.

According to Theorem 4.3.4, $\frac{\|\hat{\beta}_B - \hat{\beta}_{LS}\|}{\|\hat{\beta}_{LS}\|}$ is bounded away from zero as $N \to \infty$, so

$$\lim_{N \to \infty} \frac{\|\hat{\beta}_B - \hat{\beta}_{LS}\|}{\|\hat{\beta}_{LS}\|}$$

must also be bounded away from zero avoiding the ELS behavior.
Remark. The BOB hyper-\(g/n\) prior also does not suffer from ELS when the same sufficient conditions hold. This is a result that can be verified easily using a modification of Theorem 3.6.2 along the lines of Theorem 4.3.4.

4.3.2 Significance of the Order of Orthogonalization

We assumed in the last subsection that the group of predictors corresponding to the exploding coefficient vector \(\beta_s\) from the sequence of problems \(\{\Psi_N\}\) in (4.5) has been predetermined to be the most important block. Under such an assumption, many nice properties of the block hyper-\(g\) and block hyper-\(g/n\) priors carry over to the BOB hyper-\(g\) and BOB hyper-\(g/n\) priors. It is an intriguing fact that the same conclusions do not persist if the order of orthogonalization is switched. Consider a simple situation with a pair of nested models \(M_1\) and \(M_2\) with \(M_1 \subset M_2\) such that \(X_s\) is identified as the most important block among all the groups of predictors in \(M_1\), but the extra block of predictors in \(M_2\) is incorrectly assigned more importance than \(X_s\). So \(X_s\) is the first block in the block Gram–Schmidt orthogonalization process for model \(M_1\) (i.e., \(X_s = Q_1\)), but it is the second block in case of model \(M_2\). (4.11) displays the exact order in which the orthogonalization is implemented in the two models.

\[
\begin{align*}
\text{Model } M_1 & \\
Q_1 &= X_s \\
Q_2 &= (I - P_{Q_1})X_{i_1} \\
& \vdots \\
Q_l &= (I - P_{Q_1} - \ldots - P_{Q_{l-1}})X_{i_l} \\
\text{Model } M_2 & \\
Q_1^* &= X_{i_{l+1}} \\
Q_2^* &= (I - P_{Q_1^*})X_s \\
& \vdots \\
Q_l^* &= (I - P_{Q_1^*} - \ldots - P_{Q_{l-1}^*})X_{i_{l-1}} \\
Q_{l+1}^* &= (I - P_{Q_1^*} - \ldots - P_{Q_{l+1}^*})X_{i_l} \quad (4.11)
\end{align*}
\]
Theorem 4.3.5. Consider two models $M_1$ and $M_2$

\[ M_1 : \quad y = \alpha_1 + X_{i_1} \beta_{i_1} + \ldots + X_{i_l} \beta_{i_l} + \epsilon \]
\[ M_2 : \quad y = \alpha_1 + X_{i_1} \beta_{i_1} + \ldots + X_{i_l} \beta_{i_l} + X_{i_{l+1}} \beta_{i_{l+1}} + \epsilon \]

where each $\beta_{i_j}$ is a vector of length $p_{i_j}$ and $i_j$ is a unique index from the set \{1, 2, \ldots, k\} such that the index $i_j$ appears in both models. Assume that a BOB hyper-$g$ prior is imposed on the reparameterized versions of models $M_1$ and $M_2$ described in (4.11). Then for the sequence $\{\Psi_N\}$ defined in (4.5), when $n > (a + p_s + 1) \lor (2l + 3)$, $a \geq 3$ and Condition 4.3.1 holds, the Bayes factor $BF(M_2 : M_1)$ goes to zero as $N \to \infty$.

Proof. As before, we can rewrite the two models as

\[ M_1 : \quad y = \alpha_1 + Q_1 \gamma_1 + Q_2 \gamma_2 + \ldots + Q_l \gamma_l + \epsilon \]
\[ M_2 : \quad y = \alpha_1 + Q^*_1 \gamma^*_1 + Q^*_2 \gamma^*_2 + \ldots + Q^*_{l+1} \gamma^*_{l+1} + \epsilon \]

where the orthogonal blocks $Q_i$ and $Q^*_i$ are defined by (4.11).

Define $[R_j^{(e)}]^2$ in the same way as in Theorem 4.3.2 and notice that $(R_1^{(1)})^2 \to 1$ and $(R_1^{(2)})^2 + (R_2^{(2)})^2 \to 1$ as $N \to \infty$ due to Lemma 4.3.1 and Condition 4.3.1. Hence,

\[
\lim_{N \to \infty} BF(M_2 : M_1) = \lim_{N \to \infty} \frac{a - 2 \int_{(0,1)}^{l+1} \left[ \prod_{j=1}^{l+1} (1 - t_j) \frac{a + p_j}{2} \right] \left( 1 - \sum_{j=1}^{l+1} t_j [R_j^{(2)}]^2 \right)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^l} \left[ \prod_{j=1}^{l} (1 - t_j) \frac{a + p_j}{2} \right] \left( 1 - \sum_{j=1}^{l} t_j [R_j^{(1)}]^2 \right)^{-\frac{n-1}{2}} dt}
\]

\[
= C_1 \lim_{N \to \infty} \left. \frac{\int_{(0,1)}^{l+1} \left[ \prod_{j=1}^{l+1} (1 - t_j) \frac{a + p_j}{2} \right] \left( 1 - t_1 [R_1^{(2)}]^2 - t_2 [R_2^{(2)}]^2 \right)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^l} \left[ \prod_{j=1}^{l} (1 - t_j) \frac{a + p_j}{2} \right] \left( 1 - t_1 [R_1^{(1)}]^2 \right)^{-\frac{n-1}{2}} dt} \right|_{(0,1)}^{l+1}
\]

106
\[ C_2 \lim_{N \to \infty} \int_{(0,1)^2} \left[ \prod_{j=1}^2 (1 - t_j)^{-\frac{n-1}{2}} \right] (1 - t_1[R_1^{(1)}]^2 - t_2[R_2^{(2)}]^2)^{-\frac{n-1}{2}} dt_1 dt_2. \]

The last two steps follow from Lemma 4.3.3 along with the additional assumption of Condition 4.3.1. To apply Lemma 4.3.3 we need \( n > 2l + 3. \)

If we define \( M = (R_1^{(2)})^2 + (R_2^{(2)})^2, \) \( L = (R_1^{(1)})^2, \) then as \( N \to \infty \) we have \( M \to 1, \)
\( L \to 1 \) and \( \frac{1-L}{1-M} = \frac{1-[R_1^{(1)}]^2}{1-[R_1^{(2)}]^2 + [R_2^{(2)}]^2} = \frac{O(1)/O(\|\beta\|_2^2)}{O(1)/O(\|\beta^*_1\|_2^2)} = O(1) = D, \) a finite non-zero constant. Furthermore, when \( n > a + p_s + 1, \) all the prerequisites for applying Lemma 4.3.2 are satisfied and we conclude that \( \lim_{N \to \infty} BF(M_2 : M_1) = 0. \)

The Bayes factor \( BF(M_2 : M_1) \) equals zero in the limit which indicates that the peculiar behavior associated with the CLP does not disappear with all conceivable versions of the BOB hyper-g prior.

A generalization of Theorem 4.3.5 shows that if any number of extra blocks of predictors from the larger model are considered more important than the group of predictors in block \( X_s \) and \( X_s \) is orthogonalized with respect to the new blocks, then \( CLP \) persists in the BOB hyper-g model when the sample is sufficiently large. Before stating this result, first consider the following pair of nested models \( M_1 \) and \( M_2, \) where \( X_s \) is considered to be the \( d^{th} \) most important block in the larger model \( M_2. \)

Assume that \( r, l \geq 1, d \geq 2 \) and \( j_1, j_2, \ldots, j_{d-1} \in \{ l+1, l+2, \ldots, l+r \}. \)

\[
\begin{align*}
\text{Model } M_1 & \quad \text{Model } M_2 \\
Q_1 = X_s & \quad Q'_1 = X_{ij_1} \\
Q_2 = (I - P_{Q_1})X_i & \quad Q'_2 = (I - P_{Q'_1})X_{ij_2} \\
\vdots & \quad \vdots
\end{align*}
\]

(4.12)
\[ Q_d^* = (I - P_{Q_i} - \ldots - P_{Q_{i-1}})X_s \]
\[ Q_t = (I - P_{Q_1} - \ldots - P_{Q_{i-1}})X_t \]
\[ Q_{t+r}^* = (I - P_{Q_i} - \ldots - P_{Q_{i+r-1}})X_{t+r} \]

**Corollary 4.3.1.** Consider two models \( M_1 \) and \( M_2 \)

\[ M_1 : \ y = \alpha 1 + X_{i_1}\beta_{i_1} + \ldots + X_{i_s}\beta_{i_s} + \epsilon \]
\[ M_2 : \ y = \alpha 1 + X_{i_1}\beta_{i_1} + \ldots + X_{i_s}\beta_{i_s} + \ldots + X_{i_{t+r}}\beta_{i_{t+r}} + \epsilon \]

where each \( \beta_{i_j} \) is a vector of length \( p_{i_j} \) and \( i_j \) is a unique index from the set \( \{1, 2, \ldots, k\} \) such that the index \( i_{i'} = s \) appears in both models. Assume that a BOB hyper-g prior is imposed on the reparameterized models \( M_1 \) and \( M_2 \) described in (4.12). Then for the sequence \( \{\Psi_N\} \) defined in (4.5), when \( n > (a + p_s + 2d - 3) \lor (2l + 2r + 1), a \geq 3 \) and Condition [4.3.1] holds, the Bayes factor \( BF(M_2 : M_1) \) goes to zero as \( N \to \infty \).

It is straightforward to come up with a version of Corollary 4.3.1 for the BOB hyper-g/n prior, which reveals that it suffers from the CLP under identical conditions.

**Corollary 4.3.2.** Consider two models \( M_1 \) and \( M_2 \)

\[ M_1 : \ y = \alpha 1 + X_{i_1}\beta_{i_1} + \ldots + X_{i_s}\beta_{i_s} + \epsilon \]
\[ M_2 : \ y = \alpha 1 + X_{i_1}\beta_{i_1} + \ldots + X_{i_s}\beta_{i_s} + \ldots + X_{i_{t+r}}\beta_{i_{t+r}} + \epsilon \]

where each \( \beta_{i_j} \) is a vector of length \( p_{i_j} \) and \( i_j \) is a unique index from the set \( \{1, 2, \ldots, k\} \) such that the index \( i_{i'} = s \) appears in both models. Assume that a BOB hyper-g/n prior is imposed on the reparameterized models \( M_1 \) and \( M_2 \) described in (4.12). Then for the sequence \( \{\Psi_N\} \) defined in (4.5), when \( n > (a + p_s + 2d - 3) \lor (2l + 2r + 1), a \geq 3 \) and Condition [4.3.1] holds, the Bayes factor \( BF(M_2 : M_1) \) goes to zero as \( N \to \infty \).
The proofs for Corollaries 4.3.1 and 4.3.2 rely on a minor modification of the steps in the proof of Theorem 4.3.5.

**Theorem 4.3.6.** For the regression model described by (4.4), (3.9) and (4.2),
\[
E(\tau | y) = \begin{pmatrix}
E \left( \frac{q_1}{1+q_1} | y \right) \hat{\tau}_{1,LS} \\
\vdots \\
E \left( \frac{q_k}{1+q_k} | y \right) \hat{\tau}_{k,LS}
\end{pmatrix}.
\]

Suppose that $X_s$ is considered as the $l$th ($l \geq 2$) most important block of predictors in the sequence $\{\Psi_N\}$ defined in (4.5) so that $Q_l = X_s$. When $n > 1 + \sum_{j=1}^{l} p_j + l(a - 2)$, $a \geq 3$ and Condition 4.3.1 holds, $E \left( \frac{q_i}{1+q_i} | y \right) \to 1$ for $i = 1, 2, \ldots, l$ and for all other $i$, $E \left( \frac{q_i}{1+q_i} | y \right) \to \Delta_i$ with $\frac{2}{a+p_i} \leq \Delta_i < 1$, as $N \to \infty$.

The proof of this theorem is in Appendix B.4. The conclusion about shrinkage factors for the transformed regression coefficients $\tau$ does not directly relate to the ELS behavior for estimates of $\beta$. But the following corollary shows that a modified form of ELS, which we call the Componentwise Essentially Least Squares estimation or CELS, affects inference in the sequence $\{\Psi_N\}$. In this situation, we have the Bayes estimates (under squared error loss) coinciding with least squares estimates for some components of $\beta$, while the remaining components experience shrinkage in their coefficient estimates depending on the quality of the predictor variables. Ideally, we would want only the Bayes estimate of $\beta_s$ to be close to its least squares counterpart, but the next corollary demonstrates that CELS affects coefficient estimates in all blocks orthogonalized prior to $X_s$.

**Corollary 4.3.3.** (CELS) Under the same setup as in Theorem 4.3.6 for the sequence $\{\Psi_N\}$, we have $\lim_{N \to \infty} \frac{\|\widehat{\beta}_{i,B} - \widehat{\beta}_{i,LS}\|}{\|\widehat{\beta}_{i,LS}\|} = 0$, for $i = 1, 2, \ldots, l \geq 2$ and the limit is non-zero for all other $i$. 

109
Proof. The relation $\beta = U^{-1}\tau$ implies that

$$\frac{||\hat{\beta}_{i,B} - \hat{\beta}_{i,LS}||}{||\hat{\beta}_{i,LS}||} = \frac{||\sum_{j=1}^{k} U_{ij}(\tau_{j,B} - \tau_{j,LS})||}{||\sum_{j=1}^{k} U_{ij}\tau_{j,LS}||} = \frac{||\sum_{j=1}^{k} E(\frac{1}{1+g_j} | y) U_{ij}\tau_{j,LS}||}{||\sum_{j=1}^{k} U_{ij}\tau_{j,LS}||} \leq \frac{\sum_{j=1}^{k} E\left(\frac{1}{1+g_j} | y\right) O(||\beta_s||^2)}{O(||\beta_s||^2)} + \lim_{N \to \infty} \sum_{j=k+1}^{l} E\left(\frac{1}{1+g_j} | y\right) \frac{O(1)}{O(||\beta_s||^2)}.$$

where $U_{ij}$ is the $(i, j)^{th}$ block matrix in $U^{-1}$. $U^{-1}$ is block upper triangular since $U$ is block upper triangular and it can be shown that $U_{ii} = I \forall i$.

Since $||\beta_s|| \to \infty$ and $X_s$ is the $l^{th}$ orthogonalized block, $O(||\hat{\tau}_{j}||^2) = O(||\beta_s||^2)$ for all $i \in \{1, 2, ..., l\}$. Hence, when $i \in \{1, 2, ..., l\}$,

$$\lim_{N \to \infty} \frac{||\hat{\beta}_{i,B} - \hat{\beta}_{i,LS}||}{||\hat{\beta}_{i,LS}||} \leq \lim_{N \to \infty} \frac{\sum_{j=1}^{k} E\left(\frac{1}{1+g_j} | y\right) O(||\beta_s||^2)}{O(||\beta_s||^2)} + \lim_{N \to \infty} \sum_{j=k+1}^{l} E\left(\frac{1}{1+g_j} | y\right) \frac{O(1)}{O(||\beta_s||^2)}.$$

The first part of the last expression equals zero because $E\left(\frac{1}{1+g_j} | y\right) = 1 - E\left(\frac{g_j}{1+g_j} | y\right)$ tends to zero for each $j = 1, 2, \ldots, l$ when $n$ is sufficiently large (by Theorem 4.3.6). The second part also equals zero as $\lim_{N \to \infty} ||\beta_s|| = \infty$. This proves that $\lim_{N \to \infty} \frac{||\hat{\beta}_{i,B} - \hat{\beta}_{i,LS}||}{||\hat{\beta}_{i,LS}||} = 0$, for $i = 1, 2, \ldots, l$.

When $i = l + 1, l + 2, \ldots, k$,

$$\frac{||\hat{\beta}_{i,B} - \hat{\beta}_{i,LS}||}{||\hat{\beta}_{i,LS}||} = \frac{||\sum_{j=i}^{k} E\left(\frac{1}{1+g_j} | y\right) U_{ij}\tau_{j,LS}||}{||\sum_{j=i}^{k} U_{ij}\tau_{j,LS}||},$$

which does not involve $\beta_s$ in any way and the ratio of the finite, non-zero numerator and denominator is always non-zero, even in the limit as $N \to \infty$. \[\square\]
The same result regarding CELS in the BOB hyper-$g$ prior is also true in the BOB hyper-$g/n$ prior given the aforementioned set of sufficient conditions.

4.4 Consistency of the BOB Hyper-$g$ and BOB Hyper-$g/n$ Priors

In this section, we study the properties of the BOB hyper-$g$ and BOB hyper-$g/n$ priors with respect to the three important aspects of asymptotic consistency discussed in Liang et al. (2008): information consistency, model selection consistency and prediction consistency. The analysis of these consistency criteria is almost identical to the asymptotic analyses in Sections 3.5 and 3.6.2 with only a slight difference in the set of all possible models due to distinctive model spaces.

Consider a regression model with predetermined blocks of predictor variables and assume that the total number of blocks in the full model with all $p$ predictors is $K (\in \mathbb{N})$. The model space under consideration is of a smaller size than the usual model space containing all possible combinations of all $p$ covariates. The set of models only includes all combinations of the $K$ blocks from the full model and, thus, has $2^K$ elements. Let $z \in \Gamma_B = \{0, 1\}^K$ be a vector that denotes the presence (if $z_i = 1$) or absence (if $z_i = 0$) of the $i$th block of predictors in an arbitrary model $M_z$. Let $k_z$ denote the number of blocks in the model $M_z$.

As in Section 3.5, in order to show model selection consistency and prediction consistency in Sections 4.4.2 and 4.4.3, we have to first define a “true” model which generates the data. Let $M_T : y = \alpha_T 1 + \sum_{i \in B_T} X_{i,T} \beta_{i,T} + \epsilon$ denote the true model, where $B_T \subseteq \{1, 2, \ldots, K\}$ is the set of block indices in the true model and $|B_T| = k_T$. The “true” reparameterized model, generated by the block Gram–Schmidt process, can
be represented as $\mathcal{M}_T : y = \alpha_T 1 + \sum_{i \in B_T} Q_{i,T} \tau_{i,T} + \epsilon$, and (4.2) and (4.3) connect the two alternative forms of the same model. The standard assumption for consistency of Bayesian methods in linear models (Fernandez et al., 2001; Liang et al., 2008; Maruyama and George, 2011; Bayarri et al., 2012) is considered to hold in the form of Condition 3.5.1.

### 4.4.1 Information Consistency

The notion of information consistency arises from preventing the Information Paradox described in Section 3.2. A Bayesian normal linear regression model is information consistent, if for a fixed sample size $n$ and a fixed design, an appropriate limit on the likelihood causing $R^2_z \to 1$ for an arbitrary model $\mathcal{M}_z$ implies that $BF(\mathcal{M}_z : \mathcal{M}_0) \to \infty$. Theorems 4.4.1 and 4.4.2 verify that the BOB hyper-$g$ and hyper-$g/n$ priors are information consistent.

**Theorem 4.4.1.** Consider a regression model with a BOB hyper-$g$ prior as described in (4.4), (4.2) and (3.9). The BOB hyper-$g$ is “information consistent” when $n > k_z(a - 2) + p_z + 1$, where $k_z$ is the total number of blocks, $p_z = \sum_{j=1}^{k_z} p_{j,z}$ and $p_{j,z}$ is the size of block $X_{j,z}$ in model $\mathcal{M}_z$.

**Theorem 4.4.2.** Consider a regression model with a BOB hyper-$g/n$ prior as described in (4.4), (4.2) and (3.14). The BOB hyper-$g/n$ is “information consistent” when $n > k_z(a - 2) + p_z + 1$, where $k_z$ is the total number of blocks, $p_z = \sum_{j=1}^{k_z} p_{j,z}$ and $p_{j,z}$ is the size of block $X_{j,z}$ in model $\mathcal{M}_z$.

The proofs of the above theorems follow immediately from the proofs of Theorems 3.5.1 and 3.6.3 in Appendices A.10 and A.13 once we note that $R^2_z = \frac{y^T P_{X_z} y}{y^T y} = \sum_{i=1}^{k_z} \frac{y^T P_{Q_{i,z}} y}{y^T y} = \sum_{i=1}^{p_z} R^2_{i,z}$.
4.4.2 Model Selection Consistency

Model selection consistency and prediction consistency are derived under the usual asymptotic setting related to increasing sample size and fixed parameters. Condition 3.5.1 needs to be true for both of these consistency properties to hold in the hyper-mixtures of BOB $g$ priors. Because the original and the reparameterized design matrices are related as $X = QU$, Condition 3.5.1 on the design $X$ implies an equivalent condition on the block orthogonal $n \times p$ design matrix $Q$:

$$
\lim_{n \to \infty} \frac{1}{n} Q^T Q = D^*
$$

(4.13)

for some $p \times p$ positive definite matrix $D^*$. This follows directly from $\frac{1}{n} Q^T Q = \frac{1}{n} (U^{-1})^T (X^T X) U^{-1} = (U^{-1})^T \left[ \frac{1}{n} (X^T X) \right] U^{-1}$.

Maruyama and George (2011) show that (4.13) is a precursor to the following condition:

$$
\lim_{n \to \infty} \frac{1}{n} \tau_T Q_T^T (I - P_{Qz}) Q_T \tau_T = V_z > 0.
$$

(4.14)

Both (4.13) and (4.14) are essential to derive the subsequent results on model selection consistency. According to Fernandez et al. (2001), posterior consistency in model selection is attained when

$$
\pi(\mathcal{M}_T \mid y) \xrightarrow{P} 1 \text{ as } n \to \infty, \text{ assuming } \mathcal{M}_T \text{ is the true model},
$$

or equivalently, if

$$
BF(\mathcal{M}_z : \mathcal{M}_T) \xrightarrow{P} 0 \text{ as } n \to \infty, \text{ for any model } \mathcal{M}_z \neq \mathcal{M}_T.
$$

(4.15)

Mixture of BOB $g$ priors behave identical to the mixtures of block $g$ priors in regards to posterior model selection. Theorems 4.4.3 and 4.4.4 show that the BOB
hyper-g prior is not consistent in model selection in some situations, but the BOB hyper-g/n is always consistent. Before presenting the main results from this section, a modification of Lemma A.11.1 from Appendix A.11 is first stated, which will be used in the proofs of both the main results.

**Lemma 4.4.1.** Let $R^2_{i,z}$ and $R^2_{i,T}$ denote the $i$th component of $R^2$ under an arbitrary model $\mathcal{M}_z$ and the true model $\mathcal{M}_T$.

(i) For $i \in B_T$, $R^2_{i,T} \xrightarrow{P} F_{i,T} > 0$.

(ii) For $i \in B_T$, $R^2_{i,z} \xrightarrow{P} F_{i,z} > 0$.

(iii) There exists a set $H_z \subset B_z \setminus B_T \subset \{1, 2, \ldots, k\}$ depending on the order of block orthogonalization such that for $i \in H_z$, $R^2_{i,z} \xrightarrow{P} 0$. For all such $i$, $nR^2_{i,z} \xrightarrow{d} c\chi^2_{p_i,z} = O(1)$ for a constant $c$.

(iv) For any model $\mathcal{M}_z \supset \mathcal{M}_T$, $(\frac{1-R^2_T}{1-R^2_z})^n$ is bounded from above in probability.

The proof of this lemma is similar to the proof of Lemmas B.2 and B.3 from Maruyama and George (2011) and the proof of Lemma A.11.1. The lemma signifies that based on how the blocks in a model $\mathcal{M}_z$ are orthogonalized, the components of $R^2$ in $\mathcal{M}_z$ which belong to blocks not included in the “true” model might have a non-zero limit with increasing sample size. The other parts of the lemma are identical to Lemma A.11.1.

**Theorem 4.4.3.** Consider a regression model with a BOB hyper-g prior as described in (4.4), (4.2) and (3.9) which satisfies 4.13 and 4.14. This model does not have posterior model selection consistency for all choices of the true model.
Proof. Let $R^2_{i,T}$ and $p_{i,T}$ represent the component of $R^2$ and the number of predictors in the $i^{th}$ block $Q_{i,T}$ from the true model $\mathcal{M}_T$ while $R^2_{i,z}$ and $p_{i,z}$ denote the corresponding entities for block $Q_{i,z}$ from an arbitrary model $\mathcal{M}_z (\neq \mathcal{M}_T)$. Further assume that $B_z$ denotes the set of indices of the blocks within $\mathcal{M}_z$, and let $k_z = |B_z|$. We only prove the theorem for the case $a > 3$, as in Appendix A.11, and argue that extension to the case $2 < a \leq 3$ is analogous to the proof in Appendix A.16.

First consider the case when $\mathcal{M}_T \neq \mathcal{M}_0$. Following the notation and the steps of the proof of Theorem 3.5.2 and utilizing Lemmas A.11.1 and A.11.2, we can show that for large $m = n - 1/2$ (same as having large $n$),

$$BF(\mathcal{M}_z : \mathcal{M}_T) \approx O(m^{k_T-q_z}) \exp \left[ -m \log \left( \frac{m(1-R^2_z)}{m-b_z} \right) + m \log \left( \frac{m(1-R^2_T)}{m-b_T} \right) + \sum_{i \in B_z} b_{i,z} \log \left( \frac{b_{i,z}(1-R^2_z)}{R^2_{i,z}(m-b_z)} \right) - \sum_{i \in B_T} b_{i,T} \log \left( \frac{b_{i,T}(1-R^2_T)}{R^2_{i,T}(m-b_T)} \right) \right]$$

$$= O(m^{k_T-q_z}) \exp \left[ m \log \left( \frac{1-R^2_T}{1-R^2_z} \right) + (b_T - b_z) \log m - \sum_{i \in B_z} b_{i,z} \log(R^2_{i,z}) + O(1) \right]$$

due to a Laplace approximation. Here $q_z = k_z - L_z$, and $L_z$ is the number of components $R^2_{i,z}$ going to zero, while $R^2_j = \sum_{i \in B_j} R^2_{i,j}$, $b_{i,j} = \frac{a + p_{i,j}}{2} - 2$ and $b_j = \sum_{i \in B_j} b_{i,j}$ for $j \in \{z, T\}$.

Case 1: $\mathcal{M}_z \not\supset \mathcal{M}_T$

In this case, $R^2_{z} < R^2_{T}$ (see Lemmas B.1 and B.2 in Maruyama and George (2011)) indicating that $(\frac{1-R^2_T}{1-R^2_z}) < 1$. For $i \in H_z$, $R^2_{i,z} \rightarrow 0$ and $mR^2_{i,z} = O(1)$ (from Lemma 4.4.1), which leads to

$$\lim_{m \rightarrow \infty} BF(\mathcal{M}_z : \mathcal{M}_T) = \lim_{m \rightarrow \infty} O(m^s)O(f^m).$$

Here $0 < f < 1$, and so regardless of the choice of $s \in \mathbb{R}$, $\lim_{m \rightarrow \infty} BF(\mathcal{M}_z : \mathcal{M}_T) = 0.$
Case 2: $\mathcal{M}_z \supset \mathcal{M}_T$

This situation arises when $\mathcal{M}_z$ contains all the design blocks in $\mathcal{M}_T$ and has at least one more block not in $\mathcal{M}_T$. By Lemma 4.4.1 (iii) and (v), $(1 - \frac{R_{i,z}^2}{1 - R_z^2})^m = O(1)$ and $mR_{i,z}^2 = O(1)$ for all $i \in H_z \subset B_z \setminus B_T$. Then the logarithm of the Bayes factor is approximated as

$$m \log \left(1 - \frac{R_T^2}{1 - R_z^2}\right) + (b_T - b_z) \log m - \sum_{i \in B_z} b_{i,z} \log (R_{i,z}^2) + (k_T - q_z) \log m$$

$$= O(1) + \left[(k_T - q_z) + \frac{1}{2} \sum_{i \in B_T} (p_{i,T} - p_{i,z}) + \frac{1}{2} \sum_{i \in B_z \setminus B_T} (0 - p_{i,z})\right. \left. + (a - 4) \frac{k_T - k_z}{2} + \sum_{i \in H_z} \left(\frac{a + p_{i,z}}{2} - 2\right)\right] \log m \quad \text{(since } mR_{i,z}^2 = O(1)\text{)}$$

$$= O(1) + \left[(k_T - k_z + L_z - \frac{1}{2} \sum_{i \in B_z \setminus B_T} p_{i,z} + (a - 4) \frac{k_T - k_z}{2}\right. \left. + \frac{1}{2} \sum_{i \in H_z} p_{i,z} + \frac{a - 4}{2} L_z\right] \log m \quad (p_{i,z} = p_{i,T} \forall i \in B_T)$$

$$= O(1) + \left[(a/2 - 1)(k_T - k_z + L_z) - \frac{1}{2} \sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z}\right] \log m$$

Now, $a > 2$ and $q_z = k_z - L_z \geq k_T$ for any model $\mathcal{M}_z \supset \mathcal{M}_T$, which implies that the expression

$$s = (a/2 - 1)(k_T - k_z + L_z) - \frac{1}{2} \sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z} \leq 0.$$  

When the above expression is strictly negative, \(\lim_{m \to \infty} BF(\mathcal{M}_z : \mathcal{M}_T) = \lim_{m \to \infty} O(m^s) = 0\), otherwise \(BF(\mathcal{M}_z : \mathcal{M}_T) \to^d W_z\) for a non-degenerate random variable $W_z$. It is possible for $s$ to equal zero and this happens, for example, when $H_z = B_z \setminus B_T$ so that $q_z = k_T$ and $B_z \setminus (B_T \cup H_z) = \phi$, the null set. In such a situation, model selection consistency does not hold.

116
When $\mathcal{M}_T = \mathcal{M}_0$, $BF(\mathcal{M}_z : \mathcal{M}_T) = BF(\mathcal{M}_z : \mathcal{M}_0) \overset{d}{\rightarrow} W_z$ for any model $\mathcal{M}_z$, and $P(\lim_{n \to \infty} BF(\mathcal{M}_z : \mathcal{M}_0) = 0) = 0$ so that the BOB hyper-$g$ prior is again model selection inconsistent.

**Theorem 4.4.4.** Consider a regression model with a BOB hyper-$g/n$ prior as described in (4.4), (4.2) and (3.14) which satisfies (4.13) and (4.14). This posterior distribution is consistent in model selection.

**Proof.** We imitate the proof of Theorem 3.6.4 appearing in Appendix A.14 to show that (4.15) holds.

When $\mathcal{M}_T \neq \mathcal{M}_0$, using a Laplace approximation as in Appendix A.14 we get

$$\lim_{m \to \infty} BF_{BOB}\text{-}g/n(\mathcal{M}_z : \mathcal{M}_T) = \lim_{m \to \infty} O(m^{k_T-k_z+\frac{a}{2}(q_z-k_T)}) BF_{BOB}\text{-}g(\mathcal{M}_z : \mathcal{M}_T),$$

where the subscripts denote the corresponding Bayes factors for the BOB hyper-$g$ and BOB hyper-$g/n$ priors.

**Case 1:** $\mathcal{M}_z \not\supset \mathcal{M}_T$

It has been shown in the proof of Theorem 4.4.3 that $BF_{BOB}\text{-}g(\mathcal{M}_z : \mathcal{M}_T)$ goes to zero as $m \to \infty$ ($n \to \infty$) at an exponential rate and hence $BF_{BOB}\text{-}g/n(\mathcal{M}_z : \mathcal{M}_T) \to 0$ for all such models $\mathcal{M}_z$, regardless of the values of $k_T$, $k_z$ and $q_z$.

**Case 2:** $\mathcal{M}_z \supset \mathcal{M}_T$

The log Bayes factor in this case equals

$$O(1) + \left[(a/2 - 1)(k_T - k_z + L_z) - \frac{1}{2} \sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z}\right] \log m$$

$$+ \left[\frac{a}{2}(q_z - k_T) + k_T - k_z\right] \log m$$

$$= O(1) + \left[\frac{1}{2} \sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z} - (k_T - k_z + L_z) + k_T - k_z\right] \log m$$

$$= O(1) - \left[L_z + \frac{1}{2} \sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z}\right] \log m$$

117
Observe that $L_z$ and $\sum_{i \in B_z \setminus (B_T \cup H_z)} p_{i,z}$ can never be zero simultaneously. If $L_z = 0$, then $H_z = \phi$ and there is at least one $i \in B_z \setminus B_T$ such that $p_{i,z} > 0$ as $M_z \supset M_T$. The other term can equal zero only if $H_z = B_z \setminus B_T$ and clearly $L_z > 0$ in such case. This ensures that $\lim_{n \to \infty} BF_{BOBHg/n}(M_z : M_T) = 0$ for any model $M_z$ proving model selection consistency when $M_T \neq M_0$.

The case $M_T = M_0$ is exactly similar to Case 2, but with the extra restriction that $B_T = \phi$. The posterior distribution under the BOB hyper-$g/n$ prior is thus model selection consistent for any choice of the true model. \hfill \Box

The BOB hyper-$g/n$ prior is model selection consistent while the BOB hyper-$g$ prior is not, because of the same argument in favor of the hyper-$g/n$ mixing distribution in the ordinary $g$ and block $g$ priors. The hyper-$g/n$ prior converges to a stable limit, whereas the hyper-$g$ prior degenerates to a point mass at 0 as $n \to \infty$ under Condition 3.5.2 or (4.13). Notice that the conditions in (4.13) and (4.14) are important and have been used in the previous two theorems through the application of Lemma 4.4.1.

### 4.4.3 Prediction Consistency

The Bayes-optimal prediction $\tilde{y}_n^*$ of the unknown, true response $y^*$ corresponding to a new vector of predictors $x^* \in \mathcal{R}^p$ is the model averaged prediction

$$\tilde{y}_n^* = E(\alpha \mid y) + \sum_{z \in \Gamma_B} \pi(M_z \mid y)x^{*T}E(\beta \mid y, M_z),$$

where $y$ is the vector of $n$ data points observed simultaneously with the $n \times p$ matrix of covariate values $X$. Prediction consistency is achieved when $\tilde{y}_n^* \xrightarrow{P} E(y^*) = \alpha_T + x^{*T}\beta_T$ as $n \to \infty$. The BOB hyper-$g$ and BOB hyper-$g/n$ priors are both prediction...
consistent under Bayesian model averaging (BMA), as displayed by the next set of results.

**Theorem 4.4.5.** Consider a regression model with a BOB hyper-\(g\) prior as described in (4.4), (4.2) and (3.9) which satisfies 4.13 and 4.14. The predictions under BMA are consistent for this model.

**Proof.** The proof is similar to the proof of Theorem 3.5.3. First, observe that block orthogonality of the reparameterized design gives

\[
E(\tau \mid y, M_z) = E\left[ E(\tau \mid y, g, M_z) \right] = \int \left( \frac{g_{i1}}{1+g_{i1}} \hat{\tau}_{i1,z,LS} \right) \cdots \left( \frac{g_{ikz}}{1+g_{ikz}} \hat{\tau}_{ikz,z,LS} \right) \pi(g \mid y, M_\gamma) dg
\]

and \(E(\alpha \mid y) = \hat{\alpha}_{LS}\) assuming \(B_z = \{i_1, i_2, \ldots, i_{k_z}\}\).

When \(M_T = M_0\), \(\hat{\tau}_{z,LS} \overset{P}{\to} 0\) and \(\hat{\alpha}_{LS} \overset{P}{\to} \alpha_T\) for every \(M_z\) since least squares estimators are consistent. Thus the model averaged prediction \(\hat{y}_n^*\) converges to \(E(y^*) = \alpha_T\).

Let \(\Omega\) denote the set of all models \(M_z\) (including \(M_T\)) such that \(\lim_{n \to \infty} \pi(M_z \mid y) > 0\). Theorem 4.4.3 shows that the set \(\Omega\) can only include some models \(M_z\) of the form \(M_z \supseteq M_T\). Thus, \(\lim_{n \to \infty} \sum_{z: M_z \in \Omega} \pi(M_z \mid y) = 1\). When \(M_T \neq M_0\), the least squares estimates for the reparameterized coefficients in \(M_z \in \Omega\) are consistent, so that \(\hat{\tau}_{i,z,LS} \overset{P}{\to} \tau_{i,T}\) for \(i \in B_T\) and \(\hat{\tau}_{i,z,LS} \overset{P}{\to} 0\) for \(i \notin B_T\).

We simply need to show that Corollary 3.5.1 is true under this setup, i.e.,

\[
\lim_{n \to \infty} E\left( \frac{g_i}{1+g_i} \mid M_z, y \right) = 1 \quad \forall \ i \in B_T \text{ when } M_z \in \Omega, \text{ while } 0 \leq E\left( \frac{g_i}{1+g_i} \mid M_z, y \right) \leq 1
\]
for all other $i$. But $\beta$ and $\tau$ are related as $\beta_z = U^{-1}\tau_z$ so that $\hat{\beta}_{z,LS} = U^{-1}\hat{\tau}_{z,LS}$ and $E(\beta \mid y, g, M_z) = U^{-1}E(\tau \mid y, g, M_z)$. Hence,

$$
\lim_{n \to \infty} \hat{y}_n^* = E(\alpha \mid y) + \lim_{n \to \infty} \sum_{z \in \Gamma_B} \pi(M_z \mid y)x^T U^{-1}E(\tau \mid y, M_z)
$$

$$
= \alpha_T + x^T U^{-1} \tau_T \lim_{n \to \infty} \sum_{z : M_z \in \Omega} \pi(M_z \mid y)
$$

$$
= \alpha_T + x^T \beta_T = E(y^*)
$$

which makes the BOB hyper-$g$ prior consistent in prediction under BMA.

The proof of the theorem will be complete after verifying that Corollary 3.5.1 holds. The only requirement for Corollary 3.5.1 to be true is that the relevant component $R^2_{i,z}$ is strictly greater than zero in the limit as $n \to \infty$ (see proof of Lemma 3.5.1 in Appendix A.12). Lemma 4.4.1 (i) and (ii) confirm that $\lim_{n \to \infty} R^2_{i,z} > 0$ for all $i \in B_T$ and all $z$ such that $M_z \supseteq M_T$.

**Theorem 4.4.6.** Consider a regression model with a BOB hyper-$g/n$ prior as described in (4.4), (4.2) and (3.14) which satisfies 4.13 and 4.14. The predictions under BMA are consistent for this model.

**Proof.** The derivation follows directly from the proof of Theorem 3.6.5 in Appendix A.15. As the BOB hyper-$g/n$ prior is *model selection consistent*, $\lim_{n \to \infty} \sum_{z \in \Gamma_B} \pi(M_z \mid y) = \lim_{n \to \infty} \pi(M_T \mid y) = 1$. We only need to show that

$$
\lim_{n \to \infty} E \left( \frac{g_i}{1 + g_i} \mid M_T, y \right) = 1, \ \forall \ i \in B_T
$$

(4.16)

in the BOB hyper-$g/n$ prior and then use the steps in the proof of Theorem 4.4.5 to demonstrate *prediction consistency*. The derivation of (4.16) has already been done in Appendix A.15 and in the proof of Theorem 4.4.5 earlier. \qed
4.5 BOB Hyper-\( g \) Prior over all Block Permutations

The definition of the BOB \( g \) prior hinges on two distinct features, both of which strongly determine the performance of the associated model. In order to achieve the best performance, firstly, an appropriate choice of blocks (groups) of predictor variables is necessary, and secondly, the groups need to be ordered according to their relative importance in the regression model. In many regression problems it is possible to identify latent theoretical constructs that connect sets of covariates. The latent constructs justify the suitability of a common scale parameter for each of these blocks of predictors and motivate a natural block structure in the design. However, even when the proper group structure is known, it is a challenging task to order the covariate blocks accurately in the block Gram–Schmidt process. As demonstrated in Section [4.3.2], inference with the BOB hyper-\( g \) or hyper-\( g/n \) prior suffers if the blocks are poorly arranged in the implementation of the block Gram–Schmidt orthogonalization. This weakness of the mixtures of BOB \( g \) priors can be ameliorated if we define the prior to be invariant to the order of orthogonalization. In this section we suggest two new, closely related formulations of the BOB \( g \) prior based on all possible arrangements of the blocks of predictors, which obviate predetermining the “correct” ordering of blocks for sensible inference. We call each such prior the block orthogonalized permuted block \( g \) prior or the BOPB \( g \) prior.

4.5.1 Prior with Equal Weights on Permutations

Consider a regression model with \( K \) blocks for the total \( p \) predictor variables in the full model. We consider the same model space \( \Gamma_B = \{0, 1\}^K \) as in Section [4.4] that consists of \( 2^K \) models resulting from the inclusion or exclusion of entire blocks of
covariates. The subscript $z \in \Gamma_B$ in a model $M_z$ is used to denote the set of blocks present in the particular model. Any model $M_z$ is thus defined by its $k_z$ blocks of active predictors, where $k_z = \sum_{i=1}^{K} I(z_i = 1)$.

The first form of the new order-invariant BOPB $g$ prior assigns the same prior weight to each arrangement of covariate blocks in the block Gram–Schimdt process. For a model with $k$ blocks, a block $g$ type prior is assigned to the vector of regression coefficients over all $k!$ permutations of the $k$ distinct blocks. The BOPB $g$ prior on a model $M_z$ consisting of $k_z$ blocks is defined as:

$$
\pi(\mathbf{z} \mid M_z, \sigma^2) = \frac{1}{k_z!} N(U_z^{(1)} \beta_z \mid 0, A_z^{(1)} \sigma^2) + \frac{1}{k_z!} N(U_z^{(2)} \beta_z \mid 0, A_z^{(2)} \sigma^2) + \cdots \\
\cdots \cdots + \frac{1}{k_z!} N(U_z^{(k_z!)} \beta_z \mid 0, A_z^{(k_z!)} \sigma^2) \quad (4.17)
$$

where $N(x \mid \mu, \Sigma)$ denotes a Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$ on the random variable $x$. Here $U_z^{(j)}$, $j = 1, 2, \ldots, k_z!$, represents the upper triangular matrix resulting from the block Gram–Schimdt process with respect to the $j^\text{th}$ permutation of the blocks from the original design $X_z$. For the BOPB $g$ prior, $A_z^{(j)}$ is the usual prior covariance matrix for the BOB $g$ prior when the reparameterized design $Q_z$ is derived from the $j^\text{th}$ permutation of the blocks. The BOPB hyper-$g$ and BOPB hyper-$g/n$ priors are obtained by adding prior distributions on the latent scale vector $g$ as in (3.9) and (3.14).

The Bayes factor comparing model $M_z$ to the null model $M_0$

$$
BF(M_z : M_0) = \frac{p(y \mid M_z)}{p(y \mid M_0)} = \frac{1}{p(y \mid M_0)} \int p(y \mid M_z, \beta_z, \alpha, \sigma^2) \pi(\beta_z \mid M_z, \sigma^2) \pi(\alpha, \sigma^2) d\beta_z d\alpha d\sigma^2
$$
is a weighted average of the individual Bayes factors arising from each unique permuta-
tion. Observe that

\[
BF(M_z : M_0) = \frac{p(y \mid M_z)}{p(y \mid M_0)} = \frac{\sum_{j=1}^{k_z} p(y \mid M_z, j)}{k_z! p(y \mid M_0)}
\]

\[
= \frac{1}{k_z!} \sum_{j=1}^{k_z} BF^{(j)}(M_z : M_0),
\]

where \(p(y \mid M_z, j)\) is the component of the marginal likelihood due to the \(j^{th}\) per-
mutation of blocks,

\[
BF^{(j)}(M_z : M_0)
\]

\[
= \frac{1}{p(y \mid M_0)} \int p(y \mid M_z, \beta_z, \alpha, \sigma^2) N(U_z^{(j)} \beta_z; 0, A_z^{(j)} \sigma^2) \pi(\alpha, \sigma^2) d\beta_z d\alpha d\sigma^2
\]

\[
= \left( \frac{a - 2}{2} \right)^{k_z} \int_{(0,1)^{k_z}} \left[ \prod_{i=1}^{k_z} (1 - t_i)^{a + p_i z} \right] \left(1 - \sum_{i=1}^{k_z} t_i [R_i^{(z,j)}]^2 \right)^{-\frac{n-1}{2}} dt,
\]

and \([R_i^{(z,j)}]^2 = \frac{y^T Q_{i,z} y}{y^T y}\) is the \(i^{th}\) component of \(R^2\) from model \(M_z\). The orthogonal
blocks \(Q_{i,z}^{(j)}\) can be derived as in (4.2) when the blocks \(X_{i,z}\) from model \(M_z\) are
arranged according to the \(j^{th}\) permutation, \(j = 1, \ldots, k_z!\).

### 4.5.2 Prior on Model Space Encompassing all Permutations

We can take a second view of the model space, identifying the model by its blocks
of active predictors and the permutations of these blocks. The model space con-
tains all \(k!\) permutations for each model having \(k\) blocks, \(k = 0, 1, \ldots, K\), and the
permutations decide the order in which the \(k\) blocks are orthogonalized. Thus, the
size of this larger model space is \(|M| = \sum_{k=0}^{K} \binom{K}{k} k! = \sum_{k=0}^{K} k P_k = e \Gamma(K + 1, 1)\),
where \(\Gamma(s, x) = \int_x^\infty t^{s-1} e^{-t} dt\) is the upper incomplete gamma function. The expres-
sion \(e \Gamma(K + 1, 1)\) is always guaranteed to be an integer for any \(K \in \mathbb{N} \cup \{0\}\) due
to the extra exponential factor. In this setup, we index a model by a subscript $z$ to denote the set of included blocks (as in Section 4.5.1) and a superscript $j$ to indicate a specific permutation of those blocks.

The second form of the BOPB $g$ prior is defined as:

$$
\begin{align*}
\mathbf{y} | \alpha, \beta, \sigma^2 & \sim N(\alpha \mathbf{1} + X_z \beta_z, \sigma^2 I) \\
\pi(\beta_z | M_z^i, \sigma^2) & = N(U_z^{(j)} \beta_z; 0, A_z^{(j)} \sigma^2), \quad j = 1, 2, \ldots, k_z! \\
\pi(\alpha, \sigma^2) & \propto \frac{1}{\sigma^2}
\end{align*}
$$

where $k_z$ is the number of blocks in the unordered model $M_z$ and $M_z^i$ refers to the model with the $j^{th}$ permutation of the blocks in $M_z$. $U_z^{(j)}$ and $A_z^{(j)}$ are defined in (4.17). The BOB hyper-$g$ and BOB hyper-$g/n$ priors result from the prior distributions (3.9) and (3.14) respectively on the parameter $g$. As before, the Bayes factor for the model $M_z^j$ is given by:

$$
BF(M_z^j : M_0) = BF^{(j)}(M_z : M_0)
$$

$$
= \left( \frac{a - 2}{2} \right)^{k_z} \int_{(0,1)^{k_z}} \left[ \prod_{i=1}^{k_z} (1 - t_i)^{a + \nu_i \frac{z - 2}{2}} \right] \left( 1 - \sum_{i=1}^{k_z} t_i[R_z^{(j)}]_i^2 \right)^{-\frac{n-1}{2}} dt.
$$

This variation of the BOB $g$ prior is not order-invariant in the typical sense, since the prior on each model $M_z^j$ depends on the order of block orthogonalization within the model. Nevertheless, since all feasible arrangements of the blocks of predictors are included in the model space as separate models, the “correct” arrangement is also guaranteed to contribute to inference.

4.5.3 Inference with BOPB Hyper-$g$ Priors

The benefit of using a BOPB hyper-$g$ or hyper-$g/n$ prior distribution with prior component (4.17) or (4.18) on the regression coefficients is that the choice of the order
in which the blocks are orthogonalized does not have any adverse effect. The pitfalls of incorrect judgment in sorting the blocks, exhibited in the form of CLP and CELS in the BOB hyper-\(g\) prior over a single arrangement, are avoided. The permutations are the key to avoiding these problems. For a given set of \(k\) design blocks, \((k - 1)!\) permutations will have the “largest” coefficient block in the first position, placing us in the ambit of Theorem 4.3.2 (avoiding the CLP) and 4.3.4 (avoiding the ELS). We turn to an adjustment of our earlier theory to formally handle the permutations.

Another prominent characteristic of these priors is that prior specification is now invariant to the order of block orthogonalization, in addition to the usual invariance to blockwise linear reparameterizations. Prior (4.17) is invariant to the order of orthogonalization since it averages over all possible arrangements while prior (4.18) treats distinct block arrangements as separate models in the model space.

Although the prior distributions of the form (4.17) and (4.18) make the BOPB hyper-\(g\) prior invariant to the order of orthogonalization, it is not immediate that the procedure does not suffer from the CLP and CELS. We show in the next two lemmas that in the situation when one set of regression coefficients is huge compared to the rest and should be identified as the most important block, the BOPB hyper-\(g\) prior concentrates posterior probability only on those models where the important block has been correctly identified. We prove the results only for prior form (4.18), the implications on prior form (4.17) are the same. Under appropriate limits, the entire posterior probability is placed on the components corresponding to the “correct” permutations which contain the most important block in the first position.

Consider the sequence of regression problems \(\{\Psi_N\}\) described in (4.5) with a prior distribution of the form (4.18) and (3.9) assigned to each model in the sequence. The
model space $\mathcal{M}$ consists of $\mathcal{e}\Gamma(K + 1, 1)$ models arising from entire blocks of predictors being included or excluded from consideration together with all permutations of the set of included blocks. Let $I^z_s$ represent the set of block permutations from an unordered model $\mathcal{M}_z$ which includes all covariates in $X_s$, where $X_s$ is the first and the most important block in the block Gram–Schmidt process. If $k_z$ is the number of design blocks in $\mathcal{M}_z$, then $|I^z_s| = (k_z - 1)!$.

**Lemma 4.5.1.** Consider the sequence of problems $\{\Psi_N\}$ defined in (4.5) with a BOPB hyper-$g$ prior as in (4.18) and (3.9) on each model in the sequence. Let $l_1$ and $l_2$ denote permutation indices for an unordered model $\mathcal{M}_z$ containing block $X_s$, such that $l_1 \in I^z_s$ and $l_2 \notin I^z_s$. When Condition 4.3.1 holds, $a \geq 3$ and $n > (2k_z + 1) \lor (a + p_z + 2q - 3)$, where $q \geq 2$ is the position of block $X_s$ in model $\mathcal{M}_z^l$, $\lim_{N \to \infty} BF(\mathcal{M}_z^l : \mathcal{M}_z^l) = 0$.

**Proof.** The Bayes factor $BF(\mathcal{M}_z^l : \mathcal{M}_z^l)$ comparing model $\mathcal{M}_z^l$ to model $\mathcal{M}_z^l$ has the limit

$$
\lim_{N \to \infty} BF(\mathcal{M}_z^l : \mathcal{M}_0) = \frac{\int_{(0,1)^k} \prod_{i=1}^{k_z} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{k_z} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}{\int_{(0,1)^k} \prod_{i=1}^{k_z} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{k_z} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}
$$

$$= C_1 \lim_{N \to \infty} \frac{\int_{(0,1)^q} \prod_{i=1}^{q} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{q} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}{\int_{(0,1)^q} \prod_{i=1}^{q} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{q} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}
$$

(by Lemma 4.3.3 as $n > 2k_z + 1$)

$$= C_2 \lim_{N \to \infty} \frac{\int_{(0,1)^q} \prod_{i=1}^{q} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{q} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}{\int_{(0,1)^q} \prod_{i=1}^{q} (1 - t_i)^{a + p_z - 2} \left( 1 - \sum_{i=1}^{q} t_i [R_i(z_{t_1})]^2 \right)^{n-1} dt}
$$

126
where \( q \) denotes the maximum index \( i \) with \( \lim_{N \to \infty} [R_i^{(z,l_2)}]^2 > 0 \). It is easy to see that \( q \geq 2 \) since \( l_2 \notin I_z \). As \( N \to \infty \), \( L = [R_i^{(z,l_1)}]^2 \to 1 \), \( M = \sum_{i=1}^{q} [R_i^{(z,l_2)}]^2 \to 1 \) and
\[
\frac{1-L}{1-M} = \frac{1-\sum_{i=1}^{q} [R_i^{(z,l_2)}]^2}{\sum_{i=1}^{q} [R_i^{(z,l_2)}]^2} = O(1) / O(1) = O(1).
\]
We can show with a little effort that the \( O(1) \) term equals a finite non-zero constant in the limit.

Lemma 4.3.2 can be applied whenever \( n > a + p_s + 2q - 3 \), and thus the Bayes factor tends to zero as \( N \to \infty \).

**Lemma 4.5.2.** Consider the sequence of problems \( \{\Psi_N\} \) defined in (4.5) with a BOPB hyper-\( g \) prior as in (4.18) and (3.9) on each model in the sequence. Let \( M^* \) denote a model with \( X_s \) as the first block in the block Gram–Schmidt process and \( M' \) be any other model which does not have \( X_s \) as the first block. For sufficiently large \( n \), when Condition 4.3.1 holds and \( a \geq 3 \),
\[
\lim_{N \to \infty} BF(M' : M^*) = 0.
\]

**Proof.** Let \( k^* \) and \( k' \) denote the number of blocks, \( p_i^* \) and \( p'_i \) the size of the \( i^{th} \) block, and \( [R_i^*]^2 \) and \( [R_i']^2 \) the component of \( R^2 \) due to the \( i^{th} \) orthogonal block in models \( M^* \) and \( M' \) respectively. It is easy to see that \( p_1^* = p_s \), and since the sequence \( \{\Psi_N\} \) has \( k \) blocks in total, \( k^* \in \{1, 2, \ldots, k\} \) and \( k' \in \{0, 1, \ldots, k\} \).

**Case 1:** \( M' \) does not contain the block \( X_s \).

If \( M' \neq \mathcal{M}_0 \), then
\[
\lim_{N \to \infty} BF(M' : M^*) = \lim_{N \to \infty} BF(M' : \mathcal{M}_0) / BF(M^* : \mathcal{M}_0)
\]
\[
= \lim_{N \to \infty} \left( \frac{a-2}{2} \right) ^{k' - k^*} \int_{(0,1)^{k'}} \left[ \prod_{i=1}^{k'} (1 - t_i)^{\frac{a + p_i'}{2} - 2} \right] \left( 1 - \sum_{i=1}^{k'} t_i [R_i']^2 \right) ^{-\frac{n-1}{2}} dt \int_{(0,1)^{k^*}} \left[ \prod_{i=1}^{k^*} (1 - t_i)^{\frac{a + p_i^*}{2} - 2} \right] \left( 1 - \sum_{i=1}^{k^*} t_i [R_i^*]^2 \right) ^{-\frac{n-1}{2}} dt
\]
\[
= 0
\]

127
when \( n > k^*(a - 2) + \sum_{i=1}^{k^*} p_i^* + 1 \) (see proof of Theorem 3.5.1 in Appendix A.10).

Since \( \sum_{i=1}^{k^*} [R_i']^2 \to 0 \) and \( \sum_{i=1}^{k^*} [R_i^*]^2 \to 1 \) as \( N \to \infty \), the numerator is finite for all \( N \), but the denominator diverges to \( \infty \) with \( N \).

If \( M' = \mathcal{M}_0 \), \( BF(\mathcal{M}_0 : M^*) \to 0 \) as \( N \to \infty \) like before, when \( n > k^*(a - 2) + \sum_{i=1}^{k^*} p_i^* + 1 \).

Case 2: \( M' \) contains the block \( X_s \).

Case 2A: \( M' \) and \( M^* \) are distinct permutations of the same set of blocks.

This is exactly the same situation as in Lemma 4.5.1 and so \( \lim_{N \to \infty} BF(M' : M^*) = 0 \) when \( a \geq 3 \), \( n > a + p_s + 2k - 3 \) and Condition 4.3.1 holds. Note that the condition on \( n \) is sufficient to ensure that the conclusion of Lemma 4.5.1 holds for any position of the block \( X_s \) within model \( M' \).

Case 2B: \( M' \) and \( M^* \) have a different collection of blocks.

\[
\lim_{N \to \infty} BF(M' : M^*) = \lim_{N \to \infty} \frac{BF(M' : \mathcal{M}_0)}{BF(M^* : \mathcal{M}_0)} = \lim_{N \to \infty} \left( \frac{a - 2}{2} \right)^{k' - k^*} \frac{\int_{(0,1)^{k'}} \left[ \prod_{i=1}^{k'} (1 - t_i) \right] \frac{\prod_{i=1}^{k'} (1 - t_i) \frac{a + p_i'}{2} - 2}{\prod_{i=1}^{k'} (1 - t_i) \frac{a + p_i}{2} - 2}}{\int_{(0,1)^{k^*}} \left[ \prod_{i=1}^{k^*} (1 - t_i) \right] \frac{\prod_{i=1}^{k^*} (1 - t_i) \frac{a + p_i^*}{2} - 2}{\prod_{i=1}^{k^*} (1 - t_i) \frac{a + p_i}{2} - 2}} \left( 1 - \sum_{i=1}^{k'} t_i [R_i']^2 \right)^{-\frac{n-1}{2}} dt'.
\]

Since \( M' \) contains block \( X_s \) in a position other than the first one, \( \lim_{N \to \infty} [R_i']^2 > 0 \) for at least two \( i \), while only \( \lim_{N \to \infty} [R_i^*]^2 > 0 \) among all \( [R_i^*]^2 \), \( i = 1, \ldots, k^* \). Using Lemmas 4.3.2 and 4.3.3 as in the proof for Case 2A, \( \lim_{N \to \infty} BF(M' : M^*) = 0 \) when \( a \geq 3 \), \( n > a + p_s + 2k - 3 \) and Condition 4.3.1 holds.

Hence, when \( n \) is sufficiently large such that both sets of restrictions on \( n \) (from Cases 1 and 2) are satisfied, the Bayes factor \( BF(M' : M^*) \) under the BOPB hyper-\( g \)-prior converges to zero in the limit.
Lemmas 4.5.1 and 4.5.2 indicate that when $||\beta_s|| \to \infty$, the BOPB hyper-$g$ prior allocates posterior probability (given a suitable prior on the model space) entirely to the set of permutations where $X_s$ is the first block. The models with “incorrect” order of block orthogonalization, which suffer from the CLP and CELS, do not get any posterior weight in the limit. Hence, when $N \to \infty$ in the sequence $\{\Psi_N\}$ in (4.5), these models cannot affect model-averaged coefficient estimates or play a detrimental role in model selection.

The limit that drives the CLP and ELS can be extended to allow different blocks of coefficients to grow at different rates. In the upcoming Lemmas 4.5.3 and 4.5.4 we show that the posterior concentrates on a unique set of models/permutations when the design blocks can be ordered according to importance.

Consider a new sequence of regression problems $\{\Psi_N\} = \{X_1(N), \ldots, X_k(N), \alpha_N, \beta_1(N), \ldots, \beta_k(N), \epsilon_N\}$ similar to the one in (4.5). In this sequence, the groups of regression coefficients grow in size simultaneously, but at different rates while the design is unaltered. $\beta_1$ grows in size at the fastest rate, $\beta_2$ at the next highest rate and so on. $\beta_{k-1}$ is the slowest growing set of coefficients and $\beta_k$ is fixed throughout the sequence. Thus

$$\Psi_N = \{X_1, \ldots, X_k, \alpha, \beta_1(N), \ldots, \beta_{k-1}(N), \beta_k, \epsilon\} \quad \text{(4.19)}$$

with $\lim_{N \to \infty} ||\beta_i|| = \infty$ and $\lim_{N \to \infty} \frac{||\beta_{i+1}||}{||\beta_i||} = 0$ for all $i = 1, 2, \ldots, k-1$.

Among the $k!$ models containing all $k$ active design blocks, the BOPB hyper-$g$ posterior concentrates on the “true” arrangement of covariate blocks. Lemma 4.5.3 establishes this result for the case when $a = 3$ and each block $\beta_i$ is of size one. In this situation, there are $k = p$ blocks and $p_i = 1$, $\forall \ i = 1, \ldots, p$. Extension of the
result to the general case with arbitrary \( a \in (2, 4] \) and arbitrary \( p_i \geq 1 \) for all \( k \), is currently being investigated.

**Lemma 4.5.3.** Consider the sequence of problems \( \{ \Psi_N \} \) defined in (4.19) with a BOPB hyper-\( g \) prior as in (4.18) and (3.9) on each model in the sequence. Assume that \( a = 3 \) and \( p_i = 1 \) \( \forall i \) so that there are exactly \( p \) blocks in the design. Let \( \mathcal{M}_1 \) denote the model with the design blocks orthogonalized in the correct order of importance \( \{X_1, X_2, \ldots, X_p\} \) and \( \mathcal{M}_l \) denote any other model with a different permutation of the same \( p \) blocks. When Condition 4.3.1 holds and \( n > 2p + 1 \), \( \lim_{N \to \infty} BF(\mathcal{M}_l : \mathcal{M}_1) = 0 \).

The proof of this lemma is given in Appendix B.5.

Assume the same setup as in Lemma 4.5.3. Let \( S \) denote the set containing only two models, one with blocks \( \{X_1, X_2, \ldots, X_p\} \) and the other with blocks \( \{X_1, X_2, \ldots, X_{p-1}\} \), both models having design blocks orthogonalized in that specific order.

**Lemma 4.5.4.** Consider the sequence of problems \( \{ \Psi_N \} \) defined in (4.19) with a BOPB hyper-\( g \) prior as in (4.18) and (3.9) on each model in the sequence. Assume that \( a = 3 \) and \( p_i = 1 \) \( \forall i \) so that there are exactly \( p \) blocks in the design. Let \( M^* \in S \) and \( M' \not\in S \) be any model of an arbitrary size. When Condition 4.3.1 holds and \( n \) is sufficiently large, \( \lim_{N \to \infty} BF(M' : M^*) = 0 \).

The proof of this lemma is provided in Appendix B.6. The posterior distribution under a BOPB hyper-\( g \) prior defined by (4.18) and (3.9) focuses only on the models that include the important blocks, and among such models, only on the arrangements that correctly recognize the relative importance of the blocks. In a practical regression problem, coefficients will differ in size, but will be of fixed magnitude, and the BOB hyper-\( g \) posterior will allocate a lot of probability to the model with the important
blocks of covariates with the “correct” order of block orthogonalization. The same behavior is also observed under the BOPB hyper-\(g\) prior defined by \((4.17)\) and \((3.9)\), which substantiates our faith in models where such prior distributions are imposed. Since the models with the most important blocks in the “correct” permutation get most of the posterior weight (they get all the posterior weight in the limit), these models contribute the most to model-averaged inferences.

Identical conclusions hold for the BOPB hyper-\(g/n\) prior defined by \((4.17)\) and \((3.14)\) or by \((4.18)\) and \((3.14)\) and versions of all the lemmas from this section related to the BOB hyper-\(g/n\) prior follow directly from previous developments. Once again, the same set of sufficient conditions is adequate for the new theoretical results. We recommend using the BOPB hyper-\(g/n\) prior in applications instead of the BOPB hyper-\(g\) prior because (under mild and standard conditions) the hyper-\(g/n\) prior converges to a limit as \(n \to \infty\) while the hyper-\(g\) prior does not, and because of the consistency arguments in favor of the prior scaled by sample size (see Section 4.4). The BOPB \(g\) prior is simply another form of the BOB \(g\) prior, and a mixture of BOPB \(g\) priors has the same behavior in regards to the three main kinds of asymptotic consistency properties discussed in Section 4.4 as the correpnding mixture of BOB \(g\) priors.

4.6 Application to Boston Housing Data

This section studies the behavior of the BOPB hyper-\(g/n\) prior in an applied example. The Boston housing data set appearing in Harrison and Rubinfeld (1978) consists of 506 observations on 14 variables concerning housing values in the suburbs of Boston in 1970. A brief description of the data set is provided in Appendix B.7. The
median value of owner-occupied homes is the dependent variable which is regressed on the thirteen other covariates. Following the analysis in [Harrison and Rubinfeld (1978)], the predictor variables \( \text{dis}, \text{rad}, \text{lstat} \) and the response variable \( \text{medv} \) are log transformed, while variables \( \text{nox}^2 \) and \( \text{rm}^2 \) are included as predictors instead of the corresponding linear terms. All of the predictors and the response are centered and scaled to have zero mean and unit variance.

4.6.1 Specification of the BOPB Hyper-\( g/n \) Prior

Implementation of the BOPB hyper-\( g/n \) prior requires grouping the predictor variables into blocks, and five different block structures are contrasted in this problem. The five block structures are listed below.

**Form 1:** Block 1- \( \log(lstat), \text{crim} \); Block 2- \( \text{chas}, \text{nox}^2, \text{rm}^2, \log(\text{dis}), \log(\text{rad}), \text{tax}, \text{ptratio}, b \); Block 3- \( \text{zn}, \text{indus}, \text{age} \).

**Form 2:** Block 1- \( \text{zn}, \text{indus}, \text{age} \); Block 2- \( \text{crim}, \text{chas}, \text{nox}^2, \text{rm}^2, \log(\text{dis}), \log(\text{rad}), \text{tax}, \text{ptratio}, b, \log(lstat) \).

**Form 3:** Block 1- \( \log(lstat), \text{crim} \); Block 2- \( \text{zn}, \text{indus}, \text{chas}, \text{nox}^2, \text{rm}^2, \text{age}, \log(\text{dis}), \log(\text{rad}), \text{tax}, \text{ptratio}, b \).

**Form 4:** Block 1- \( \log(lstat) \); Block 2- \( \text{crim}, \text{zn}, \text{indus}, \text{chas}, \text{nox}^2, \text{rm}^2, \text{age}, \log(\text{dis}), \log(\text{rad}), \text{tax}, \text{ptratio}, b \).

**Form 5:** Block 1- \( \text{nox}^2 \); Block 2- \( \text{crim}, \text{zn}, \text{indus}, \text{chas}, \text{rm}^2, \text{age}, \log(\text{dis}), \log(\text{rad}), \text{tax}, \text{ptratio}, b, \log(lstat) \).

The motivation behind most of these choices comes from an examination of the \( t \) statistics (used in the \( t \) test) associated with each of the predictor variables. A classical regression analysis with the data set reveals that most of the predictor variables in
the model are significant while only a few of them are not so important. The \( t \) statistics from the covariates are huge for a couple of predictors and relatively big for a lot of the other ones. The blocks in Form 1 are constructed based on this feature. Covariates in block 1 are believed to be highly important since they have large absolute values of the \( t \) statistic. The covariates in block 3 are considered to be the least important due to low \( t \) statistic values, while the rest in block 2 with medium to big absolute \( t \) values are moderately important. The block structure is chosen so that the estimated coefficients from block 1 will be left unchanged from least squares estimates, whereas coefficients from block 3 will be shrunk strongly toward zero, highlighting the flexibility in the shrinkage pattern of the blockwise prior.

The other group structures are chosen with the same rationale. The structure in Form 2 separates only the unimportant predictor variables while Form 3 places the very important covariates in a block of their own. The last two block structures, Form 4 and Form 5, isolate exactly one highly significant predictor from the rest to form a block by itself in either case. The structure in Form 1 seems to take advantage of the blocking strategy in the most effective way by placing the highly significant as well as the insignificant predictors in distinct blocks.

The BOPB hyper-\( g/n \) prior imposed on the model with any of the block structures (Forms 1-5) is the same prior as described in (4.18):

\[
\begin{align*}
\pi(y \mid \alpha, \beta_z, \sigma^2) &= N(\alpha 1 + X_z \beta_z, \sigma^2 I) \\
\pi(\beta_z \mid M^i_z, \sigma^2) &= N(U^{(j)}_z \beta_z; 0, A^{(j)}_z \sigma^2) \\
\pi(\alpha, \sigma^2) &\propto \frac{1}{\sigma^2} \\
\pi(g) &= \left(\frac{a - 2}{2n}\right)^{k_z} \prod_{i=1}^{k_z} \left(1 + \frac{g_i}{n}\right)^{-a/2}
\end{align*}
\]
where \( j \in \{1, 2, \ldots, k_z\} \) denotes a block permutation index and \( k_z, A_z^{(j)}, U_z^{(j)} \) have the usual definition as in (4.17) and (4.18).

The model space under consideration in the BOPB \( g \) prior formulation is the set of all models with entire blocks in and out of the model and all possible permutations for orthogonalization given each set of blocks. Thus the size of the model space is \( e\Gamma(K+1, 1) \) as mentioned in Section 4.5.2, with \( K \) being the total number of candidate blocks in the design. We assume a Bernoulli(1/2) block inclusion prior on the model space so that all combinations of blocks have the same probability \( a \) prior. But in this extended model space, each unique block combination has multiple representations through different permutations for block orthogonalization. The prior probability for any specific block combination is divided equally among all the distinct permutations.

### 4.6.2 Cross-validated Predictive Performance

This subsection compares the predictive performance of the new BOPB hyper-\( g/n \) prior with the performances of existing \( g \) priors and mixtures of \( g \) priors. The data set containing 506 observations is first randomly split into 11 equal parts of 46 observations each in order to carry out an eleven-fold cross-validation. The cross-validated predictive sum of squared errors (SSE) are calculated for the five different forms of the BOPB hyper-\( g/n \) prior and contrasted with the SSE values of existing \( g \) priors and their mixtures. All five forms of the BOPB hyper-\( g/n \) prior have the hyperparameter choice of \( a = 3 \) while the \textit{robust} prior of Bayarri et al. (2012) is specified with their recommended hyperparameter values of \( a = \frac{1}{2}, b = 1 \) and \( \rho = \frac{1}{p+1} \). The prior distribution on the \( 2^{13} = 8,192 \) models in the model space for the usual Bayesian methods mentioned in the left column of Table 4.1 is taken to be uniform.
Table 4.1 shows the improvement in predictive SSE when any of the different forms of the new prior is assigned to the linear model. The predictive SSE from a simple least squares fit of the full model with all 13 predictors is also mentioned in the table. Among the standard Bayesian methods, the Zellner–Siow full-based prior does the best job in prediction closely followed by the RIC and the hyper-$g$ priors. Surprisingly, the least squares fit of the full linear model gives a lower cross-validation error than the more complex traditional Bayesian methods. The block structure in prior Form 1 yields the lowest SSE among all methods. Prior Form 2, which moves the insignificant predictors into a separate block, also does a commendable job in prediction.

<table>
<thead>
<tr>
<th>Prior on $\beta$</th>
<th>SSE</th>
<th>Prior on $\beta$</th>
<th>SSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$-prior ($g = n$)</td>
<td>109.0293</td>
<td>$g$-prior ($g = p^2$)</td>
<td>108.5533</td>
</tr>
<tr>
<td>EB-Local</td>
<td>108.6240</td>
<td>EB-Global</td>
<td>108.6164</td>
</tr>
<tr>
<td>EB-Global</td>
<td>108.6164</td>
<td>ZS-Full</td>
<td>108.4286</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>108.7108</td>
<td>Hyper-$g$ ($a = 3$)</td>
<td>108.6254</td>
</tr>
<tr>
<td>Hyper-$g$ ($a = 3$)</td>
<td>108.6254</td>
<td>Hyper-$g$ ($a = 4$)</td>
<td>108.5802</td>
</tr>
<tr>
<td>Robust</td>
<td>108.7328</td>
<td>Least Squares</td>
<td>108.2285</td>
</tr>
<tr>
<td>BOPB Form 1</td>
<td>106.7322</td>
<td>BOPB Form 2</td>
<td>106.8235</td>
</tr>
<tr>
<td>BOPB Form 3</td>
<td>107.9949</td>
<td>BOPB Form 4</td>
<td>107.9304</td>
</tr>
<tr>
<td>BOPB Form 5</td>
<td>108.1448</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Table showing cross-validated predictive SSE under different methods. The SSE values for the Bayesian models are obtained by model averaging. All five forms of the BOPB hyper-$g/n$ prior have the hyperparameter value of $a = 3$.

An important point to note here is that the model space for the existing methods and that for the mixtures of the BOPB $g$ prior are not the same. Traditional priors are associated with a model space consisting of $2^p$ models, where any of the total $p$
covariates can be included or excluded in a model. The BOPB hyper-\(g/n\) prior has a model space of size \(e\Gamma(K + 1, 1)\), where any of the total \(K\) blocks can be included or excluded and given a particular combination of \(k\) blocks, all \(k!\) block permutations define distinct models. Uniform priors on these model spaces do not allocate prior probabilities to the models in the same way and so the effect of the model prior on prediction errors is worth investigating. Moreover, there is a big discrepancy in the sizes of the model spaces; the model space for the common priors has 8,192 models while the BOB hyper-\(g/n\) over block permutations has only \(e\Gamma(3, 1) = 5\) models for prior Forms 2, 3, 4 and 5 and \(e\Gamma(4, 1) = 16\) models for prior Form 1.

<table>
<thead>
<tr>
<th></th>
<th>BOPB Hyper-(g/n) ((a = 3))</th>
<th>Hyper-(g) ((a = 3))</th>
<th>Hyper-(g/n) ((a = 3))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form 1</td>
<td>106.7322</td>
<td>106.8158</td>
<td>106.8229</td>
</tr>
<tr>
<td>Form 2</td>
<td>106.8235</td>
<td>106.8158</td>
<td>106.8229</td>
</tr>
<tr>
<td>Form 3</td>
<td>107.9949</td>
<td>108.1508</td>
<td>108.1598</td>
</tr>
<tr>
<td>Form 4</td>
<td>107.9304</td>
<td>108.1508</td>
<td>108.1598</td>
</tr>
<tr>
<td>Form 5</td>
<td>108.1448</td>
<td>108.1523</td>
<td>108.1614</td>
</tr>
</tbody>
</table>

Table 4.2: Table showing cross-validated predictive SSE (under BMA) for the hyper-\(g\), hyper-\(g/n\) and BOPB hyper-\(g/n\) priors. The model averaging is done over all possible permutations and combinations of the set of design blocks appearing in the specific prior form.

Although different permutations of the same set of blocks lead to distinct priors in the BOPB hyper-\(g/n\) prior, the ordinary hyper-\(g\) and hyper-\(g/n\) priors are unaffected by block permutations and all permutations of a given combination of blocks result in an identical prior. Table 4.2 reveals the role of the model space in lowering predictive errors in the BOPB hyper-\(g/n\), hyper-\(g\) and hyper-\(g/n\) priors. The cross-validated prediction errors for the mixtures of ordinary \(g\) priors decrease slightly in the case of
the smaller model space and the prediction errors get closer to values corresponding to the new prior. However, with the exception of prior Form 2, the BOPB hyper-$g/n$ still furnishes lower predictive SSE values than the ordinary mixtures of $g$ priors when model averaging is carried out over the same curtailed model space for each prior.

The BOPB hyper-$g/n$ prior in Form 1 dominates all other priors in terms of predictive accuracy in the reduced as well as the complete model space. The degree of improvement associated with a BOPB hyper-$g/n$ prior is diminished when estimates are averaged over the smaller model space, hinting at a positive effect of the reduced set of models. But in the usual situation with a complete model space, as in a general implementation of a standard prior, the predictive errors are significantly higher compared to the new prior. Appropriate choice of group structure in the design is essential for good behavior of the BOPB hyper-$g/n$ prior, and an intelligent choice, like prior Form 1 in this example, enhances the performance of the model remarkably.

### 4.7 Summary

In this chapter, we modified the setting of the basic block $g$ prior and proposed a new version of the block $g$ prior on a reparameterized design with orthogonal design blocks. There are at least two motivating reasons to justify the introduction of the new block orthogonalized block (BOB) $g$ prior. The first is that the block orthogonal design simplifies expressions for posterior summaries and expedites posterior inference. The second reason is that some formal results related to the basic block $g$ priors in Chapter 3 depend on the assumption of a blockwise orthogonal regression design. These same results hold in the BOB $g$ prior without the extra assumption.
The novel BOB $g$ prior is imposed on a transformed block orthogonal design to begin with and so the extra assumption of block orthogonality needed for amenable posterior expressions becomes irrelevant. The BOB hyper-$g$ and hyper-$g/n$ priors possess desirable asymptotic consistency properties and are shown to avoid the CLP and ELS when the orthogonalization of blocks is carried out in the “correct” order. In practice, we recommend orthogonalizing the blocks of covariates according to their relative order of importance in the regression model, to avert the CLP and a new, unwanted phenomenon called CELS. The complication of ranking blocks accurately is avoided by a BOB hyper-$g$ (or hyper-$g/n$) prior over the set of all block permutations, called the BOPB hyper-$g$ (or hyper-$g/n$) prior. This prior considers every possible arrangement for block orthogonalization, including the “correct” arrangement and is formally demonstrated to focus its probability on the model with the “correct” arrangement of blocks. Models with properly ordered design blocks have the main contribution to the calculations for posterior summaries and consequently, the adverse effects of erroneous ordering only have a negligible impact on posterior inference.
Chapter 5: Science Driven Priors on the Strength of Linear Regression

In Chapter 1 we briefly discussed how specification of traditional priors on the regression coefficients $\beta$ in a linear model might not be meaningful in situations where they impart unnecessarily strong (and often inplausible) information about certain regions of the parameter space. This chapter demonstrates how some common, popular priors have unwanted implications on the strength of linear relationship in regression. Use of these “flawed” priors for model comparison and inference is often unwarranted and alternative prior distributions need to be developed to address their drawbacks. We propose a new class of science driven priors in this chapter as a viable solution to the deficiencies of the usual priors.

5.1 Implications of Standard Priors

Consider the regression model described by

$$ y = \beta^T x + \epsilon $$

where $\beta$ and $x$ are both $p \times 1$ vectors denoting the unknown regression parameters and the covariate vector respectively. The observed error $\epsilon$ is assumed to be normally distributed with variance $\sigma^2_\epsilon$. Further assume that $x \sim N(0, I_p)$ is independent of $\epsilon$.
which makes this model consistent with the joint distribution

$$(y, x)^T | \Sigma_{yx} \sim N\left(0, \Sigma_{yx} = \begin{pmatrix} \sigma_y^2 & \beta^T \\ \beta & I_p \end{pmatrix}\right)$$

where $\sigma_y^2 = \sigma^2 + \beta^T \beta$. In this particular setup we work with $E((y, x^T)^T) = 0$ and $Var(x) = I_p$, with the understanding that any regression model can be easily transformed into this setting. The prior on the error variance $\sigma^2$ is inverse-gamma, i.e., $\sigma^2 \sim IG(a_1, b_1)$, for some suitably chosen hyperparameters $a_1$ and $b_1$ such that $E(\sigma^2) = \frac{1}{b_1(a_1 - 1)}$ if $a_1 > 1$.

Let $\rho^2$, the proportion of variation in $y$ explained by the regression component $\beta^T x$ in the model, be the population analogue of the coefficient of determination $R^2$. Due to the independence assumption in the model, the total variation in $y$ can be decomposed as $Var(y) = Var(\beta^T x) + Var(\epsilon)$, also represented as $\sigma_y^2 = \sigma_x^2 + \sigma_\epsilon^2$, where $\sigma_x^2 = \beta^T \beta$. This leads to the following expression for $\rho^2$:

$$\rho^2 = \frac{\sigma_x^2}{\sigma_y^2} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\epsilon^2} = \frac{\beta^T \beta}{\beta^T \beta + \sigma_\epsilon^2}.$$

As discussed in Chapter 1, many commonly used priors for $\beta$ resemble

$$\beta | c, \Sigma, \sigma_\epsilon^2 \sim N(0, c\sigma_\epsilon^2 \Sigma). \quad (5.1)$$

The matrix $\Sigma$ may be $I_p$, $(X^T X)^{-1}$ or have some other form determined by the user. A closely-related prior is specified to be conditionally independent of the variance of the error term:

$$\beta | c, \Sigma, \sigma_\epsilon^2 \sim N(0, c\Sigma).$$

As mentioned in Chapter 1, $c$ may be fixed a priori or it may be given a prior distribution that results in a thick-tailed marginal prior for $\beta$. 140
It is instructive to study the behavior of the implied prior on $\rho^2$ under various default priors for $\beta$. Under prior (5.1) that conditions on $\sigma^2_\epsilon$, the common choice of $\Sigma = I_p$ results in

$$\beta^T \beta \mid \sigma^2_\epsilon \sim \text{Gamma}\left(\frac{p}{2}, 2c\sigma^2_\epsilon\right) \text{ (with mean } c\rho \sigma^2_\epsilon\text{)}$$

and

$$\frac{\rho^2}{1 - \rho^2} \mid \sigma^2_\epsilon \sim \text{Gamma}\left(\frac{p}{2}, 2c\right),$$

the latter of which is independent of $\sigma^2_\epsilon$. Then the implied prior density for $\rho^2$ is

$$\pi(\rho^2) = \frac{(2c)^{-p/2}}{\Gamma(p/2)} \left(\frac{\rho^2}{1 - \rho^2}\right)^{p/2 + 1} \rho^{-4} \exp\left(-\frac{\rho^2}{2c(1 - \rho^2)}\right), \quad 0 < \rho^2 < 1$$

for all choices of priors for $\sigma^2_\epsilon$.

The left panel of Figure 1.1 illustrates the behavior of this prior when $c = 1$ as the number of predictors ($p$) increases. Table 5.1 illustrates the behavior of the prior’s first two central moments (computed numerically). When $p = 1$, the prior is fairly flat over most of the range of $\rho^2$ and has an infinite spike at $\rho^2 = 0$. This behavior may be reasonable in a situation where there is no prior information available about the strength of the linear relationship between the predictor and the response and shrinkage of this parameter toward zero is desired. When $p = 2$, the prior density for $\rho^2$ is finite at zero and flat over most of its range, with a slight accumulation of prior mass near large values of $\rho^2$. As the number of predictors increases, the behavior changes dramatically: prior mass concentrates around one, effectively declaring that we are nearly certain \textit{a priori} that all variation in the response can be explained by the predictor variables. While we typically collect predictor variables because we think they will explain variation in the response, we rarely expect them to explain
almost all of the variation, even with a sizable number of “significant” predictors. (A notable exception is a carefully designed and controlled physical experiment where inputs to a system have a known relationship with the response and $\sigma^2_\epsilon$ is thought to predominantly represent measurement error.)

<table>
<thead>
<tr>
<th>Model Size</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(\rho^2)$</td>
<td>0.34</td>
<td>0.54</td>
<td>0.66</td>
<td>0.78</td>
<td>0.84</td>
</tr>
<tr>
<td>$SD(\rho^2)$</td>
<td>0.26</td>
<td>0.24</td>
<td>0.19</td>
<td>0.12</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Size</th>
<th>10</th>
<th>12</th>
<th>15</th>
<th>18</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E(\rho^2)$</td>
<td>0.89</td>
<td>0.91</td>
<td>0.93</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>$SD(\rho^2)$</td>
<td>0.05</td>
<td>0.04</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 5.1: Prior mean and standard deviation of $\rho^2$ as a function of model size when $\beta \sim N(0, \sigma^2_\epsilon I_p)$.

The following lemma provides a formal description of the absurd behavior of the prior density (5.2) on $\rho^2$ when the number of predictors ($p$) increases.

**Lemma 5.1.1.** Consider a prior distribution on $\rho^2$ with density of the form (5.2). The distribution on $\rho^2$ converges to a degenerate distribution with all of its mass at 1 as $p \to \infty$.

**Proof.** Recall that the prior in (5.2) is generated from the distribution

$$\frac{\rho^2}{1-\rho^2} \sim \Gamma \left( \frac{p}{2}, 2c \right),$$

and hence $\rho^2/(1-\rho^2) \overset{d}{=} Z$, where $Z$ is a Gamma($p/2, 2c$) random variable. This means that $\rho^2 \overset{d}{=} Z/(1+Z)$ because the map $g : t \mapsto \frac{t}{1+t}$ is continuous.

Since a gamma random variable exhibits the property of infinite divisibility, we can express $Z$ as $Z = \sum_{i=1}^{p} Z_i$, where the $Z_i$’s are i.i.d. Gamma($1/2, 2c$) random variables.
So,
\[ \rho^2 = \rho^2(p) \overset{d}{=} \frac{\sum_{i=1}^{p} Z_i}{1 + \sum_{i=1}^{p} Z_i} = \frac{\frac{1}{p} \sum_{i=1}^{p} Z_i}{\frac{1}{p} \sum_{i=1}^{p} Z_i} \cdot \]

By the Strong Law of Large Numbers, \( \frac{1}{p} \sum_{i=1}^{p} Z_i \rightarrow E(Z_1) = c \) almost surely as \( p \rightarrow \infty \). This implies that the RHS of the last expression converges almost surely to \( c/c = 1 \) as \( p \rightarrow \infty \) due to the continuous mapping theorem. As a result, the sequence of random variables \( \rho^2(p) \) converges in distribution to 1, or equivalently, to a degenerate distribution with mass at 1 when \( p \rightarrow \infty \). Note that almost sure convergence guarantees convergence in law for the RHS in the above display.

The behavior of the prior on \( \rho^2 \) as \( p \) grows large is driven by the behavior of \( \beta^T \beta \) in the same limit. In particular, the expected squared \( L_2 \) norm of the regression coefficients grows linearly in \( p \):
\[ E(\beta^T \beta \mid \sigma^2_\epsilon) = p \sigma^2_\epsilon. \]
It is tempting to ameliorate this effect by choosing \( c \) to scale with \( p \). One quick attempt at a fix is to assign \( c = \frac{1}{p} \) so that \( E(\beta^T \beta \mid \sigma^2_\epsilon) \) is constant across models of different sizes. Unfortunately, this specification forces the implied prior on \( \rho^2 \) to concentrate around \( \frac{1}{2} \) as \( p \rightarrow \infty \) as demonstrated in Figure 1.1. This follows from the fact that \( \text{Var}(\beta^T \beta \mid \sigma^2_\epsilon) = 2 \sigma^4_\epsilon/p \) causing the prior on \( \rho^2 = \beta^T \beta / (\beta^T \beta + \sigma^2_\epsilon) \) to converge to a degenerate distribution at \( \frac{1}{2} \) in the limit as \( p \rightarrow \infty \). We see that it is not possible to adjust the scale of the prior to maintain both a constant mean and a constant variance for \( \beta^T \beta \) as \( p \) increases, leading to the passing of unreasonably strong prior information to \( \rho^2 \) in the limit.

The same complication appears in other standard priors, like priors of the form \( \beta \sim N(0, cI_p) \) that do not condition on \( \sigma^2_\epsilon \). In this case, \( \beta^T \beta \sim \text{Gamma}(p/2, 2c) \) and the implied prior on \( \rho^2 \) now depends on the prior on \( \sigma^2_\epsilon \) because
\[ \frac{\rho^2}{1 - \rho^2} \mid \sigma^2_\epsilon \sim \text{Gamma}\left(\frac{p}{2}, \frac{2c}{\sigma^2_\epsilon}\right). \]
The traditional choice of a prior on $\sigma^2_\epsilon$ is the inverse-gamma prior, but it does not lead to a closed form expression for the implied marginal density function of $\rho^2$. The non-standard choice of a gamma prior on $\sigma^2_\epsilon$ induces a generalized three-parameter Beta distribution on $\rho^2$ which has a density available in closed form. When $\sigma^2_\epsilon \sim \text{Gamma}(a_1, b_1)$, the induced prior on $\rho^2$ is $\text{G3B}(p/2, a_1, b_1/2c)$ with density

$$
\pi(\rho^2) = \frac{\Gamma(p/2 + a_1) (2c)^{a_1} \rho^{p-1/2}(1 - \rho^2)^{a_1-1}}{\Gamma(p/2)\Gamma(a_1) b_1^{a_1} \left[ \frac{2c}{b_1} (1 - \rho^2) + \rho^2 \right]^{p/2+a_1}}, \quad 0 < \rho^2 < 1.
$$

Table 5.2 illustrates the effect of increasing $p$ on the marginal moments of $\rho^2$ under both the inverse-gamma and gamma priors on $\sigma^2_\epsilon$. For the illustration, we choose an inverse-gamma prior with parameters 3 and $\frac{1}{2}$ and a gamma prior with both parameters equal to 1, so that $E(\sigma^2_\epsilon) = 1$ and $\text{Var}(\sigma^2_\epsilon) = 1$ for both prior choices. As before, we use numerical integration to calculate the moments. The inflation in $\beta^T \beta$ as $p$ grows pushes the prior mass toward one as before. Rescaling the covariance matrix has the same effect again of failing to remedy the problem (results not shown).

<table>
<thead>
<tr>
<th>(A)</th>
<th>Model Size</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E(\rho^2)$</td>
<td>0.38</td>
<td>0.80</td>
<td>0.90</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>SD($\rho^2$)</td>
<td>0.29</td>
<td>0.15</td>
<td>0.08</td>
<td>0.06</td>
<td>0.04</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(B)</th>
<th>Model Size</th>
<th>1</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E(\rho^2)$</td>
<td>0.43</td>
<td>0.81</td>
<td>0.90</td>
<td>0.93</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>SD($\rho^2$)</td>
<td>0.32</td>
<td>0.17</td>
<td>0.09</td>
<td>0.07</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 5.2: Prior mean and standard deviation of $\rho^2$ as a function of model size when (A) $\beta \sim N(0, I_p)$ and $\sigma^2_\epsilon \sim \text{IG}(3, \frac{1}{2})$, and (B) $\beta \sim N(0, I_p)$ and $\sigma^2_\epsilon \sim \text{Gamma}(1, 1)$. 

144
Zellner’s $g$ prior (Zellner, 1986), which has a distribution of the form (5.1) with 
$\Sigma = (X^TX)^{-1}$ where $X$ is the $n \times p$ matrix of covariates, demonstrates similar behavior
in the implied prior on $\rho^2$. The prior on $\beta$ as well as the implied prior on $\rho^2$ now
depend on both $n$ and $p$. Several choices for the unknown parameter $c$ have been
considered in recent literature, and Section 3.1.1 briefly reviews some of the popular
methods of choosing $c$. Under our model assumptions, each row of $X$ is independently
distributed as $N(0, I_p)$. The unit information prior (Kass and Wasserman, 1995) sets
$c = n$, making this specification asymptotically equivalent to our earlier example
where $\beta \sim N(0, \sigma^2 \epsilon I_p)$, since $n(X^TX)^{-1} \rightarrow I_p$ as $n \rightarrow \infty$. The choice of $c = p^2$
(Foster and George, 1994) is used to calibrate the prior based on the risk inflation
criterion.

<table>
<thead>
<tr>
<th>Model Size</th>
<th>$E(\rho^2)$</th>
<th>$SD(\rho^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>0.13</td>
</tr>
<tr>
<td>10</td>
<td>0.48</td>
<td>0.15</td>
</tr>
<tr>
<td>15</td>
<td>0.71</td>
<td>0.13</td>
</tr>
<tr>
<td>20</td>
<td>0.95</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 5.3: Prior mean and standard deviation of $\rho^2$ as a function of model size when
$n = 20$ and (A) $\beta \sim N(0, \sigma^2 \epsilon (X^TX)^{-1})$, and (B) $\beta \sim N(0, p^2 \sigma^2 \epsilon (X^TX)^{-1})$.

Table 5.3 displays the marginal moments of $\rho^2$ for the two cases $c = 1$ and $c = p^2$
when $n = 20$, which again indicate the strange behavior of the induced prior on $\rho^2$.
As in the earlier example with prior form (5.1), scaling the variance in the $g$ prior by
$\sigma^2 \epsilon$ causes the distribution of $\rho^2$ to be invariant to the prior on $\sigma^2 \epsilon$. All the moments
have been calculated by Monte Carlo integration.
It is clear that the most commonly used standard priors on $\beta$ induce highly informative priors on $\rho^2$ for models with even a modest number of predictor variables. Disturbingly, the prior information concentrates in a region to which, in almost all cases, we would want to assign relatively little prior mass. While it may be tempting to not worry about this phenomenon—with the view that, with enough data, the prior will be dominated by the likelihood—we argue that the use of priors that behave in this way should be avoided for (at least) two reasons.

First, the purest descriptions of Bayesian methods emphasize their ability to coherently incorporate prior information into an analysis. It seems at odds with best modeling practices to knowingly include prior information that is orthogonal to actual belief, especially when alternative priors exist.

Second, while the likelihood may dominate the prior asymptotically with respect to targeted intra-model inferences, it is well known that prior-data conflict can have a substantial impact on the marginal likelihood. In turn, such conflict can strongly and adversely affect the comparison of models via the Bayes factor. These effects cascade to the posterior distribution on the model space, from which the weights for Bayesian model averaged prediction are derived. As there is typically uncertainty about which predictors belong in the model, the use of a prior that behaves sensibly with respect to $\rho^2$ is important.

The rest of this chapter is devoted to the development of prior distributions for $\beta$ for which the implied prior on $\rho^2$ is reasonable. Our approach first posits a prior for $\rho^2$ itself. This allows prior information about likely values of $\rho^2$ (with the sample $R^2$ statistic in mind) to be included in the model when it is available. It will also allow us to be non-informative about $\rho^2$ when appropriate. Given the prior on $\rho^2$,
we then construct a prior for the “rest” of $\beta$, with the goal of being non-informative about the aspects of $\beta$ for which we have little prior information. Computational considerations are addressed, and we investigate the performance of the prior on a substantive example.

5.2 Priors on $\rho^2$

An alternative strategy is to place a prior distribution directly on $\rho^2$ and then pass this information to the regression coefficients. Naturally the strongest form of prior information about $\rho^2$ that we could inject into the model is to place all of the prior mass on one particular value of $\rho^2$. Recall that $\beta$ and $\rho^2$ are related as $\beta^T \beta = \sigma^2 \rho^2 / (1 - \rho^2)$, which is equivalent to an equation of a $p$-dimensional hypersphere in $\beta$ when $\sigma^2$ and $\rho^2$ are fixed and known. Conditional on $\sigma^2$ and for a fixed $\rho$, the above relation restricts all of the prior mass for the regression vector $\beta$ on a $p$-dimensional hypersphere with radius $\sigma \rho / \sqrt{1 - \rho^2}$. In practical situations one cannot expect to have such strong prior information, but this nice structure can be used as a building block for constructing meaningful prior distributions. When we have problem specific prior information to pin down a tight range of values for $\rho^2$, a possible formulation of prior distribution would place most of its mass in a “ring” surrounding a circular region or hypersphere. Weaker prior information would correspond to a greater diffusion of probability mass away from the mode at the circular part.

Our preferred approach is to model $\rho^2$ directly with a parametric prior distribution. A natural choice is $\rho^2 \sim \text{Beta}(a, b)$ where the hyperparameters $a$ and $b$ are chosen to reflect prior information about $\rho^2$ when available and to be non-informative about $\rho^2$ otherwise. Define $\delta^2 = \beta^T \beta$, so that
\[
\rho^2 = \frac{\delta^2}{\delta^2 + \sigma^2}
\]
and
\[
\delta^2 = \sigma^2 \frac{\rho^2}{1 - \rho^2}.
\]

It follows that \(\frac{\rho^2}{1 - \rho^2} \sim \text{IB}(a, b)\), an inverted-beta (or beta-prime) distribution, which leads to
\[
\delta^2 \mid \sigma^2 \sim \text{GIB}(a, b, \sigma^2)
\]
with
\[
\pi(\delta^2 \mid \sigma^2) = \frac{\Gamma(a + b)}{(a\Gamma(b)} \sigma^2 \delta^{2a - 1} (\delta^2 + \sigma^2)^{-a-b}, \delta^2 \geq 0 \quad (5.3)
\]
where \(\text{GIB}(\cdot)\) stands for a generalized inverted-beta distribution. The generalized inverted-beta distribution on \(\delta^2\) induces the following prior on \(\delta = \sqrt{\beta^T \beta}\) when conditioned on \(\sigma^2\):
\[
\pi(\delta \mid \sigma^2) = \frac{2\sigma^2}{Be(a, b)} \delta^{2a - 1} (\sigma^2 + \delta^2)^{-a-b}, \delta \geq 0.
\]
where \(Be(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} dt = \Gamma(a)\Gamma(b)/\Gamma(a + b)\). The parameter \(\delta\) can be interpreted as the radius of the p-dimensional hypersphere or simply the distance of \(\beta\) from the origin.

It is interesting that specific choices of \(a\) and \(b\) in the prior for \(\rho^2\) result in well-known distributions on \(\delta\) when \(\sigma^2\) is known. For example, the particular choice of \(a = \frac{1}{2}\) and \(b = \frac{\nu}{2}\) induces the half \(t\)-distribution with the density
\[
\pi(\rho \mid \sigma^2) = \frac{2}{\sqrt{\pi^2}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{\delta^2}{\sigma^2}\right)^{-\left(\frac{\nu+1}{2}\right)}, \delta \geq 0.
\]

It is also possible to induce a half-Cauchy distribution [Gelman 2006; Polson and Scott 2012b] on \(\delta\)
\[
\pi(\delta \mid \sigma^2) = \frac{2\sigma^2}{\pi(\sigma^2 + \delta^2)}, \delta \geq 0.
\]
with a Beta\(\left(\frac{1}{2}, \frac{1}{2}\right)\) prior on \(\rho^2\). Such a prior choice for \(\rho^2\) is a bit unrealistic as it allocates substantial probability to the regression model being either really good or really bad, but it has close connections to the prior on the shrinkage factor in the \textit{horseshoe prior} \cite{Carvalho:2010}.

Another special case is when the prior on \(\rho^2\) is uniform over \((0, 1)\) which can be thought of as a non-informative prior on \(\rho^2\). When \(\rho^2 \sim \text{Beta}(1, 1)\),
\[
\pi(\delta^2 | \sigma^2) = \frac{\sigma^2}{(\sigma^2 + \delta^2)^2}, \delta^2 \geq 0,
\]
and
\[
\pi(\delta | \sigma^2) = \frac{2\sigma^2\delta}{(\sigma^2 + \delta^2)^2}, \delta \geq 0.
\]
The distribution on \(\delta^2\) is sometimes called the “log-logistic” distribution (because \(\log \delta^2\) follows a logistic distribution), which is popular in the survival analysis literature. The implied prior distribution on \(\delta\) is known as the Dagum distribution or the Burr Type III distribution. A Dagum distribution with parameters \((a, b, p)\) has a pdf of the form
\[
f(x; a, b, p) = \frac{ap}{x} \left(\frac{x}{\xi}\right)^{ap} \frac{1}{\left[\left(\frac{x}{\xi}\right)^a + 1\right]^{p+1}}, \quad x > 0.
\]
The induced prior on \(\delta\) in this case corresponds to a Dagum \((2, \sigma, 1)\) distribution.

### 5.3 Priors on \(\beta\)

It has been established in Section 5.2 that a Beta\((a, b)\) prior on \(\rho^2\) implies a generalized inverted-beta prior on the squared norm of the regression coefficients \((\delta^2 = \beta^T \beta)\) and the following distribution on \(\delta\) when \(\sigma^2\) is known
\[
\pi(\delta | \sigma^2) = \frac{2\sigma^2b}{Be(a, b)} \delta^{2a-1}(\sigma^2 + \delta^2)^{-(a+b)}, \delta \geq 0.
\]
In order to obtain a prior on \(\beta\), we need to make assumptions on how the total size \(\delta\) is distributed among the individual coefficients \(\beta_i\). One sensible assumption is that of
isotropy for the distribution of $\boldsymbol{\beta}$, so that given a $p$-dimensional hypersphere of radius $\delta$, the $\boldsymbol{\beta}$ coefficients are uniformly spread out over the $(p - 1)$-dimensional surface of the sphere. This is reasonable when a priori no single coefficient is expected to be more important than any of the other coefficients. The conditional prior is therefore,

$$
\pi(\beta | \delta) = \frac{\Gamma(p/2)}{2\pi^{p/2}\delta^{p-1}}, \quad \beta \in \{z \in \mathbb{R}^p : z^T z = \delta^2\},
$$

which leads to a prior $\pi(\beta | \sigma_i^2)$ where the elements of $\beta$ are exchangeable:

$$
\pi(\beta | \sigma_i^2) = \pi(\beta, \delta | \sigma_i^2) \quad \text{(since } \delta = \sqrt{\beta^T \beta} \text{ is fixed given } \beta) \\
= \pi(\beta | \delta) \pi(\delta | \sigma_i^2) \\
= \frac{\Gamma(p/2)}{B\epsilon(a, b)} \frac{\sigma_i^{2b}}{\pi^{p/2}} (\beta^T \beta)^{a-p/2} (\sigma_i^2 + \beta^T \beta)^{-(a+b)}, \quad \beta \in \mathbb{R}^p. 
$$

(5.4)

A detailed derivation of the density $\pi(\beta | \sigma_i^2)$ in (5.4) using the hyperspherical representation of the vector of regression coefficients $\beta$ is given in Appendix C.1. The prior density (5.4) on $\beta$ given $\sigma_i^2$ corresponds to a spherical distribution whose behavior depends primarily on the value of the hyperparameter $a$ relative to the number of predictor variables $p$. The three cases $a < p/2$, $a = p/2$ and $a > p/2$ are described separately below. Each case corresponds to a different behavior of the prior at the origin. In two cases, the prior can be represented as a scale mixture of normals (West, 1987). Such a representation of the distribution of regression coefficients in a normal linear model is well known to aid posterior computation. In this case, a data augmentation Gibbs sampler with closed-form (conditionally normal) updates for the regression coefficients can be constructed by including updates of the latent scale parameters. We describe such samplers in Section 5.5.2.
Representations and Properties of the Prior: \( a < p/2 \)

When \( a < p/2 \) that the prior density (5.4) can be represented as a normal scale mixture of the form

\[
\beta | v, \sigma^2_x \sim N(0, v\sigma^2_x I_p)
\]

\[
v \sim G_v(\cdot)
\]  

(5.5)

where \( G_v(\cdot) \) has the density

\[
g_v(v) = \frac{\Gamma(p/2) 2^{-b} v^{-(b+1)}}{Be(a, b) \Gamma(b + p/2)} \, _1F_1(a + b; b + p/2; -1/2v), \ v > 0
\]

and \( _1F_1(A, B, z) \) is the Kummer confluent hypergeometric function with \( B > A > 0 \). The derivation of the scale mixture representation is provided in Appendix C.2 and utilizes the multiplicative rule for Laplace transforms of function convolutions (suggested by Dr. Jim Berger, personal communication, October 20, 2012).

Prior (5.5) on the scale factor \( v \) can be further broken down by adding another step of hierarchy in the Bayesian model with the introduction of a new latent variable \( \lambda \)

\[
\beta | v, \sigma^2_x \sim N(0, v\sigma^2_x I_p)
\]

\[
v | \lambda \sim IG\left(b, \frac{2}{\lambda}\right)
\]

\[
\lambda \sim Beta\left(a, \frac{p}{2} - a\right).
\]  

(5.6)

This follows directly from the derivation of the scale mixture form (5.5) in Appendix C.2 where we show that \( v \sim G_v(\cdot) \) with density

\[
g_v(v) = \frac{\Gamma(p/2) 2^{-b} v^{-(b+1)}}{\Gamma(a) \Gamma(b) \Gamma(p/2 - a)} \int_0^1 (1 - y)^{p/2 - a - 1} e^{-y/2} y^{a+b-1} dy
\]

\[
= \int_0^1 \left[ \frac{\lambda^b}{\Gamma(b) 2^b} v^{-(b+1)} e^{-\lambda/2} \right] \lambda^{a-1} (1 - \lambda)^{p/2 - a - 1} d\lambda.
\]
The first part of the last display is the density of an inverse-gamma distribution and the second part is the density of a Beta distribution with the aforementioned parameters.

The hierarchical model structure in (5.6) can be partially collapsed by marginalizing over one of the latent variables and obtaining \( \pi(\beta \mid \lambda, \sigma^2, \epsilon) \) as
\[
\int_0^\infty \pi(\beta \mid v, \sigma^2, \epsilon) \pi(v \mid \lambda) \, dv.
\]
Then
\[
\beta \mid \lambda, \sigma^2 \sim t_{2b}\left(0, \frac{\lambda \sigma^2}{2b} I_p\right)
\]
\[
\lambda \sim \text{Beta}\left(\frac{p}{2} - a, \frac{p}{2}\right),
\]
where \( t_\nu(\mu, \Sigma) \) denotes a multivariate t-distribution with \( \nu \) degrees of freedom, location parameter \( \mu \) and scale matrix \( \Sigma \). A data augmentation Gibbs sampler may be applied for sampling from the posterior distribution with one of the mixture representations described by (5.5), (5.6) and (5.7).

When \( a < p/2 \), the joint density function is unbounded at the origin because
\[
\lim_{\|\beta\| \to 0} \pi(\beta \mid \sigma^2) = \infty.
\]
But the marginal density functions of the individual coefficients \( \beta_i \) have a finite, non-zero value at the origin. Though it is complicated to derive closed form expressions for the marginal densities of \( \beta_i, i = 1, \ldots, p \), evaluating any of these densities at 0 is not cumbersome when \( a > \frac{1}{2} \) since
\[
\pi_{\beta_i}(0 \mid \sigma^2) = \int_{\mathbb{R}^{p-1}} \frac{\Gamma(p/2) \sigma^2 b}{Be(a, b) \pi^{p/2}} (\beta^T_i \beta_{-i})^{a-p/2} (\sigma^2 + \beta^T_i \beta_{-i})^{-(a+b)} d\beta_{-i}
\]
\[
= \frac{\Gamma((p-1)/2) \sigma^2 b}{Be(a, b) \pi^{p/2}} \frac{Be(a^*, b^*) \pi^{(p-1)/2}}{\Gamma((p-1)/2) \sigma^{2b^*}}
\]
\[
\times \frac{\Gamma((p-1)/2) \sigma^2 b^*}{Be(a^*, b^*) \pi^{(p-1)/2}} (\beta^T_i \beta_{-i})^{a^*-(p-1)/2} (\sigma^2 + \beta^T_i \beta_{-i})^{-(a^*+b^*)} d\beta_{-i}
\]
\[
= \frac{\Gamma(p/2) \Gamma(a-1/2) \Gamma(b+1/2)}{\Gamma((p-1)/2) \Gamma(a) \Gamma(b) \sqrt{\pi \sigma^2}} > 0 \quad \text{(if } a > \frac{1}{2})
\]

152
where $\beta_{-i} = \{ \beta_k : k \neq i \}$, $a^* = a - \frac{1}{2}$ and $b^* = b + \frac{1}{2}$. The joint density $\pi(\beta \mid \sigma^2_\epsilon)$ is unimodal with the mode located at $\beta = 0$ because of the infinite spike.

**Representations and Properties of the Prior: $a = p/2$**

When $a = p/2$, $\beta \mid \sigma^2_\epsilon \sim t_{2b}(0, (\sigma^2_\epsilon/2b)I_p)$, a multivariate $t$ distribution with $2b$ degrees of freedom with density

$$
\pi(\beta \mid \sigma^2_\epsilon) = \frac{\Gamma(b + p/2)}{\Gamma(b)\sigma^p_\epsilon \pi^{p/2}} \left(1 + \frac{\beta^T \beta}{\sigma^2_\epsilon}\right)^{-(b+p/2)}, \quad \beta \in \mathbb{R}^p.
$$

The mode of this distribution is at $\beta = 0$ and the joint density function is finite at the mode, with the value

$$
\pi(0 \mid \sigma^2_\epsilon) = \frac{\Gamma(b + p/2)}{\Gamma(b)\sigma^p_\epsilon \pi^{p/2}} > 0.
$$

Each of the marginal distributions for $\beta_i, i = 1, 2, \ldots, p$ is a univariate $t$-distribution with finite density at the origin.

The multivariate $t$ distribution has a well-known representation as a scale mixture of normal distributions where the mixing distribution for the scale parameter is inverse-gamma:

$$
\beta \mid v, \sigma^2_\epsilon \sim N(0, v\sigma^2_\epsilon I_p),
$$

$$
v \sim IG(b, 2).
$$

**Representations and Properties of the Prior: $a > p/2$**

When $a > p/2$, the joint density function is zero at the origin and the mode of the distribution occurs along a hypersphere with radius $\delta = \sqrt{\beta^T \beta} = \sigma_\epsilon \sqrt{\frac{a-p/2}{b+p/2}}$. As before, we can show that the marginal density of each $\beta_i$ conditional on $\sigma^2_\epsilon$ is finite and non-zero at the origin provided $a > \frac{1}{2}$. But this is not necessarily a restriction
on the value of $a$ as $a > p/2 \geq 1/2$ in any regression model considered. The prior has no scale mixture of normals representation when $a > p/2$.

5.4 Modeling with a Science Driven Prior on $\rho^2$

5.4.1 Model Framework

Suppose we describe the basic linear regression model as

$$y = X\beta + \epsilon$$

where $X = (x_1, \ldots, x_p)$ is a $n \times p$ design matrix, and the $n \times 1$ vector $\epsilon$ is a vector of independent normal errors. This model is simply the matrix version of the regression setting that we considered in Sections 5.1 and 5.2 without the restriction that $\text{Var}(x_i) = 1$ for all $i = 1, 2, \ldots, p$. Without loss of generality, we first center and scale all variables so that every column of the observed matrix $X$ and the response $y$ have mean zero and variance 1. Due to this centering we remove the intercept term from the model. Further assume that the matrix $X$ is of full column rank, i.e., $\text{rank}(X_{n \times p}) = p$.

We work with a reparameterization of the original design $X$ so that all likelihood contours in the modified design $W$ (say) are spherical. The reparameterization can be done easily by the transformation $W = X(X^TX)^{-\frac{1}{2}}$ which converts the original regression parameters $\beta$ to the new set of parameters $\eta = (X^TX)^{\frac{1}{2}}\beta$ and the regression model to

$$y = W\eta + \epsilon$$

which has spherical likelihood contours in $\eta$ because $W^TW = I_p$. Recall that the relationship between $\rho^2$ and $\delta^2 = \beta^T\beta$ from the original model is $\rho^2 = \delta^2/(\delta^2 + \sigma^2)$.
(see Section 5.2), but it depends on the assumption of unit variance for all predictor variables. Following the transformation to the new regression design, we have \( \text{Var}(W) = \frac{1}{n-1} I_p \) and \( \eta^T \eta = \beta^T (X^T X) \beta \) which implies that \( \eta \) and \( \delta^2 \) are connected through the expression

\[ \rho^2 = \frac{\eta^T \eta}{\eta^T \eta + (n-1) \sigma^2}. \]

Henceforth, for notational convenience, we rename \( \eta^T \eta \) as \( \delta^2 \) so that now \( \delta \) represents the \( L_2 \) norm of the parameter \( \eta \). When \( \rho^2 \sim \text{Beta}(a, b) \), we have

\[ \delta^2 | \sigma^2 \sim \frac{(n-1) \sigma^2 \rho^2}{1 - \rho^2} | \sigma^2 \sim \text{GIB}(a, b, (n-1) \sigma^2). \]

As mentioned in Section 5.1, the model specification is completed with an inverse-gamma prior on the error variance \( \sigma^2 \), so that \( \sigma^2 \sim \text{IG}(a_1, b_1) \) for suitable hyperparameters \( a_1 \) and \( b_1 \). We call this version of science driven prior the \( R \)-prior.

### 5.4.2 Transformation to Hyperspherical Coordinates

The structure of the \( R \)-priors defined in Section 5.4.1 allows for the construction of efficient computational strategies for model fitting. To implement the sampling strategy we work with the spherical representation (for \( p \) dimensions) of the regression parameter \( \beta \) instead of the usual Euclidean form. Any vector \( \mathbf{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m \) has a representation in terms of its Euclidean norm \( r = \| \mathbf{x} \| \) and the \( (m-1) \) angles \( \mathbf{\theta} = (\theta_1, \ldots, \theta_{m-1})^T \) that uniquely defines the vector. The hyperspherical representation of \( \mathbf{x} \) is given by:

\[
\begin{align*}
x_1 &= r \cos(\theta_1) \\
x_2 &= r \sin(\theta_1) \cos(\theta_2) \\
x_3 &= r \sin(\theta_1) \sin(\theta_2) \cos(\theta_3)
\end{align*}
\]
\[ x_{m-1} = r \sin(\theta_1) \sin(\theta_2) \cdots \cos(\theta_{m-1}) \]
\[ \text{and, } \quad x_m = r \sin(\theta_1) \sin(\theta_2) \cdots \sin(\theta_{m-1}) \]

where the angles \( \theta_i \in [0, \pi] \), for \( i = 1, 2, \ldots, m-2 \) and \( \theta_{m-1} \in [0, 2\pi] \). We call the \( m \)-tuple \((r, \theta)\) the representation of the \( m \)-dimensional vector \( \mathbf{x} \) in the spherical coordinate system.

The \( p \)-dimensional parameter \( \boldsymbol{\eta} \) can thus be characterized in terms of spherical coordinates with \((p-1)\) angular components \( \phi_1, \phi_2, \ldots, \phi_{p-1} \) and a size component \( ||\boldsymbol{\eta}|| = \sqrt{\boldsymbol{\eta}^T \boldsymbol{\eta}} = \delta \) as described above. Conditional on \( \delta \) and \( \sigma^2_\epsilon \), the entire prior mass of \( \boldsymbol{\eta} \) lies on the surface of a \( p \)-dimensional hypersphere of radius \( \delta \). We specify that the transformed regression coefficient \( \boldsymbol{\eta} \) is uniformly distributed on the \((p-1)\)-dimensional surface. The surface area of a \( p \)-dimensional hypersphere of radius \( \delta \) is
\[ \frac{1}{\Gamma(p/2)} 2 \pi^{p/2} \delta^{p-1}, \]
so that the density of \( \boldsymbol{\eta} \) given \( \delta \) is
\[ \pi(\boldsymbol{\eta} | \delta) = \frac{\Gamma(p/2)}{2 \pi^{p/2} \delta^{p-1}}, \quad \boldsymbol{\eta} \in \{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^T \mathbf{z} = \delta^2\}. \quad (5.8) \]

The last assumption implies that the regression coefficients \( \eta_i, i = 1, 2, \ldots, p \), in the reparameterized model are identically distributed for a given \( \delta \), but they are not independent of each other because of the constraint on their total size. This generates a form of spherical symmetry in the \( \boldsymbol{\eta} \) coefficients lying uniformly on the surface of the hypersphere that makes the joint distribution of \( \boldsymbol{\eta} \) exchangeable. Prior (5.8) translates to the following induced prior on the angular components \( \boldsymbol{\phi} = (\phi_1, \ldots, \phi_{p-1})^T \) in the hyperspherical characterization of \( \boldsymbol{\eta} \) (see relation (C.1) in Appendix C.1).
\[ \pi(\phi | \delta^2) = \frac{\Gamma(p/2)}{2^{p/2} \pi^{p/2}} \prod_{j=1}^{p-2} \sin^j(\phi_{p-1-j}) \]

where \(0 \leq \phi_i \leq \pi\) for \(i = 1, 2, \ldots, p - 2\) and \(0 \leq \phi_{p-1} < 2\pi\). The form of the above prior indicates that all the \((p - 1)\) angles are independent of each other as well as independent of the parameter \(\delta^2\).

### 5.4.3 Rotation of the Coordinate Axes

The efficient sampling algorithm that we introduce later hinges on a further reparameterization of the model and the associated parameters. The goal of the transformation is to ensure that the least squares estimate of the parameter \(\eta, \hat{\eta}_{LS} = W^T y\), has only one non-zero Euclidean coordinate (the first component) while the rest are all zero. If we denote the hyperspherical representation of the vector \(\hat{\eta}_{LS}\) by \((r, \theta)\), this is equivalent to rotating the coordinate axes appropriately to make all angles \(\theta_i\) equal to zero in the new system. The idea of rotation is crucial to streamline sampling from the posterior distribution limited to the surface of a hypersphere. The transformation is implemented by rotating each axis with the objective of making the least squares estimate \(\hat{\eta}_{LS}\) point directly toward the top or the “north pole” of the \(p\)-dimensional hypersphere. In the rotated design, any spherical likelihood contour (a rotation preserves the spherical likelihood contours) and the surface of the hypersphere intersect at a fixed angle \(\theta^*_1\), the first angle in the spherical representation of the rotated regression parameter. Hence the \((p - 1)\)-dimensional sphere arising from the intersection of any spherical contour and the \(p\)-dimensional hypersphere of a specified radius is uniquely determined and characterized by the angle \(\theta^*_1\) only, and we denote this sphere as \(C_{\theta^*_1}\).
The steps undertaken to implement the rotation are detailed below.

(i) First we determine the least squares estimate $\hat{\eta}_{LS} = W^T y$ in the spherical
 system. This entails knowledge of $(p - 1)$ angular components $\theta$ and a size component $r = \sqrt{\hat{\eta}_{LS}^T \hat{\eta}_{LS}}$ for the vector $\hat{\eta}_{LS}$.

(ii) Construct rotation matrices $H_i$, $i = 1, 2, ..., p - 1$, where each matrix $H_i$ rotates
 the coordinate axes in one direction at a time, to align along the direction with
 $\hat{\eta}_{LS}$ at the "north pole". Specifically, for each $i \in \{1, 2, ..., p - 1\}$, we create the $i^{th}$ rotation matrix as

$$H_i = \begin{pmatrix}
1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\
0 & 1 & \cdots & \cdots & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & \cos(\theta_i) & \sin(\theta_i) & \cdots & 0 \\
0 & \cdots & -\sin(\theta_i) & \cos(\theta_i) & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \cdots & \cdots & \cdots & 0 & 1
\end{pmatrix}$$

(iii) Calculate the final rotation matrix $R$ as a combination of the individual rotation
 matrices $H_i$ as $R = \prod_{i=1}^{p-1} H_i$. Since any rotation matrix is orthogonal, we have
 $R^{-1} = R^T$.

(iv) After determining the complete rotation matrix $R$, we reparameterize the re-
 gression model as

$$y = W \eta + \epsilon = WR^T R \eta + \epsilon = V \tau + \epsilon$$

where $\tau = R \eta$ and $V = WR^T$.

Observe that in the new regression setting, $\hat{\tau}_{LS} = (V^T V)^{-1} V^T y = RW^T y = R \hat{\eta}_{LS}$,
a simple rotation of $\hat{\eta}_{LS}$ that constrains $\hat{\tau}_{LS}$ to be of the form $(||\hat{\tau}_{LS}||, 0, \ldots, 0)^T$.  

158
More significantly, \( \tau^T \tau = \eta^T R^T R \eta = \eta^T \eta = \delta^2 \), displaying that spherical symmetry is preserved in the new regression parameter \( \tau \). Let \((\delta, \phi)\) denote the parameter \( \tau \) in the spherical coordinate system from here on so that \( \phi_1, \phi_2, \ldots, \phi_{p-1} \) now relate to the parameter \( \tau \) (and not \( \eta \)).

### 5.5 Sampling from the \( R \)-Posterior

Posterior inference with the \( R \)-prior specification on a linear model is not straightforward and might be computationally intensive without the transformation, as the posterior distributions for the regression parameters are not available in closed form. Direct application of MCMC methods—for example, coordinate-wise Metropolis-Hastings updates—without the model reparameterization is possible, but the resulting chains may converge slowly. We suggest two methods for carrying out inference in the \( R \)-posterior. The first one involves executing a Gibbs sampler in a reparameterized design after rotating the coordinate axes. The motivation behind reparameterizing the model is to develop a sampling algorithm that is efficacious in terms of convergence, speed and storage. The second method is the application of the standard Gibbs sampling algorithm in an augmented model using the scale mixture representations derived in Section 5.3.

#### 5.5.1 Gibbs Sampling in the Reparameterized Design

For the linear model \( y = V \tau + \epsilon \) described in Section 5.4.3, the least squares estimator of \( \tau \) is of the form \( \hat{\tau}_{LS} = (\delta_0, 0, \ldots, 0)^T \) for some \( \delta_0 = ||\hat{\tau}_{LS}|| \) and hence the log-likelihood (given \( \sigma^2 \)) can be simplified as

\[
-\frac{1}{2\sigma^2}(y - V \tau)^T(y - V \tau) = -\frac{1}{2\sigma^2}\left[(y - V \hat{\tau}_{LS})^T(y - V \hat{\tau}_{LS}) + (\tau - \hat{\tau}_{LS})^T(\tau - \hat{\tau}_{LS})\right]
\]
since $V^T V = I_p$. Recall that $(\delta, \phi)$ denotes $\tau$ in the spherical coordinate system and so $\tau = (\delta \cos(\phi_1), \delta \sin(\phi_1) \cos(\phi_2), \ldots, \delta \sin(\phi_1) \sin(\phi_2) \cdots \sin(\phi_{p-1}))^T$. Thus,

$$(\tau - \hat{\tau}_{LS})^T (\tau - \hat{\tau}_{LS}) = \sum_{i=1}^{p} (\tau_i - \hat{\tau}_{i,LS})^2$$

$$= (\tau_1 - \delta_0)^2 + \sum_{i=2}^{p} (\tau_i - 0)^2$$

$$= [\delta \cos(\phi_1) - \delta_0]^2 + \delta^2 \sin^2(\phi_1)$$

$$= \delta^2 + \delta_0^2 - 2\delta_0 \delta \cos(\phi_1)$$

which implies that the likelihood in the transformed model depends only on one component ($\phi_1$) of $\phi$ when conditioned on $\delta$ and $\sigma^2$. After the rotation of the axes, only the full conditional of $\phi_1$ in the Gibbs sampler gets updated using information from the data, while the rest of the angles have full conditionals that are independent of the likelihood. This speeds up sampling because at each iteration of the Gibbs sampler, we need to draw from distributions for the three parameters $\sigma^2$, $\delta^2$ and $\phi_1$ only, irrespective of the number of predictors ($p$) in the model. The fixed dimensional parameterization not only helps to learn about the posterior distribution rapidly by discarding $(p - 2)$ parameters while performing the MCMC steps, but also promotes efficient storage of posterior samples in high dimensional problems with large values of $p$.

The transformed regression coefficients $\tau_i, i = 1, \ldots, p$, are functions of the hyperspherical parameters $\delta^2$ and $\phi$ so that $\tau = \tau(\phi, \delta)$. All the angles in $\phi$ range from 0 to $\pi$ apart from the last one which varies between 0 and $2\pi$. For computational simplicity, we take the range of all the angles to be the same, i.e., from 0 to $\pi$ and add a new random variable $\gamma \in \{-1, 1\}$ to the model that decides which half of the real axis $\phi_{p-1}$ lies in. Renaming the restricted version of the parameter $\phi_{p-1}$ as $\phi^*_{p-1}$,
we follow the convention that $\gamma = -1$ actually corresponds to $\phi_{p-1} = 2\pi - \phi^*_p$
while $\gamma = 1$ means $\phi_{p-1} = \phi^*_p$. Since the prior on $\phi_{p-1}$ is uniform on $(0, 2\pi)$, the
implied prior on $\gamma$ is discrete uniform on $\{-1, 1\}$. Let $\phi^* = (\phi_1, \ldots, \phi_{p-2}, \phi^*_p)^T$ and
it follows that $\tau = \tau(\phi^*, \delta, \gamma)$.

We sample from the posterior distributions of the unknown regression parameters
using the Gibbs sampler:

**Full conditional of $\delta^2$**

Standard Metropolis–Hastings (MH) algorithms are suitable to draw samples from
the full conditional of $\delta^2$.

$$
\pi(\delta^2 \mid y, \sigma^2, \gamma, \phi^*) \propto (\delta^2)^{a-1} \left[ (n - 1)\sigma^2 + \delta^2 \right]^{-(a+b)} 
\times \exp\left[ -\frac{1}{2\sigma^2} (y - V \tau)^T (y - V \tau) \right], \quad \delta^2 \geq 0.
$$

A random-walk MH algorithm is used to generate samples from the posterior of $\delta^2$.
At the $t^{th}$ iteration, a proposed value of $\delta^2$ is drawn from a distribution which is
uniform on $(\delta^2(t-1) - B, (\delta^2(t-1) + B)$, where $B$ is a predetermined step size. The
proposed positive value of $\delta^2$ is either accepted or rejected according to the usual MH
acceptance ratio. Negative proposed values of $\delta^2$ are rejected, as the MH acceptance
ratio is zero.

**Full conditional of $\sigma^2_\epsilon$**

Posterior samples of $\sigma^2_\epsilon$ can be drawn conveniently using the Adaptive Rejection
Sampling (ARS) method on a transformed variable. The full conditional of $\sigma^2_\epsilon$ is:

$$
\pi(\sigma^2_\epsilon \mid y, \delta^2, \gamma, \phi^*) \propto \left[ (n - 1)\sigma^2_\epsilon + \delta^2 \right]^{-(a+b)} (\sigma^2_\epsilon)^{b-(a_1+1)} 
\times \frac{1}{\sigma^2_\epsilon} \exp\left[ -\frac{1}{b_1\sigma^2_\epsilon} - \frac{1}{2\sigma^2_\epsilon} (y - V \tau)^T (y - V \tau) \right]
$$
\(\propto \left[(n - 1)\sigma^2_\epsilon + \delta^2\right]^{-(a+b)}\sigma^{-2a_1-n-2}_\epsilon \times \exp\left[-\frac{1}{\sigma^2_\epsilon} \left\{\frac{1}{b_1} + \frac{1}{2}(y - V\tau)^T(y - V\tau)\right\}\right], \quad \sigma^2_\epsilon > 0.\)

It is shown in Appendix C.3 that the full conditional of \(1/\sigma^2_\epsilon\) is log-concave when \(n > 2(b + 1) - 2a_1\). So for a sufficiently large sample size, we can use ARS to get a sample from the full conditional of \(1/\sigma^2_\epsilon\), and invert this value to get a sample from the full conditional of \(\sigma^2_\epsilon\).

**Full conditional of \(\phi_2, \phi_3, \ldots, \phi^*_p\)**

This posterior is most convenient to sample from because of the nice structure induced in the density by the axis rotations. When conditioned on \(\phi_1\), the full conditional is independent of the data as the spherical likelihood contours intersect the hypersphere along a circular region characterized by a fixed angle \(\phi_1\), over which the likelihood is constant throughout. If \(C_{\phi_1}\) denotes the \((p - 1)\)-dimensional circle generated at the intersection of a spherical contour and the \(p\)-dimensional hypersphere of radius \(\delta\), varying any of the angles \(\phi_2, \phi_3, \ldots, \phi^*_p\) has no effect on the likelihood over the region \(C_{\phi_1}\). So sampling from the posterior is equivalent to sampling from the independent prior distributions on each of the \((p - 2)\) angles:

\[
\begin{align*}
[\phi_2, \phi_3, \ldots, \phi^*_p \mid y, \sigma^2_\epsilon, \delta^2, \gamma, \phi_1] & \overset{d}{=} [\phi_2, \phi_3, \ldots, \phi^*_p \mid C_{\phi_1}, \gamma] \\
& \overset{d}{=} [\phi_2, \phi_3, \ldots, \phi^*_p \mid \gamma]
\end{align*}
\]

where

\[
\pi(\phi_2, \phi_3, \ldots, \phi^*_p \mid \gamma) = \frac{\Gamma\left(\frac{p-1}{2}\right)}{\pi^{(p-1)/2}} \sin^{p-3}(\phi_2) \sin^{p-4}(\phi_3) \cdots \sin(\phi_{p-2})
\]
over the parameter space \((0, \pi)^{p-2}\). Since \(\phi^*_p\) does not explicitly appear in the joint prior density, the prior distribution on \(\phi_{p-1}\) is uniform over \((0, \pi)\) when \(\gamma = 1\) and uniform over \((\pi, 2\pi)\) when \(\gamma = -1\).

**Full conditional of \(\phi_1\)**

Again the MH algorithm is used to draw from the full conditional of \(\phi_1\):

\[
\pi(\phi_1 \mid \mathbf{y}, \sigma^2, \delta^2, \gamma, \phi_2, \phi_3, \ldots, \phi^*_p) \propto \sin^{p-2}(\phi_1) \exp \left[ \frac{1}{2\sigma^2} (\mathbf{y} - V\tau)^T (\mathbf{y} - V\tau) \right],
\]

for \(0 \leq \phi_1 \leq \pi\).

We implement the Metropolis–Hastings algorithm with one of two different choices of proposal distributions for \(\phi_1\). But first \(\phi_1\) is scaled so that the range of the new parameter \(\omega = \phi_1/\pi\) is between 0 and 1.

(a) The first proposal distribution is based on the random-walk MH algorithm. We draw \(\omega^{(t)}\) from Uniform \((\omega^{(t-1)} - \Omega, \omega^{(t-1)} + \Omega)\), where \(\Omega (\leq 1)\) is a predetermined step size and \(\omega^{(t-1)}\) is the sample from the last (current) iteration. To avoid direct rejection of the \(\omega^{(t)}\) samples proposed outside the interval \((0,1)\), we use a wrapping-around strategy that ensures that all proposed \(\omega\) values are utilized in the Markov chain. For any \(0 < B < \Omega\), if \(\omega^{(t)} = -B\) is proposed, we immediately change the proposal to \(\omega^{(t)} = 1 - B\), which obviously lies within the proper parameter space \((0,1)\). Likewise, a proposal \(\omega^{(t)} = 1 + B\) is first converted to \(\omega^{(t)} = B\) before calculating the acceptance ratio. Proposals \(\omega^{(t)}\) are transformed back to \(\phi_1^{(t)} = \pi \omega^{(t)}\) which are then accepted or rejected depending on the MH acceptance probability.

(b) The second method involves generating a proposal \(\omega^{(t)}\) from a Beta \((\omega^{(t-1)} \lambda, (1 - \omega^{(t-1)}) \lambda)\) distribution which has mean \(\omega^{(t-1)}\) and variance \(\frac{\omega^{(t-1)}(1-\omega^{(t-1)})}{1+\lambda}\). The
value of $\lambda$ is chosen to be large so that proposals are close to the current iteration when $\omega^{(t-1)}$ is small (very close to zero) or large (very close to one). The large value of $\lambda$ stabilizes the acceptance rate for proposals when an iteration $\omega^{(t)}$ is close to either 0 or 1.

**Full conditional of $\gamma$**

The parameter $\gamma$ is used to determine the quadrant in which $\phi_{p-1}$ belongs. Conditional on all the angles $\phi$, the full conditional of $\gamma$ is the same as the prior on $\gamma$ which is discrete uniform over two values:

$$[\gamma \mid y, \sigma^2, \delta^2, \phi^*] \overset{d}{=} \text{Uniform} \{ -1, 1 \}.$$

An important observation here is that when sampling from the full conditionals of $\delta^2, \phi_1, \sigma^2_\epsilon$ and $\gamma$, the current value of the vector $\phi^*_{-1} = (\phi_2, \ldots, \phi^*_{p-1})^T$ does not affect the Markov chain in any way. So we fix $\phi^*_{-1}$ at the beginning and skip sampling from the full conditional of $\phi^*_{-1}$ in the Gibbs sampling steps. Once the Markov chain is deemed to have reached its stationary distribution and posterior samples of a sufficiently large size have been collected, it is straightforward to draw from the individual priors of $\phi_i, i = 1, 2, \ldots, p-1$, and get samples of $\tau$ by converting $((\delta^2)^{(t)}, \phi^{(t)})$ to the Euclidean form for each iteration $t$.

**5.5.2 The Data Augmentation Gibbs Sampler**

The innovative reparameterization of the model allows us to develop a useful and economical MCMC method. Nevertheless, when $a < p/2$, it is also possible to work with the original model and the prior (5.4) on $\beta$ using the scale mixture representations outlined in (5.5), (5.6) and (5.7). The case when $a = p/2$ can be handled
similarly using the normal scale mixture form of a multivariate t-distribution. Though it is possible to perform Gibbs sampling with the mixture representations (5.5) and (5.7), the full conditionals for some of the parameters in both representations are complicated distributions which are difficult to sample from. A Metropolis within Gibbs sampling algorithm can be implemented to sample from these complex distributions, but tuning proposal distributions (to achieve reasonable acceptance rates for the proposed samples) for the Metropolis step requires some effort. However, it is easy to sample from the remaining full conditionals which are standard distributions.

We suggest using the data augmentation Gibbs sampler on the regression model described by the scale mixture form in (5.6) when $a < p/2$, since the full conditionals for the latent variables in this form are simpler to manage. The full model is

$$
\begin{align*}
\mathbf{y} \mid \beta, \sigma^2_{\epsilon} & \sim N(X\beta, \sigma^2_{\epsilon}I_n) \\
\beta \mid v, \sigma^2_{\epsilon} & \sim N(0, v\sigma^2_{\epsilon}I_p) \\
v \mid \lambda & \sim IG\left(b, \frac{2}{\lambda}\right) \\
\lambda & \sim Beta\left(a, \frac{p}{2} - a\right) \\
\text{and, } \sigma^2_{\epsilon} & \sim IG(a_1, b_1).
\end{align*}
$$

The full conditional distributions for the Gibbs method are as follows:

(i) **Full Conditional of $\beta$, defined on $\mathbb{R}^p$:**

$$
\pi(\beta \mid \mathbf{y}, v, \lambda, \sigma^2_{\epsilon}) = N\left(\beta \mid (X^TX + \frac{1}{v}I_p)^{-1}X^T\mathbf{y}, (X^TX + \frac{1}{v}I_p)^{-1}\sigma^2_{\epsilon}\right)
$$

(ii) **Full Conditional of $v$, defined on $(0, \infty)$:**

$$
\pi(v \mid \mathbf{y}, \beta, \lambda, \sigma^2_{\epsilon}) = IG\left(v \mid b + \frac{p}{2}, 2(\lambda + \frac{\beta^T\beta}{\sigma^2_{\epsilon}})^{-1}\right)
$$
(iii) **Full Conditional of** $\sigma^2_\epsilon$, **defined on** $(0, \infty)$:

$$
\pi(\sigma^2_\epsilon \mid y, \beta, \lambda, v) = \text{IG} \left( \sigma^2_\epsilon \mid a_1 + \frac{n}{2} \left[ \frac{1}{b_1} + \frac{1}{2} (y - X\beta)^T (y - X\beta) \right]^{-1} \right)
$$

(iv) **Full Conditional of** $\lambda$:

$$
\pi(\lambda \mid y, \beta, v, \sigma^2_\epsilon) \propto \lambda^{a+b-1} (1 - \lambda)^{p/2-a-1} e^{-\lambda/2v}, \ 0 < \lambda < 1
$$

This full conditional corresponds to a confluent hypergeometric (CH) distribution **(Gordy 1998)**, which is a special case embedded within the class of compound confluent hypergeometric (CCH) distributions **(Gordy 1998, Armagan et al. 2011)**. Polson and Scott **(2009)** refer to the class of CCH distributions as the hypergeometric-beta (HB) distributions. The full conditional corresponds to a CH($a + b, p/2 - a, 1/2v$) distribution with the normalized density

$$
\pi(\lambda \mid y, \beta, v, \sigma^2_\epsilon) = \frac{\lambda^{a+b-1} (1 - \lambda)^{p/2-a-1} \exp(-\lambda/2v)}{Be(a + b, p/2 - a) \text{$_1$F$_1$(a + b, b + p/2, -1/2v)}}. \quad (5.9)
$$

Polson and Scott **(2009)** suggest utilizing a rejection sampling algorithm to draw samples from the HB distribution. Since the CH distribution is a special case of the HB/CCH distribution, the same sampling technique will work to obtain posterior samples of $\lambda$. They show that (5.9) is bounded by $M\lambda^{a+b-1} (1 - \lambda)^{p/2-a-1} / Be(a + b, p/2 - a)$, for an appropriately chosen finite constant $M > 0$.

So we can generate proposals of $\lambda$ from a Beta($a + b, p/2 - a$) distribution and use $M \times \text{Beta}(\lambda \mid a + b, p/2 - a)$ as the envelope for (5.9) in the rejection sampler.

When $a = p/2$, the prior on $\beta$ is a multivariate $t$-distribution which can be expressed as a normal scale mixture with an inverse-gamma mixing distribution. The full conditionals in this situation will either be normal or inverse-gamma and a fast, efficient Gibbs chain can be run effortlessly.
5.6 Model Uncertainty

The MCMC techniques discussed in the last two sections are based on sampling from a fixed design corresponding to a particular model. But the true data generating model is never known to the researcher beforehand in any real world problem and there is always some uncertainty in choosing the best model to explain the variation in the data. For estimation and prediction purposes, it is common in a Bayesian setting to average over posterior summaries from a range of models. The summary from each model is weighted appropriately to reflect the belief of the researcher and the ability of the given model to explain the observed data pattern. Bayesian model averaging (BMA) requires a prior distribution on the space of all possible models. Subjective elicitation of prior probabilities for all $2^p$ probable models becomes difficult for large or even moderately large problems and default priors are set on the model space in such cases. Let $\mathcal{M}$ denote the set of all $2^p$ models that can be constructed from all viable combinations of the $p$ candidate predictors. A popular default prior on $\mathcal{M}$ is the Bernoulli variable inclusion prior where a predictor variable is included or excluded from a model with a fixed probability $w$ (say) $\in (0, 1)$, independent of the other predictors. We follow this strategy in the remainder so that the prior probability of any model $M$ of size $k$ is $\pi(M) = w^k(1-w)^{p-k}$, $k \in \{0, 1, \ldots, p\}$. We note that other priors on model space can be easily accommodated (e.g., [Scott and Berger (2010)]).

The Gibbs sampler for the reparameterized model, discussed in the previous section, can be suitably modified in this situation by adding the model space to the existing parameter space and treating each model as a random entity. A similar strategy is adopted in the $MC^3$ algorithm of [Madigan and York (1995)] and the SSVS algorithm of [George and McCulloch (1993, 1997)]. The Markov chain now traverses
the set of models as well, visiting models with high posterior probabilities more often
than the ones with low posterior probabilities. However, inclusion of the model space
within the parameter space comes with additional implications as the Markov chain
now has a state space with varying dimensions. This happens because the number
of components in $\phi$ in the transformed design is a function of the size of the model.
Depending on whether a proposed model in the Gibbs sampler is of a higher or lower
dimension, the dimension of $\phi$ also increases or decreases simultaneously. To accom-
modate for this restriction, we design the sampling steps to allow a joint update of
the model and some of the angles together in the following manner.

**Full Conditional of $\delta^2$**

The full conditional of $\delta^2$ given a particular model is similar to the one described
in Section 5.5, the only difference now being that some of the parameters are model
specific:

$$\pi(\delta^2 | y, \sigma^2_\epsilon, \phi, M, \gamma) \propto (\delta^2)^{a-1} \left( (n-1)\sigma^2_\epsilon + \delta^2 \right)^{-(a+b)}$$

$$\times \exp \left[ -\frac{1}{2\sigma^2_\epsilon} (y - V_M \tau)^T (y - V_M \tau) \right], \quad \delta^2 \geq 0.$$  

where $V_M$ describes the rotated design matrix originating from a transformation of
the original design $X_M$ corresponding to predictors from the model $M$. The random-
walk Metropolis algorithm with proposals from a uniform distribution is again used
to draw samples from this density.

**Full Conditional of $\sigma^2_\epsilon$**

This full conditional is also identical to the one outlined in the preceding section,
but the density again depends on the given model $M$:  

168
\[ \pi(\sigma^2 \mid y, \delta^2, \phi, M, \gamma) \propto \left( (n-1)\sigma^2 + \delta^2 \right)^{-(a+b)} \sigma^{2b-2a_1-n-2} \times \exp \left[ -\frac{1}{\sigma^2} \left\{ \frac{1}{b_1} + \frac{1}{2} (y - V_M \tau)^T (y - V_M \tau) \right\} \right], \sigma^2 > 0. \]

**Full Conditional of \( \phi_1 \)**

We can use one of the two proposal distributions laid out in Section 5.5 within a random-walk Metropolis sampler to draw samples from the density

\[ \pi(\phi_1 \mid y, \sigma^2, \delta^2, \phi_{-1}, M, \gamma) \propto \sin^{\left| M \right|-2}(\phi_1) \exp \left[ -\frac{1}{2\sigma^2} (y - V_M \tau)^T (y - V_M \tau) \right] \]

defined on \( 0 \leq \phi_1 \leq \pi \), where \( |M| \) denotes the size of the given model \( M \).

**Full Conditional of \((M, \phi_{-1})\)**

This is the only full conditional which is inherently different from all the full conditionals appearing in Section 5.5. The joint update of \( M \) and \( \phi_{-1} \) is necessary due to the complication of differing dimensions. We factor the density as

\[ \pi(M, \phi_{-1} \mid y, \delta^2, \phi_1, \sigma^2, \gamma) \propto \pi(\phi_{-1} \mid M, y, \delta^2, \phi_1, \sigma^2, \gamma) \pi(M \mid y, \delta^2, \phi_1, \sigma^2, \gamma) \]

from which we sample with a Metropolis–Hastings algorithm described in Appendix C.4. The proposal distribution \( q((M, \phi_{-1}) \to (M^*, \phi^*_{-1})) \) to update the current iteration \((M, \phi_{-1})\) to the new values \((M^*, \phi^*_{-1})\) is also chosen so that it factors as \( q((M, \phi_{-1}) \to (M^*, \phi^*_{-1})) = p(M^* \mid M) p(\phi^*_{-1} \mid M^*) \). We choose a simple and symmetric proposal distribution \( p(M^* \mid M) = p(M \mid M^*) \) where a future model \( M^* \) is suggested by randomly adding or deleting a single explanatory variable from the current model \( M \), so the newly proposed model always differs from the current model by exactly one predictor. The proposal density \( p(\phi^*_{-1} \mid M^*) \) is chosen so that the
angles in \( \phi_{-1} \) are independent and each angle has a uniform proposal distribution on \((0, \pi)\).

**Full Conditional of \( \gamma \)**

The full conditional of \( \gamma \) (whenever the parameter exists in the model) for any arbitrary model \( M \) is discrete uniform over two possible values

\[
\left[ \gamma \mid y, \sigma^2_\epsilon, \delta^2, \phi, M \right] \overset{d}{=} \text{Unif}\{ -1, 1 \}.
\]

However, the combination of sampling steps mentioned above work only for models of size strictly greater than 2 \((p > 2)\) when the parameter \( \phi_{-1} \) is actually present in the model. For models of size \( p = 2 \), \( \phi_{-1} \) ceases to exist and we have to modify the algorithm slightly. We update from the full conditionals of \( \delta^2, \sigma^2_\epsilon, \phi_1 \) and \( \gamma \) as usual, but the last step of the Gibbs sampler deals only with the full conditional of a model (since \( \phi_{-1} \) does not appear in the model). So we have

\[
\pi(M \mid y, \delta^2, \phi_1, \sigma^2_\epsilon, \gamma) \propto p(y \mid M, \delta^2, \phi_1, \sigma^2_\epsilon, \gamma) \pi(M)
\]

and the MH acceptance probability for a proposed model \( M^* \) is

\[
\alpha = \min \left\{ 1, \frac{p(y \mid M^*, \delta^2, \phi_1, \sigma^2_\epsilon, \gamma) \pi(M^*)}{p(y \mid M, \delta^2, \phi_1, \sigma^2_\epsilon, \gamma) \pi(M)} \right\}
\]

when we use the same symmetric proposal density \( p(M^* \mid M) \) as before.

In any model of size \( p = 1 \), the entire parameter vector \( \phi \) is missing. We instead consider a new parameter \( I(\tau > 0) \) in the model because given a value of \( \delta^2, \tau \) can take either of two values: \( \delta \) or \(-\delta\). The parameter \( \gamma \) is no longer required and is also not included in the model. So the Gibbs sampler is set up to sample from the following full conditionals:

170
(i) \( \pi(\delta^2 \mid y, \sigma^2, M, I(\tau > 0)) \).

(ii) \( \pi(\sigma^2 \mid y, \delta^2, M, I(\tau > 0)) \).

(iii) \( \pi(I(\tau > 0) \mid y, \sigma^2, M, \delta^2) \).

(iv) \( \pi(M \mid y, \sigma^2, \delta^2, I(\tau > 0)) \).

If we wish to persist with the original parameter \( \phi (= \phi_1) \), the new parameter \( I(\tau > 0) \) can be replaced by a discrete distribution on \( \phi_1 \), with equal prior probability of 0.5 on the choices of \( \phi_1 = 0 \) and \( \phi_1 = \pi \). The MH acceptance probability for (iv) is calculated in the same way as in the case when \( p = 2 \).

For the unique model of size zero \( (p = 0) \) with no predictors, neither \( \phi \) nor \( \gamma \) shows up in the model and furthermore, the distribution of \( \delta \) becomes degenerate at 0. So the Gibbs sampler works with only two steps: (a) sampling from the full conditional of \( \sigma^2 \) keeping the value of \( \delta^2 \) fixed at 0 and (b) proposing a new model \( M \) and a new \( \delta^2 \) jointly which is accepted according to the MH acceptance ratio. Note that the proposal for \( M \) is always a model with exactly one predictor.

A unique characteristic of the Gibbs sampling algorithm discussed here is that migration from one non-null model to any other non-null model in the sampler takes place holding \( \rho^2 \) fixed in both models. Since \( \rho^2 \) is completely determined with the knowledge of \( \sigma^2 \) and \( \delta^2 \), the proposed model \( M^* \) from this full conditional will share the identical \( \rho^2 \) value as the current model \( M \). This is indeed a very sensible feature which is not common in prevalent MCMC methods. A model of higher (lower) dimension is usually favored (disfavored) by a goodness of fit criterion due to the presence (lack) of additional predictors that further improve (deteriorate) the model fit. This sampler cuts across models of different sizes in a much more meaningful manner as it
fixes the strength of linear model fit and allows the data to choose the more suitable covariates.

The scale mixture representation of $\beta$ in the case when $a \leq p/2$ can be extended similarly to develop a data augmentation Gibbs sampler that deals with model uncertainty. The usual Gibbs sampling steps for model-specific regression parameters need to be supplemented by a simultaneous update of the model in each iteration. Due to the recurring issue of uneven parameter dimensions across different models, the updates for $\beta$ and $M$ need to be carried out simultaneously.

5.7 Application to US Crime Data

In this section we analyze the performance of the $R$-prior on the US crime data set from 1960 and investigate the practical benefits of using this new form of science driven prior. We assess how the procedure functions under a variety of Beta prior specifications (by suitably adjusting the hyperparameters) on the strength of linear regression, with different priors concentrating most of the mass of $\rho^2$ on different regions of the unit interval.

5.7.1 US Crime Data

Originally the US crime data was analyzed by Isaac Ehrlich [Ehrlich 1973], where he scientifically analyzed the rates for crimes committed in 47 of the 50 U.S. states in the year 1960. The study stressed on the deterrent effects of punishment on seven FBI index crimes: murder, rape, robbery, assault, burglary, larceny and theft. The crime data set has 15 candidate predictor variables which were used by Ehrlich to examine the relationship between aggregate levels of punishment and crime rates. Most of the 15 predictor variables reported in the study were socio-economic variables like family
income, number of families earning below half of the median income, unemployment rates for urban males (per 1000 males) in the age groups 14-24 and 35-39, education level, labor force participation rate, number of young males and non-whites (per 1000) in the population, sex ratio, indicator for southern states and the state population size. The other variables included were per capita police expenditure for 1959 and 1960, average time served in state prisons and probability of imprisonment of the offenders. These variables were used to predict crime rates in different states, defined as the number of offenses reported to the police per hundred thousand population. Since 47 states were included in the study (excluded states are Hawaii, New Jersey and Alaska), the dataset has 47 cases and 15 predictor variables.

We first log transform all the predictors (apart from the single indicator variable) and then center and scale the variables to have mean zero and variance one. The classical linear model fit using all the 15 predictor variables, but without the intercept term (the Bayesian model in this analysis also does not include an intercept), leads to only 6 predictors: education level, number of young males (per 1000), unemployment rates for urban males in age group 35-39, number of non-whites (per 1000), probability of imprisonment and number of families below half the median income (per 1000) being significant at 5% level of significance. The sample $R^2$ value obtained from the full model equals 0.8688, and the adjusted $R^2$ equals 0.8073, so the data suggest that the true population correlation coefficient is on the higher side. The prior distribution on $\sigma^2$ is inverse-gamma, as mentioned in the model specification for the $R$-prior, with parameters $a_1$ and $b_1$ chosen meaningfully. The mean and variance of the inverse-gamma distribution are selected to be 0.6 and 0.4 respectively. Specifying the first two moments of this distribution uniquely determines the hyperparameters $a_1$ and $b_1$. 
The model space is defined by the inclusion or exclusion of each predictor variable and contains a total of $2^{15}$ models. We assume a Bernoulli$(w)$ variable inclusion prior on the model space so that each predictor variable is included in a model with prior probability $w$. We consider three choices of $w$ in this analysis: (i) $w = \frac{1}{2}$, (ii) $w = \frac{1}{3}$, and (iii) $w = \frac{1}{4}$.

Figure 5.1: Kernel density estimates of the posteriors of $\delta, \sigma^2$, and $\rho^2$ from the full model.
Model prior (i) corresponds to a uniform prior on the model space where all models are equally likely \textit{a priori}. Priors (ii) and (iii) place more mass on smaller models, and any predictor is more likely to be excluded than being included in a particular model. Big models are favored \textit{a posteriori} only when the data strongly favors their inclusion. We carry out posterior inference on the US crime data set using the novel and computationally efficient MCMC strategy on the reparameterized design described in Sections 5.4 and 5.6.

5.7.2 Effect of the Prior on $\rho^2$

The prior information on $\rho^2$ has a considerable effect on the posterior distributions of $\rho^2$, $\delta^2$ and $\sigma^2_\epsilon$. Figure 5.1 shows how the posterior distributions from the full model (model containing all 15 predictors) under $R$-priors vary with different choices of the hyperparameters $a$ and $b$ in the Beta prior for $\rho^2$. It is clear from Figure 5.1 that as the mean $\left(\frac{a}{a+b}\right)$ of the prior distribution increases (i.e., high values of $\rho^2$ are allocated more prior mass), the posterior density estimate of $\rho^2$ concentrates near the high sample $R^2$ value obtained from ordinary regression. When the strength of regression is assumed to be very weak in the prior, as in the case with $a = 2$ and $b = 6$, the posterior density estimate of $\rho^2$ becomes flatter and the mode shifts toward a lower value. The same effect is also noticed in the density estimate for $\delta$ which also becomes more diffuse when the prior on $\rho^2$ focuses around smaller values. Moderate strength of $\rho^2$ in the prior (when $a = 4$, $b = 4$) causes the posterior density of $\rho^2$ to shift midway between the other two cases. The posterior densities of $\sigma^2_\epsilon$ and $\delta$ are inversely related: concentration of posterior mass of $\delta$ around a higher value is accompanied by concentration of posterior mass of $\sigma^2_\epsilon$ near a smaller value. When
the prior on $\rho^2$ focuses more on large values, the $\sigma^2$ posterior gets pulled back a bit toward zero and the distribution becomes tighter and more peaked.

Figure 5.2: Kernel density estimates of the posterior of $\rho^2$ from the full model when the prior mean is fixed.

A researcher might choose to include different levels of prior information in the model through the Beta prior on $\rho^2$. Figures 5.2 and 5.3 illustrate how the posterior density of $\rho^2$ from the full model adapts to changes in the center and spread of the prior distribution. Observe that the variance of a Beta($a, b$) distribution is given by

\[
\frac{ab}{(a+b)^2(a+b+1)} = \frac{\mu(1-\mu)}{a+b+1},
\]

where $\mu = \frac{a}{a+b}$ is the mean of the same distribution. Hence for a given mean, increasing $(a + b)$ has the effect of lowering the variance of the Beta distribution. Figure 5.2 displays that when the prior mean is high, the spread of the prior distribution has very little effect on the posterior. But strong prior information
(corresponding to a large \((a + b)\) value) with a small mean drags the posterior toward the lower values and makes the density more diffuse. Figure 5.3 illustrates that the prior mean again pulls the center of the posterior toward itself, but the pull is much stronger in the case when \((a + b)\) is fixed at a large value. This is reasonable, since a higher \((a + b)\) value with the identical mean results in a tighter prior distribution with smaller variance.

![Figure 5.3](image)

Figure 5.3: Kernel density estimates of the posterior of \(\rho^2\) from the full model when \(a + b\) is fixed.

### 5.7.3 Cross-validated Predictive Performance

In this subsection we compare the predictive performances of existing Bayesian methods to that of the novel \(R\)-prior proposed in this chapter. The BAS package in R
has been used to obtain the prediction results for the standard priors popular in current statistical literature. We perform a twelve-fold cross-validation, dividing the 47 cases into eleven random groups of size 4 and one group of size 3. A variety of priors on the squared population correlation coefficient $\rho^2$ are considered, and the different Beta prior densities focus on distinct regions of the parameter space in different magnitudes. The Beta prior hyperparameters are selected to manifest many possible contrasting levels of belief on the suitability of the predictor variables in the regression model.

<table>
<thead>
<tr>
<th>Prior on $\beta$</th>
<th>MSE</th>
<th>Beta$(a, b)$ Prior on $\rho^2$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$-prior ($g = n$)</td>
<td>0.3623</td>
<td>$a = 1, b = 7$</td>
<td>0.3305</td>
</tr>
<tr>
<td>$g$-prior ($g = p^2$)</td>
<td>0.4055</td>
<td>$a = 2, b = 6$</td>
<td>0.3208</td>
</tr>
<tr>
<td>EB-Local</td>
<td>0.3516</td>
<td>$a = 4, b = 4$</td>
<td>0.3170</td>
</tr>
<tr>
<td>EB-Global</td>
<td>0.3513</td>
<td>$a = 6, b = 2$</td>
<td>0.3102</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>0.3540</td>
<td>$a = 2, b = 14$</td>
<td>0.3997</td>
</tr>
<tr>
<td>ZS-Full</td>
<td>0.3401</td>
<td>$a = 4, b = 12$</td>
<td>0.3590</td>
</tr>
<tr>
<td>Hyper-$g$ ($a = 3$)</td>
<td>0.3522</td>
<td>$a = 8, b = 8$</td>
<td>0.3296</td>
</tr>
<tr>
<td>Hyper-$g$ ($a = 4$)</td>
<td>0.3511</td>
<td>$a = 12, b = 4$</td>
<td>0.3108</td>
</tr>
</tbody>
</table>

Table 5.4: Cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{2}$.

The distributions on $\rho^2$ reflect prior faith ranging from extreme pessimism to utter optimism regarding the strength of the linear regression model in explaining the data variation. The centers (means) for the various prior choices are fixed at four distinct values representing different levels of confidence in the linear model: 0.125 (extreme pessimism), 0.25 (pessimism), 0.5 (moderate faith) and 0.75 (optimism). The same prior mean is assigned to a pair of priors in each case, but the variability within
a pair are disparate since the hyperparameter choices are different. Recall that a higher value of \((a + b)\) is associated with a lower prior variance and implies stronger prior information. Regardless of the prior distribution on the model space \((w = \frac{1}{2} \text{ or } \frac{1}{3} \text{ or } \frac{1}{4})\), an identical pattern is visible in each of the three tables. In Tables 5.4, 5.5 and 5.6 among all the \(g\)-priors and their mixtures, the full-based Zellner–Siow prior performs the best with respect to cross-validated predictive mean squared errors (MSE). Clearly the \(R\)-priors are doing a much better job at prediction in this example than all of the \(g\) priors when the prior information imparted on \(\rho^2\) is sensible. As long as the prior on \(\rho^2\) is not too strongly pessimistic (low center with little spread), the \(R\)-prior is superior in prediction under any choice of \(w\). Observe that even if the prior mean is low but the prior information is weak (large prior variance), the posterior can adapt sufficiently well to deliver satisfactory predictions. The failure of the \(R\)-posterior predictions in the case of pessimistic priors with small spread is not surprising, since a prior plays a significant role in determining the behavior of the posterior. Strong and inaccurate prior information invariably deteriorates the

<table>
<thead>
<tr>
<th>Prior on (\beta)</th>
<th>MSE</th>
<th>Beta((a, b)) Prior on (\rho^2)</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(g)-prior ((g = n))</td>
<td>0.4019</td>
<td>(a = 1, b = 7)</td>
<td>0.3530</td>
</tr>
<tr>
<td>(g)-prior ((g = p^2))</td>
<td>0.4452</td>
<td>(a = 2, b = 6)</td>
<td>0.3412</td>
</tr>
<tr>
<td>EB-Local</td>
<td>0.3882</td>
<td>(a = 4, b = 4)</td>
<td>0.3344</td>
</tr>
<tr>
<td>EB-Global</td>
<td>0.3875</td>
<td>(a = 6, b = 2)</td>
<td>0.3289</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>0.3924</td>
<td>(a = 2, b = 14)</td>
<td>0.4023</td>
</tr>
<tr>
<td>ZS-Full</td>
<td>0.3642</td>
<td>(a = 4, b = 12)</td>
<td>0.3673</td>
</tr>
<tr>
<td>Hyper-(g) ((a = 3))</td>
<td>0.3883</td>
<td>(a = 8, b = 8)</td>
<td>0.3368</td>
</tr>
<tr>
<td>Hyper-(g) ((a = 4))</td>
<td>0.3853</td>
<td>(a = 12, b = 4)</td>
<td>0.3259</td>
</tr>
</tbody>
</table>

Table 5.5: Cross-validated predictive MSE (under BMA) for different procedures when \(w = \frac{1}{3}\).
quality of any Bayes procedure. Nonetheless, an immense improvement in prediction is noticeable when the subjective prior on $\rho^2$ is not terribly inconsistent with the truth. The “optimistic” priors on $\rho^2$, and even the “moderately optimistic” priors, display tremendous enhancement in performance compared to the traditional priors in all three tables, which substantiates the potential of rational science driven prior elicitation.

<table>
<thead>
<tr>
<th>Prior on $\beta$</th>
<th>MSE</th>
<th>Beta($a, b$) Prior on $\rho^2$</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g$-prior ($g = n$)</td>
<td>0.4235</td>
<td>$a = 1, b = 7$</td>
<td>0.3604</td>
</tr>
<tr>
<td>$g$-prior ($g = p^2$)</td>
<td>0.4608</td>
<td>$a = 2, b = 6$</td>
<td>0.3556</td>
</tr>
<tr>
<td>EB-Local</td>
<td>0.4105</td>
<td>$a = 4, b = 4$</td>
<td>0.3411</td>
</tr>
<tr>
<td>EB-Global</td>
<td>0.4099</td>
<td>$a = 6, b = 2$</td>
<td>0.3386</td>
</tr>
<tr>
<td>ZS-Null</td>
<td>0.4149</td>
<td>$a = 2, b = 14$</td>
<td>0.4117</td>
</tr>
<tr>
<td>ZS-Full</td>
<td>0.3808</td>
<td>$a = 4, b = 12$</td>
<td>0.3760</td>
</tr>
<tr>
<td>Hyper-$g$ ($a = 3$)</td>
<td>0.4101</td>
<td>$a = 8, b = 8$</td>
<td>0.3427</td>
</tr>
<tr>
<td>Hyper-$g$ ($a = 4$)</td>
<td>0.4063</td>
<td>$a = 12, b = 4$</td>
<td>0.3377</td>
</tr>
</tbody>
</table>

Table 5.6: Cross-validated predictive MSE (under BMA) for different procedures when $w = \frac{1}{4}$.

5.8 Summary

In this chapter, we identify an undesirable modeling implication on the distribution of the population correlation coefficient $\rho$ when common priors are imposed on a linear model. We propose a version of a science driven prior called the $R$-prior, which rectifies the unwanted implication by directly modeling $\rho^2$ and apportions probabilities to the parameter space compatible with sensible prior belief. We suggest a novel MCMC method for inference based on a reparameterization of the model that scales well in high dimensional problems. The $R$-prior can also be represented as a normal
scale mixture in certain situations, which opens up the possibility of working with a data augmentation Gibbs sampler as another alternative for posterior inference. The new procedure demonstrates exemplary performance even when mild subjective prior information is passed on to the model through a reasonable distribution on $\rho^2$. 
Chapter 6: Discussion and Future Work

This dissertation has focused on the Bayesian linear model. In it, we have examined currently popular prior distributions and found them lacking from a theoretical perspective. This has led to the development of new classes of prior distributions that have preferable theoretical properties. These new priors have been shown to have better empirical performance as well. In addition, a class of science driven priors with appealing performance has been developed.

In Chapter 3 we introduced the block $g$ prior as a remedy to the undesirable behaviors displayed by many standard priors, including the popular hyper-$g$ prior of [Liang et al. (2008)]. A major factor behind the success of the ordinary hyper-$g$ prior is the availability of closed form marginal likelihoods for models and the resultant computational ease in exploring the posterior over the model space. The computational advantages of the standard mixtures of $g$ priors are lost with the block $g$ priors and we have to resort to Gibbs sampling, Monte Carlo integration or numerical integration of the partially-marginalized posterior distributions to make posterior inference. The number of blocks is highly influential in determining the computational complexity of the problem and the computing burden increases rapidly with the number of blocks. A blockwise orthogonal design simplifies the block hyper-$g$ posterior and allows faster computations, but the procedure still remains more complex than the hyper-$g$. 

182
A practical question is how to best select the groups or blocks of predictors in the design. Our analysis for the block $g$ priors so far have been based on identifying predictor variables related to one another through a latent or theoretical construct. Predictors measuring the same construct are placed in the same block, as they are more likely to have comparable coefficient sizes once the predictors are normalized. An identical scale parameter for these related explanatory variables would thus appear reasonable. In some situations, grouping the predictor variables based on the absolute values of the $t$ statistics in the least squares regression table (corresponding to $t$ tests) generates a data-dependent prior that performs satisfactorily in inference, as is evident in the data analysis example in Section 4.6. In the absence of any such knowledge, our preliminary empirical research suggests that placing correlated predictors in the same block leads to a better performance. There is scope for theoretical investigation as to why this choice works or if some other choice can be established as “optimal” under specific settings.

The conditional information asymptotic defined in Chapter 3 through the sequences \{\Psi_N\} in (3.5) and (3.10) is not the only form of such an asymptotic for which the relevant results in Chapter 3 hold. These sequences allow only one group of coefficients to grow in size, lending insight into a situation commonly found in practical data analysis where one set of coefficients is huge compared to the rest. The asymptotic can be easily modified to fit the situation where different groups of coefficients grow in size at different rates, similar to the situation in (4.19), and the theoretical results for traditional priors and for the block hyper-$g$ and block hyper-$g/n$ priors regarding the CLP and ELS remain valid. This modified form of the conditional information asymptotic applies to practical problems where more than one group of
predictors are believed to have big coefficients, but the sizes of the big coefficients are not comparable across groups.

Throughout Chapter 3, we have assumed that the design matrix $X_\gamma$ under model $M_\gamma$ is of full column rank, similar to the setup in Liang et al. (2008). Maruyama and George (2011) extended the definition for the ordinary $g$ prior to the $p > n - 1$ setting calling it the generalized $g$ prior. It would be interesting to see whether their approach can be suitably modified to formulate a generalized block $g$ prior. In a similar vein, while this dissertation focuses on block hyper-$g$ and block hyper-$g/n$ priors, it is clear that the benefits of these priors extend to other mixtures of $g$ priors.

The BOB $g$ prior introduced in Chapter 4 is a variation of the basic block $g$ prior, specified on a reparameterized design with orthogonal blocks. We recommend using the BOPB hyper-$g/n$ prior defined over all possible block permutations to avoid problems in modeling like the CLP and CELS, that arise as a consequence of orthogonalizing design blocks in an unfortunate order. But prior specification again depends on the choice of a block structure, and the same issue regarding the selection of “optimal” blocks of predictor variables poses a valuable question for future research.

Chapter 5 demonstrates how a science driven prior elicited on the strength of linear regression generates a sound Bayesian modeling procedure. A distribution on $\rho^2$ indirectly passes on information about the regression coefficients to the model and the subjective prior concentrates probabilities on important regions of the parameter space in a coherent manner. The $R$-prior described in Chapter 5 is one version of a science driven prior which models $\rho^2$; it is possible to come up with other versions of science driven priors that focus on different summaries of a regression model. It is worth exploring how other science driven priors perform in estimation and model
selection when the prior distributions are judiciously chosen to reflect a strong degree of awareness about the relevant parameters.

Empirical studies reveal the promising performance of the $R$-prior in finite samples, but a concrete theoretical justification of the asymptotic properties of the prior is still lacking. We believe that an inspection of the posterior concentration rate will shed further light on the effectiveness of the $R$-posterior. In an earlier version of their paper, Bhattacharya et al. (2014) discuss how many normal global-local scale mixture priors, including the Bayesian lasso prior, have suboptimal posterior contraction rates. A knowledge of the posterior contraction rate for the $R$-prior should be able to give a clear indication of the relative strength or weakness of the $R$-prior in comparison to other existing priors.
Appendix A: Appendix for Chapter 3

A.1 Proof of Theorem 3.2.1

The posterior mean of the regression coefficients is

$$\hat{\beta} = E\left( \frac{g}{1+g} \mid y \right) \hat{\beta}_{LS}$$

where $\hat{\beta}$ denotes the posterior mean of the regression coefficient $\beta$ and $\hat{\beta}_{LS}$ denotes the estimate of $\beta$ under least squares.

For the hyper-$g$ prior, the posterior expectation of the shrinkage factor can be expressed in terms of $R^2$ (see Liang et al. (2008))

$$E\left( \frac{g}{1+g} \mid y \right) = \frac{2}{p+a} \frac{2F_1\left(\frac{n-1}{2}, 2; \frac{p+a}{2} + 1; R^2\right)}{2F_1\left(\frac{n-1}{2}, 1; \frac{p+a}{2}; R^2\right)},$$

where $2F_1$ is the Gaussian Hypergeometric Function. $2F_1(a, b; c; z)$ is finite for $|z| < 1$ whenever $c > b > 0$. Here, $c - b = (p + a)/2 - 1 > 0$ since $2 < a \leq 4$ and $p > 0$. Thus, for all values of $R^2 < 1$, both numerator and denominator are finite. We use an integral representation of the $2F_1$ function.

$$E\left( \frac{g}{1+g} \mid y \right) = \frac{2}{p+a} \frac{2F_1\left(\frac{n-1}{2}, 2; \frac{p+a}{2} + 1; R^2\right)}{2F_1\left(\frac{n-1}{2}, 1; \frac{p+a}{2}; R^2\right)}$$

$$= \frac{\int_0^1 t(1-t)^{\frac{p+a}{2}-2}(1-tR^2)^{-\frac{n-1}{2}} dt}{\int_0^1 (1-t)^{\frac{p+a}{2}-2}(1-tR^2)^{-\frac{n-1}{2}} dt}.$$
Define \( m = \frac{n - 1}{2} \), \( b = \frac{n + a}{2} - 2 \) and \( z = R^2(\leq 1) \) so that we have

\[
E\left( \frac{g}{1 + g} \mid y \right) = \int_0^1 t(1-t)^b(1-tz)^{-m} dt.
\]

In our problem we have \( m > b > -\frac{1}{2} \) since \( a > 2 \) and \( p \geq 1 \) which satisfies the requirement of \( b > -1 \) required later in the proof.

\[
\text{Numerator} = \int_0^1 t(1-t)^b \left[ \sum_{k=0}^{\infty} \frac{(m+k-1)}{k} (tz)^k \right] dt
= \int_0^1 \sum_{k=0}^{\infty} \frac{(m+k-1)}{k} z^k t^{k+1} dt
= \sum_{k=0}^{\infty} \frac{(m+k-1)}{k} \int_0^1 t^{k+1} dt
= \sum_{k=0}^{\infty} \frac{(k+1)\Gamma(m+k)\Gamma(b+1)}{\Gamma(m)\Gamma(b+k+3)} z^k.
\]

(need \( b > -1 \))

Similarly we can show by interchanging the positions of the infinite sum and the integral that

\[
\text{Denominator} = \int_0^1 (1-t)^b \left[ \sum_{k=0}^{\infty} \frac{(m+k-1)}{k} (tz)^k \right] dt
= \sum_{k=0}^{\infty} \frac{(m+k-1)}{k} \int_0^1 t^{k+1} dt
= \sum_{k=0}^{\infty} \frac{\Gamma(m+k)\Gamma(b+1)}{\Gamma(m)\Gamma(b+k+2)} z^k.
\]

Thus,

\[
E\left( \frac{g}{1 + g} \mid y \right) = \frac{\sum_{k=0}^{\infty} \frac{(k+1)\Gamma(m+k)\Gamma(b+1)}{\Gamma(m)\Gamma(b+k+3)} z^k}{\sum_{k=0}^{\infty} \frac{\Gamma(m+k)\Gamma(b+1)}{\Gamma(m)\Gamma(b+k+2)} z^k} = \frac{\sum_{k=0}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \frac{1}{1+b+k+2} z^k}{\sum_{k=0}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k}.
\]

When \( m = \frac{n - 1}{2} > b + 2 = \frac{p + a}{2} \), we show that \( \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \) is increasing in \( k \). Consider the function \( D(k) = \log \Gamma(m+k) - \log \Gamma(b+k+2) \) which has the derivative, where
\( \Psi(\cdot) \) is the digamma function

\[
D'(k) = \Psi(m+k) - \Psi(b+k+2) = \int_0^\infty \left[ \frac{e^{-t}}{t} - \frac{e^{-(m+k)t}}{1 - e^{-t}} \right] dt - \int_0^\infty \left[ \frac{e^{-t}}{t} - \frac{e^{-(b+k+2)t}}{1 - e^{-t}} \right] dt = \int_0^\infty \left[ \frac{e^{-(b+k+2)t} - e^{-(m+k)t}}{1 - e^{-t}} \right] dt > 0
\]

\( \forall k \in \mathbb{N}, \) whenever \( m > b + 2. \)

This implies that \( D(k) \) is increasing in \( k \) and so is \( \exp(D(k)) \). The algebra above makes use of the following standard integral representation of the digamma function

\[
\Psi(x) = \int_0^\infty \left[ \frac{e^{-t}}{t} - \frac{e^{-tx}}{1 - e^{-t}} \right] dt, \quad \text{when } x > 0.
\]

Lemma 3.2.1 states that \( R^2 \to 1 \) as \( N \to \infty \) so that \( \lim_{N \to \infty} \hat{\beta} = \lim_{R^2 \to 1} E \left( \frac{\eta}{1+g} \mid y \right) \hat{\beta}_{LS}. \)

The proof will be complete if we can show that \( \lim_{R^2 \to 1} E \left( \frac{\eta}{1+g} \mid y \right) = 1 \). Throughout, we make use of the expression

\[
\lim_{R^2 \to 1} E \left( \frac{\eta}{1+g} \mid y \right) = \lim_{z \uparrow 1} \frac{\sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \frac{1}{1+z^{b+k+2}} z^k}{\sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k}.
\]

Case 1: \( n > p + a + 1 \)

First note that \( \frac{1}{1+z^{b+k+2}} \) is increasing in \( k \) and \( \uparrow 1 \) as \( k \to \infty \). So for any \( \eta > 0 \), \( \exists N_0 \) such that \( \forall k > N_0, \frac{1}{1+z^{b+k+2}} > 1 - \eta. \)

Hence, \( \lim_{z \uparrow 1} \frac{\sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \frac{1}{1+z^{b+k+2}} z^k}{\sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k} = \lim_{z \uparrow 1} \frac{\sum_{k=0}^{N_0} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \frac{1}{1+z^{b+k+2}} z^k + \sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \frac{1}{1+z^{b+k+2}} z^k}{\sum_{k=0}^{N_0} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k + \sum_{k=0}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k} \]

\[
\geq \lim_{z \uparrow 1} \frac{q_1 + (1 - \eta) \sum_{k=N_0+1}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k}{q_2 + \sum_{k=N_0+1}^\infty \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k}
\]

188
\[ (1 - \eta) + \lim_{z \uparrow 1} \frac{q_1 - (1 - \eta)q_2}{q_2 + \sum_{k=N_0+1}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k} \geq (1 - \eta) + 0 = (1 - \eta) \]

\[ \Rightarrow \lim_{R^2 \to 1} E \left( \frac{g}{1 + g} \mid y \right) = \lim_{z \uparrow 1} \frac{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)} \frac{1}{1 + \frac{b+k+1}{k+1}} z^k}{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)} z^k} = 1. \]

Note that \( q_1 \) and \( q_2 \) are finite numbers corresponding to the finite sums of the first \( N_0 \) terms. Also \( \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \to \infty \) as \( k \to \infty \) due to which \( \sum_{k=0}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \) and hence the denominator goes to infinity causing the second term above to vanish in the limit.

**Case 2:** \( p + a - 1 \leq n \leq p + a + 1 \)

Let \( n = p + a - 1 + 2\xi \), where \( 0 \leq \xi \leq 1 \) and define \( N_0 \) as in Case 1,

\[ E \left( \frac{g}{1 + g} \mid y \right) = \frac{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)} \frac{1}{1 + \frac{b+k+1}{k+1}} z^k}{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)} z^k} \]

\[ \lim_{R^2 \to 1} E \left( \frac{g}{1 + g} \mid y \right) = \lim_{z \uparrow 1} \frac{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)(b+k+1)} \frac{1}{1 + \frac{b+k+1}{k+1}} z^k}{\sum_{k=0}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+2)(b+k+1)} z^k}. \]

Proceeding as in Case 1, we can show that

\[ \lim_{R^2 \to 1} E \left( \frac{g}{1 + g} \mid y \right) > (1 - \eta) + \lim_{z \uparrow 1} \frac{q_1 - (1 - \eta)q_2}{q_2 + \sum_{k=N_0+1}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+1)(b+k+1)} z^k} \geq (1 - \eta), \text{ for any } \eta > 0. \]

As \( z \uparrow 1 \), the denominator of the second term becomes \( \frac{q_1 - (1 - \eta)q_2}{q_2 + \sum_{k=N_0+1}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+1)(b+k+1)} z^k} \) which tends to zero if the infinite sum \( \sum_{k=N_0+1}^{\infty} \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+1)(b+k+1)} \) diverges. It does, since

\[ \frac{\Gamma(b+k+1+\xi)}{\Gamma(b+k+1)(b+k+1)} = O(k^{-\lambda}), \text{ where } 0 \leq \lambda \leq 1 \text{ and } \sum_{k=N_0+1}^{\infty} O(k^{-\lambda}) = \infty \text{ for any } 0 \leq \lambda \leq 1. \text{ Thus,} \]

\[ 189 \]
\[ \lim_{R^2 \to 1} E \left( \frac{g}{1+g} \mid y \right) = 1. \]

Case 3: \( n < p + a - 1 \) (proving necessity of the constraint)

\[
\lim_{R^2 \to 1} E \left( \frac{g}{1+g} \mid y \right) = \lim_{R^2 \to 1} \frac{\int_0^1 t(1-t) \frac{p+a}{2} - 2(1-tR^2) \frac{n-1}{2} \, dt}{\int_0^1 (1-t) \frac{p+a}{2} - 2(1-tR^2) \frac{n-1}{2} \, dt} = \frac{\text{Beta}(2, \frac{p+a-n-1}{2})}{\text{Beta}(1, \frac{p+a-n-1}{2})} = \frac{2}{p+a+n+1}
\]

which is strictly less than 1 \( \forall \ n < p + a - 1 \) and has a minimum value of \( \frac{2}{p+a} \) when \( n = 1 \).

A.2 Proof of Theorem 3.2.2

Liang et al. (2008) show that

\[
BF(M_2 : M_0) = 2F_1 \left( \frac{n-1}{2}, 1; \frac{a+p}{2}; R_{M_2}^2 \right) \times \frac{a-2}{a+p-2}
\]

\[
BF(M_1 : M_0) = 2F_1 \left( \frac{n-1}{2}, 1; \frac{a+p_1}{2}; R_{M_1}^2 \right) \times \frac{a-2}{a+p_1-2}
\]

where \( R_{M_i}^2 \) is the coefficient of determination for model \( M_i, i = 1, 2 \).

The \( g \) prior is invariant to linear transformation of \( X \), and so we can work with an orthogonalized version of the design without loss of generality. Specifically, we consider \( Q_1 = X_1 \) and \( Q_2 = (I - P_{Q_1})X_2 \), where \( P_{Q_i} \) is the projection matrix for the column space of \( Q_1 \). Then \( X \) can be represented as \( X = QT \) for a suitable upper triangular matrix \( T \) and \( X\beta = Q\kappa \), where \( \kappa = T\beta \) also has a hyper-\( g \) prior.

Since \( T \) is upper triangular, \( ||\beta|| \to \infty \) is equivalent to \( ||\kappa|| \to \infty \) while \( \kappa_2 \) stays
fixed in the sequence. Under the block orthogonal setup, \( R_{M_1}^2 = (Q_1\tilde{\kappa}_1)^T(Q_1\tilde{\kappa}_1) \) and \( R_{M_2}^2 = R_{M_1}^2 + \frac{(Q_2\tilde{\kappa}_2)^T(Q_2\tilde{\kappa}_2)}{y^Ty} \). The term \( (Q_2\tilde{\kappa}_2)^T(Q_2\tilde{\kappa}_2) \) is constant throughout the sequence \( \{\Psi_N\} \) and so \( \frac{(Q_2\tilde{\kappa}_2)^T(Q_2\tilde{\kappa}_2)}{y^Ty} \to 0 \) as \( N \to \infty \).

\[
BF(M_2 : M_1) = \frac{BF(M_2 : M_0)}{BF(M_1 : M_0)} = \frac{a + p_1 - 2}{a + p - 2} \cdot \frac{\text{B}(n-\frac{1}{2}, 1; \frac{a+p}{2}, R_{M_2}^2)}{\text{B}(n-\frac{1}{2}, 1; \frac{a+p_1}{2}, R_{M_1}^2)}
\]

\[
= \frac{\int_0^1 (1 - t)^{\frac{a+p-2}{2}} (1 - tR_{M_2}^2)^{-\frac{n-1}{2}} dt}{\int_0^1 (1 - t)^{\frac{a+p_1-2}{2}} (1 - tR_{M_1}^2)^{-\frac{n-1}{2}} dt}
\]

Define \( b = \frac{a+p_1}{2} - 2, m = \frac{n-1}{2}, R_{M_1}^2 = z \) and \( R_{M_2}^2 = z + q \). When \( ||\beta_1|| \to \infty \), both \( R_{M_2}^2 \) and \( R_{M_1}^2 \) go to 1 which results in \( z \uparrow 1 \) and \( q \downarrow 0 \). Hence,

\[
BF(M_2 : M_1) = \frac{\int_0^1 (1 - t)^{b+\frac{p_2}{2}} [1 - t(z + q)]^{-m} dt}{\int_0^1 (1 - t)^{b} [1 - tz]^{-m} dt}.
\]

Proceeding as in Theorem 3.2.1

Numerator = \( \sum_{k=0}^{\infty} \frac{\Gamma(m + k)\Gamma(b + 1 + \frac{p_2}{2})}{\Gamma(m)\Gamma(b + k + 2 + \frac{p_2}{2})} (z + q)^k \)
and

Denominator = \( \sum_{k=0}^{\infty} \frac{\Gamma(m + k)\Gamma(b + 1)}{\Gamma(m)\Gamma(b + k + 2)} z^k \).

Thus,

\[
BF(M_2 : M_1) = \frac{\Gamma(b + 1 + \frac{p_2}{2})}{\Gamma(b + 1)} \sum_{k=0}^{\infty} \frac{\Gamma(m + k)}{\Gamma(b + k + 2)} \left\{ \frac{\Gamma(b + k + 2 + \frac{p_2}{2})}{\Gamma(b + k + 2 + \frac{p_2}{2})} \right\} (z + q)^k
\]

Hence \( \lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) = \lim_{z \to 1} \lim_{q \to 0} BF(M_2 : M_1) \)
The last step is justified when \( \lim_{q \to 0} BF(M_2 : M_1) \) exists for all \( 0 \leq z < 1 \). This holds since \( \lim_{q \to 0} BF(M_2 : M_1) = \frac{a + p_1 - 2}{a + p - 2} \frac{\mathcal{F}(\frac{m + 2 + \frac{p_2}{z}}{2})}{\mathcal{F}(\frac{m + 2}{2})} \) which exists and is finite for all \( 0 \leq z < 1 \).

\[
\lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) = \lim_{z \uparrow 1} \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \sum_{k=0}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \left\{ \frac{\Gamma(b+k+2)}{\Gamma(b+k+2+\frac{p_2}{z})} \right\} z^k.
\]

But \( \frac{\Gamma(b+k+2)}{\Gamma(b+k+2+\frac{p_2}{z})} \) decreases to 0 as \( k \to \infty \) (see Appendix A.1 for a proof). Hence given an arbitrary \( \eta > 0 \), we can find a number \( N_0 \) such that \( \forall k > N_0 \), \( \frac{\Gamma(b+k+2)}{\Gamma(b+k+2+\frac{p_2}{z})} < \eta \).

\[
\therefore BF(M_2 : M_1) = \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \times \left( q_1 + \eta \sum_{k=0}^{N_0} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \right) \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \times \left( q_2 + \sum_{k=N_0+1}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \right) z^k
\]

\[
= \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \frac{q_1 + \eta T}{q_2 + T}, \text{ with } T = \sum_{k=N_0+1}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k
\]

\[
= \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \left[ \frac{\eta (q_1 + T)}{q_2 + T} + \frac{(1 - \eta)q_1}{q_2 + T} \right].
\]

We later show that \( ||\beta_1|| \to \infty \) (or \( z \uparrow 1 \)) implies that \( T \to \infty \) when \( n \geq a + p_1 - 1 \).

\[
\implies \lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) \leq \lim_{T \to \infty} \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \left[ \frac{\eta (q_1 + T)}{q_2 + T} + \frac{(1 - \eta)q_1}{q_2 + T} \right]
\]

\[
= \eta \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)}.
\]

Hence \( \lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) \leq \eta \frac{\Gamma(b + 1 + \frac{p_2}{z})}{\Gamma(b + 1)} \), for any arbitrary choice of \( \eta > 0 \), and \( \lim_{||\beta_1|| \to \infty} BF(M_2 : M_1) = 0 \).

We now prove that, for \( n \geq a + p_1 - 1 \), \( T \to \infty \) when \( ||\beta_1|| \to \infty \).
Case 1: \( n > a + p_1 + 1 \)

Then \( m > b + 2 \) and so \( \frac{\Gamma(m+k)}{\Gamma(b+k+2)} \uparrow \infty \) as \( k \to \infty \).

\[
\lim_{z \uparrow 1} T = \sum_{k=N_0+1}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)}.
\]

So \( T \to \infty \) for such values of \( n \).

Case 2: \( a + p_1 - 1 \leq n \leq a + p_1 + 1 \)

Let \( n = a + p_1 - 1 + 2\xi \) where \( 0 \leq \xi \leq 1 \). Then \( m = \frac{a + p_1}{2} - 1 + \xi \) and \( \frac{\Gamma(m+k)}{\Gamma(b+k+2)} = \frac{\Gamma(\frac{a + p_1}{2} - 1 + \xi + k)}{\Gamma(\frac{a + p_1}{2} + k)} \). For \( 0 \leq \xi \leq 1 \), \( \frac{\Gamma(\frac{a + p_1}{2} - 1 + \xi + k)}{\Gamma(\frac{a + p_1}{2} + k)} = O(k^{-\lambda}) \) for some \( 0 \leq \lambda \leq 1 \). But \( \sum_{k=N_0+1}^{\infty} O(k^{-\lambda}) = \infty \) for such values of \( \lambda \) implying that \( T \to \infty \) as \( z \uparrow 1 \).

Case 3: \( n < a + p_1 - 1 \) (proving necessity of the constraint)

\[
\lim_{||\beta_1\to\infty||} \frac{BF(M_2 : M_1)}{\frac{\Gamma(m+k)}{\Gamma(b+k+2)}} = \frac{\int_0^1 (1 - t)^{\frac{a + p}{2} - 2 - \frac{n - 1}{2}} \frac{(1 - t R^2)}{2} dt}{\int_0^1 (1 - t)^{\frac{a + p}{2} - 2 - \frac{n - 1}{2}} \frac{(1 - t R^2)}{2} dt} = \frac{a + p_1 - n - 1}{a + p - n - 1} < 1.
\]

When \( n < a + p_1 - 1 \), the Bayes Factor \( BF(M_2 : M_1) \) is strictly less than 1 and still favors the smaller model \( M_1 \) in the limit, but not with overwhelming evidence as in the other two cases.

A.3 Proof of Corollary 3.2.1

It can be verified that

\[
\pi(\sigma^2, g \mid y) \propto (1 + g)^{-\frac{a + p}{2}} \frac{1}{\sigma^{n+1}} \exp \left[-\frac{1}{2\sigma^2} y^T (1 + g)^{-1} \frac{g}{P_X} y \right].
\]

So \( \pi(\sigma^2 \mid y) \propto \int_0^\infty (1 + g)^{-\frac{a + p}{2}} \frac{1}{\sigma^{n+1}} \exp \left[-\frac{||y||^2}{2\sigma^2} (1 - \frac{g}{1 + g} R^2) \right] \)
\[ \propto \frac{1}{\sigma^{n+1}} \int_{0}^{1} (1 - t)^{\frac{n+p}{2}-2} \exp \left[ -\frac{||y||^2}{2\sigma^2} \left\{(1 - R^2) + R^2(1 - t)\right\} \right] \, dt \]

\[ \propto \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{||y||^2(1 - R^2)}{2\sigma^2} \right] \int_{0}^{1} (1 - t)^{\frac{n+p}{2}-2} \exp \left[ -\frac{||y||^2R^2}{2\sigma^2} (1 - t) \right] \, dt. \]

Now \( ||y||^2(1 - R^2) = (n - p - 1)\sigma^2 \), which is fixed for all \( N \), so that

\[ \pi(\sigma^2 \mid y) \propto \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{(n-p-1)\sigma^2}{2\sigma^2} \right] \int_{0}^{1} x^{\frac{n+p}{2}-2} e^{-x} \, dx \propto (\sigma^2)^{\frac{n+p}{2}-1} \]

\[ \propto \frac{1}{(\sigma^2)^{(n+1)/2-(a+p)/2+1}} \exp \left[ -\frac{(n-p-1)\sigma^2}{2\sigma^2} \right] \int_{0}^{\infty} x^{\frac{n+p}{2}-2} e^{-x} \, dx. \]

As \( N \to \infty, ||y|| \to \infty \) and \( R^2 \to 1 \) so that \( \frac{||y||^2R^2}{2\sigma^2} \to \infty. \) (Lemma 3.2.1)

\[ \Rightarrow \lim_{N \to \infty} \pi(\sigma^2 \mid y) \propto \frac{1}{(\sigma^2)^{(n+1)/2-(a+p)/2+1}} \exp \left[ -\frac{(n-p-1)\sigma^2}{2\sigma^2} \right] \int_{0}^{\infty} x^{\frac{n+p}{2}-2} e^{-x} \, dx \]

\[ \propto \frac{1}{(\sigma^2)^{(n+1)/2-(a+p)/2+1}} \exp \left[ -\frac{(n-p-1)\sigma^2}{2\sigma^2} \right]. \]

Thus, the distribution of \( \lim_{N \to \infty} \pi(\sigma^2 \mid y) \) is inverse-gamma with shape parameter\( \frac{n+1-a-p}{2} \) and scale parameter \( \frac{1}{(n-p-1)\sigma^2} \). The mean of this distribution is \( \frac{(n-p-1)\sigma^2}{(n-p-a-1)}. \)

**A.4 Finding the Limit of (3.7) in Theorem 3.2.4**

Let \( P = \frac{I_1}{I_2} = \frac{\int_{0}^{1} t^{(p-1)/2} (B + t)^{-(n-1)/2} \, dt}{\int_{0}^{1} t^{(p_1-1)/2} (B^* + t)^{-(n-1)/2} \, dt}. \)

For any fixed \( B^* > 0 \) (\( z < 1 \)), \( I_2 \) is finite and so \( f_{B^*}(t) = \frac{1}{C_{B^*}} t^{(p_1-1)/2} (B^* + t)^{-(n-1)/2} \) is a proper density function with \( C_{B^*} = I_2 \) being the normalizing constant.

\[ P = \frac{1}{C_{B^*}} \int_{0}^{1} t^{(p-1)/2} \left[ \frac{p_1 + 1}{p + 1} B^* + t \right]^{-(n-1)/2} \, dt \]

\[ = E_{f_{B^*}} \left( t^{(p-p_1)/2} \left[ \frac{B^* + t}{p + 1} \right]^{(n-1)/2} \right) \]

\[ = P_{f_{B^*}}(t < \epsilon) E_{f_{B^*}} \left( t^{(p-p_1)/2} \left[ \frac{B^* + t}{p + 1} \right]^{(n-1)/2} \mid t < \epsilon \right) \]

194
$$+P_{fB^*}(t \geq \epsilon)E_{fB^*}\left(t^{(p-p_1)/2}\left[\frac{B^* + t}{p_{1+1}B^* + t}\right]^{(n-1)/2} \mid t \geq \epsilon\right) \quad \text{(for a small } \epsilon > 0)$$

$$< P_{fB^*}(t < \epsilon)\left[\epsilon^{(p-p_1)/2}\left(\frac{p + 1}{p_{1+1}}\right)^{(n-1)/2}\right] + P_{fB^*}(t \geq \epsilon)\left[\left(\frac{B^* + \epsilon}{p_{1+1}B^* + \epsilon}\right)^{(n-1)/2}\right].$$

The last inequality follows from the fact that $t^{(p-p_1)/2}$ is increasing in $t$ while $\frac{B^* + t}{p_{1+1}B^* + t}$ is decreasing in $t$ on the interval $(0,1)$. It can also be shown that for any integer $k$ so that $k > \frac{p}{p_1}, \frac{B^* + \epsilon}{p_{1+1}B^* + \epsilon} < k$ and hence,

$$P < P_{fB^*}(t < \epsilon)\left[\epsilon^{(p-p_1)/2}\left(\frac{p + 1}{p_{1+1}}\right)^{(n-1)/2}\right] + P_{fB^*}(t \geq \epsilon)k^{(n-1)/2}.$$

We will show that $P_{fB^*}(t < \epsilon) \to 1$ as $B^* \to 0$ $(N \to \infty)$ if $n > p_1 + 2$. For any arbitrary $0 < \epsilon < 1$,

$$\lim_{N \to \infty} P \leq \epsilon^{(p-p_1)/2}\left(\frac{p + 1}{p_{1+1}}\right)^{(n-1)/2}) + 0 = \epsilon^{(p-p_1)/2}\left(\frac{p + 1}{p_{1+1}}\right)^{(n-1)/2}.$$

It is possible to choose $\epsilon$ arbitrarily small, making the upper bound for the limit of $P$ arbitrarily small, and so $P \to 0$ as claimed in (3.7).

To show that $\lim_{N \to \infty} P_{fB^*}(t < \epsilon) = 1$ for any $0 < \epsilon < 1$, consider

$$\frac{P_{fB^*}(t < \epsilon)}{P_{fB^*}(t \geq \epsilon)} = \frac{\frac{1}{C_{B^*}}\int_0^\epsilon f_{B^*}(t)dt}{\frac{1}{C_{B^*}}\int_1^\epsilon f_{B^*}(t)dt} \geq \frac{\int_{\epsilon^j+1}^{\epsilon^{j+1}} t^{(p_1-1)/2}(B^* + t)^{-(n-1)/2}dt}{\int_1^\epsilon t^{(p_1-1)/2}(B^* + t)^{-(n-1)/2}dt} \quad \text{for some } j \geq 2$$

$$\geq \frac{(\epsilon^{j+1} - \epsilon^j)}{(1 - \epsilon) \sup_{\epsilon < t < 1} t^{(p_1-1)/2}(B^* + t)^{-(n-1)/2}}.$$

For $B^* < \frac{\epsilon^{j+1}(n-p_1)}{p_{1-1}}$, the above term equals $\frac{\epsilon^j \left(\frac{p_{1-1}}{\epsilon^{p_{1-1}}}(B^* + \epsilon)^{-(n-1)/2}\right)}{(1 - \epsilon) \sup_{\epsilon < t < 1} t^{(p_1-1)/2}(B^* + t)^{-(n-1)/2}}$, since the function $t^{(p_1-1)/2}(B^* + t)^{-(n-1)/2}$ is decreasing on $(\epsilon^{j+1}, 1)$.
So,

\[
\lim_{N \to \infty} \frac{P_{f_B^*}(t < \epsilon)}{1 - P_{f_B^*}(t < \epsilon)} \geq \epsilon^{(j-1)\frac{p_1-1}{2}} e^{-(j-1)(n-1)/2} \quad \text{(since } B^* \downarrow 0) \\
= \epsilon^{\frac{1}{2} (2j + (j-1)(p_1-n))}.
\]

This relation holds for any \( j \geq 2 \) and hence also for the limit as \( j \to \infty \). Since \( 0 < \epsilon < 1 \), \( \lim_{j \to \infty} \epsilon^{\frac{1}{2} (2j + (j-1)(p_1-n))} = \epsilon^{(n-p_1)/2} \lim_{j \to \infty} \epsilon^{\frac{1}{2} (2+p_1-n)} = \infty \) when \( n > p_1 + 2 \). This shows that \( \lim_{N \to \infty} P_{f_B^*}(t < \epsilon) = 1 \), for any \( 0 < \epsilon < 1 \) if \( n > p_1 + 2 \), completing the argument for the convergence of (3.7) to zero.

### A.5 Proof of Invariance of Block \( g \) Priors to Reparameterizations by Blockwise Affine Transformations

Suppose we have two separate regression problems:

**Problem 1:**

\[
\begin{align*}
\mathbf{y} \mid \alpha, \beta, \sigma^2 & \sim \mathcal{N}(\alpha \mathbf{1} + X_1 \beta_1 + \ldots + X_k \beta_k, \sigma^2 \mathbf{I}) \\
\beta \mid g, \sigma^2 & \sim \mathcal{N}(\mathbf{0}, \mathbf{A} \sigma^2) \\
\pi(\alpha, \sigma^2) & \propto \frac{1}{\sigma^2}
\end{align*}
\]

with \( \mathbf{A} = \begin{pmatrix}
g_1(X_1^T X_1)^{-1} & 0 & \ldots & 0 \\
0 & g_2(X_2^T X_2)^{-1} & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & g_k(X_k^T X_k)^{-1}
\end{pmatrix},
\]

and **Problem 2:**

\[
\begin{align*}
\mathbf{y} \mid \alpha, \gamma, \sigma^2 & \sim \mathcal{N}(\alpha \mathbf{1} + Z_1 \gamma_1 + \ldots + Z_k \gamma_k, \sigma^2 \mathbf{I}) \\
\gamma \mid g, \sigma^2 & \sim \mathcal{N}(\mathbf{0}, \mathbf{B} \sigma^2) \\
\pi(\alpha, \sigma^2) & \propto \frac{1}{\sigma^2}
\end{align*}
\]

196
We wish to show that if Problem 2 is a reparameterization of Problem 1 by a linear map acting within blocks, then inference is the same for both problems.

If the predictors in Problem 2 are a within-block linear transformation of the predictors in Problem 1, then there exist non-singular matrices $P_i$ such that $Z_i = X_i P_i$ for each $i = 1, 2, ..., k$. If we define $P = \begin{pmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{pmatrix}$, it is clear that $Z = (Z_1, Z_2, ..., Z_k)$ and $X = (X_1, X_2, ..., X_k)$ are related by $Z = XP$. So $y = \alpha_1 + Z\gamma + \epsilon$ can be rewritten as $y = \alpha_1 + XP\gamma + \epsilon$. For each $i$,

$$(Z_i^T Z_i)^{-1} = [(X_i P_i)^T (X_i P_i)]^{-1} = [P_i^T X_i^T X_i P_i]^{-1} = P_i^{-1} (X_i^T X_i)^{-1} (P_i^T)^{-1}.$$ 

It follows now that $B = P^{-1} A (P^T)^{-1}$ so that

$$\gamma \mid g, \sigma^2 \sim N(0, B\sigma^2) = N(0, P^{-1} A (P^T)^{-1} \sigma^2)$$

$$\implies P\gamma \mid g, \sigma^2 \sim N(0, P [P^{-1} A (P^T)^{-1} \sigma^2] P^T) = N(0, A \sigma^2) = \beta \mid g, \sigma^2.$$

The equivalence of these two priors signifies that inferences from Problem 1 and from Problem 2 will be exactly the same. Hence the block $g$ prior is invariant under a blockwise affine transformation of the problem in any general design.

**A.6 Proof of Lemma 3.4.2**

The proof relies on a result in, for example, [Lehmann and Romano (2005)] on stochastic ordering of random variables.
**UR: A Useful Result** (Stochastic ordering of densities/random variables)

Let $p_\theta(x)$ be a family of densities on the real line with monotone likelihood ratio in $x$. Then

1. For any $\theta < \theta'$, the cumulative distribution functions of $X$ under $\theta$ and $\theta'$ satisfy
   
   $F_{\theta'}(x) \leq F_{\theta}(x)$ for all $x$.

2. If $\psi$ is a non-decreasing function of $x$, then $E_\theta(\psi(X))$ is a non-decreasing function of $\theta$.

If $C_1$ and $C_2$ denote the normalizing constants for the two densities then

$$f_1(t_m) = \frac{1}{C_1} \int_{(0,1)} \left[ \prod_{i=1}^{k} (1-t_i)^{\frac{a+\pi_i}{2} - 2} \right] (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt_m$$

$$f_2(t_m) = \frac{1}{C_2} \int_{(0,1)} \left[ \prod_{i=1}^{k} (1-t_i)^{\frac{a+\pi_i}{2} - 2} \right] (1 - t_j R_j^2)^{-\frac{n-1}{2}} dt_m$$

So,

$$\frac{f_1(t_m)}{f_2(t_m)} = \frac{C_2 \int \prod_{i=1}^{k} (1-t_i)^{\frac{a+\pi_i}{2} - 2} (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt_m}{C_1 \int \prod_{i=1}^{k} (1-t_i)^{\frac{a+\pi_i}{2} - 2} (1 - t_j R_j^2)^{-\frac{n-1}{2}} dt_m}$$

**Case 1: $m \neq j$**

$$\frac{f_1(t_m)}{f_2(t_m)} = \frac{C_2 \int \prod_{i\neq m} (1-t_i)^{\frac{a+\pi_i}{2} - 2} (1 - \sum_{i\neq m} t_i R_i^2 - t_m R_m^2)^{-\frac{n-1}{2}} dt_m}{C_1 \int \prod_{i\neq m} (1-t_i)^{\frac{a+\pi_i}{2} - 2} (1 - t_j R_j^2)^{-\frac{n-1}{2}} dt_m}$$

Note that $(1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}}$ is a non-decreasing function of $t_m$, and so $\frac{f_1(t_m)}{f_2(t_m)}$ is a non-decreasing function of $t_m$.

Applying the Useful Result (UR) stated earlier, $f_1$ is stochastically larger than $f_2$ and, using part (2) of the UR with the strictly increasing function $\psi(t_m) = t_m$, we
have $E_{f_1}(t_m) \geq E_{f_2}(t_m)$. In fact, the ratio $\frac{f_1(t_m)}{f_2(t_m)}$ is strictly increasing and we have strict inequality in the result when $R_m^2 > 0$.

**Case 2: $m = j$**

\[
\frac{f_1(t_j)}{f_2(t_j)} = \frac{C_2}{C_1} \frac{(1-t_j)^{a+p_j-2} \int \prod_{i \neq j} (1-t_i)^{a+p_i-2} (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt_{-j}}{(1-t_j R_j^2)^{-\frac{n-1}{2}}} = \frac{C_2}{C_1} \frac{\int \prod_{i \neq j} (1-t_i)^{a+p_i-2} (1 - \sum_{i=1}^{k} t_i R_i^2)^{-\frac{n-1}{2}} dt_{-j}}{\int \prod_{i \neq j} (1-t_i)^{a+p_i-2} dt_{-j}}
\]

The function $(1 - \sum_{i \neq j} t_i R_i^2)^{-\frac{n-1}{2}}$ is non-decreasing in $t_j$ and so $\frac{f_1(t_j)}{f_2(t_j)}$ is a non-decreasing function of $t_j$. Using UR, we conclude that $E_{f_1}(t_j) \geq E_{f_2}(t_j)$. Strict inequality holds in this case when the ratio $\frac{f_1(t_m)}{f_2(t_m)}$ is strictly increasing and this happens when $R_m^2 > 0$ and $R_i^2 > 0$ for at least one $i \neq m$.

**A.7 Proof of Theorem 3.4.1**

The first part of the proof is trivial and follows directly from (3.11).

In the block orthogonal setup

\[
\pi(g \mid y) \propto \frac{\prod_{j=1}^{k} (1 + g_j)^{-\frac{a+p_j}{2}}}{[1 - \sum_{j=1}^{k} \frac{g_j}{g_j+1} R_j^2]^{(n-1)/2}}.
\]

So for any $i = 1, 2, \ldots, k$,

\[
E\left(\frac{g_i}{1+g_i} \mid y\right) = \frac{\int_{(0,1)^k} t_i \prod_{j=1}^{k} (1-t_j)^{a+p_j-2} (1 - \sum_{j=1}^{k} t_j R_j^2)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^{k} (1-t_j)^{a+p_j-2} (1 - \sum_{j=1}^{k} t_j R_j^2)^{-\frac{n-1}{2}} dt}
\]

\[
\geq \frac{\int_{(0,1)^k} t_i \prod_{j=1}^{k} (1-t_j)^{a+p_j-2} (1 - t_i R_i^2)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^{k} (1-t_j)^{a+p_j-2} (1 - t_i R_i^2)^{-\frac{n-1}{2}} dt}
\]

(by Lemma 3.4.2)
Hence \( \lim_{N \to \infty} E \left( \frac{g_i}{1 + g_i} \mid y \right) \geq \lim_{N \to \infty} \frac{2}{a + p_i} \frac{2F_1(\frac{n - 1}{2}, 2; \frac{a + p_i}{2} + 1; R_i^2)}{2F_1(\frac{n - 1}{2}, 1; \frac{a + p_i}{2}; R_i^2)} \).

As \( N \to \infty \), \( R_i^2 \to 1 \) so that for \( i = 1 \),

\[
\lim_{N \to \infty} E \left( \frac{g_1}{1 + g_1} \mid y \right) \geq \lim_{z \to 1} \frac{2}{a + p_1} \frac{2F_1(\frac{n - 1}{2}, 2; \frac{a + p_1}{2} + 1; z)}{2F_1(\frac{n - 1}{2}, 1; \frac{a + p_1}{2}; z)} = 1
\]

when \( n \geq a + p_1 - 1 \) (see Theorem 3.2.1). But \( E \left( \frac{g_1}{1 + g_1} \mid y \right) \leq 1 \), implying that \( E \left( \frac{g_1}{1 + g_1} \mid y \right) \to 1 \) in the limit.

For \( i > 1 \) and \( m \neq i \), \( E \left( \frac{g_i}{1 + g_i} \mid y \right) \)

\[
= \frac{\int_{(0,1)^k} t_i \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n - 1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n - 1}{2}} dt} \geq \frac{\int_{(0,1)^k} t_i \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - t_m R_m^2)^{-\frac{n - 1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - t_m R_m^2)^{-\frac{n - 1}{2}} dt} \quad \text{(by Lemma 3.4.2)}
\]

\[
= \frac{\text{Beta}(2, \frac{a + p_i}{2} - 1) \times 2F_1(\frac{n - 1}{2}, 1; \frac{a + p_m}{2}; R_m^2)}{\text{Beta}(1, \frac{a + p_i}{2} - 1) \times 2F_1(\frac{n - 1}{2}, 1; \frac{a + p_m}{2}; R_m^2)} = \frac{2}{a + p_i}.
\]

Thus, \( \lim_{N \to \infty} E \left( \frac{g_i}{1 + g_i} \mid y \right) \geq \frac{2}{a + p_i} \). Equality in the relation above is attained when \( R_i^2 = 0 \) or \( R_i^2 \to 0 \) and \( \lim_{N \to \infty} \sum_{j \neq i} R_j^2 < 1 \), i.e., when no linear combination of the predictors explains all of the variation in the response.

Again,

\[
E \left( \frac{g_i}{1 + g_i} \mid y \right) = \frac{\int_{(0,1)^k} t_i \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n - 1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n - 1}{2}} dt} \leq \frac{\int_{(0,1)^k} t_i \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j \neq i} R_j^2 - t_i R_i^2)^{-\frac{n - 1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{\frac{a + p_j}{2} - 2}(1 - \sum_{j \neq i} R_j^2 - t_i R_i^2)^{-\frac{n - 1}{2}} dt} \quad \text{(using a variation of Lemma 3.4.2)}
\]

200
A.8 Proof of Corollary 3.4.1

If \( y_\kappa \) where \( \kappa \) while very small values of \( y \), factor is near the lower bound

For this sequence of problems, \( 0 < \kappa < 1 \) is fixed for all \( i \neq 1 \), because

\[
\kappa_i = \frac{R_i^2}{1 - \sum_{j \neq i} R_j^2} = \frac{y^TP_Xy}{(n - p - 1)\hat{\sigma}^2 + y^TP_Xy}.
\]

If \( y^TP_Xy \) is quite large compared to \( \hat{\sigma}^2 \), then \( \kappa_i \approx 1 \) and the shrinkage factor is near 1 while very small values of \( y^TP_Xy \) relative to \( \hat{\sigma}^2 \) implies \( \kappa_i \approx 0 \) and the shrinkage factor is near the lower bound \( \frac{2}{a + p_i} \). For all values of \( 0 < \kappa_i < 1 \), we have

\[
\lim_{N \to \infty} E \left( \frac{g_i}{1 + g_i} \mid y \right) < 1, \text{ for } i \neq 1.
\]

A.8 Proof of Corollary 3.4.1

It is easy to show that

\[
\pi(\sigma^2, g \mid y) \propto \prod_{i=1}^{k} (1 + g_i)^{-\frac{a + p_i}{2}} \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{1}{2\sigma^2} y^T (I - \sum_{i=1}^{k} \frac{g_i}{1 + g_i} P_X_i) y \right].
\]

So, \( \pi(\sigma^2 \mid y) \)

\[
\propto \frac{1}{\sigma^{n+1}} \int_{(0, \infty)^k} \prod_{i=1}^{k} (1 + g_i)^{-\frac{a + p_i}{2}} \exp \left[ -\frac{\|y\|^2}{2\sigma^2} (1 - \sum_{i=1}^{k} \frac{g_i}{1 + g_i} R_i^2) \right] \, dg
\]

\[
\propto \frac{1}{\sigma^{n+1}} \int_{(0, \infty)^k} \prod_{i=1}^{k} (1 - t_i)^{-\frac{a + p_i}{2}} \exp \left[ -\frac{\|y\|^2}{2\sigma^2} (1 - \sum_{i=1}^{k} t_i R_i^2) \right] \, dt
\]

\[
\propto \frac{1}{\sigma^{n+1}} \int_{(0, \infty)^k} \prod_{i=1}^{k} (1 - t_i)^{-\frac{a + p_i}{2}} \exp \left[ -\frac{\|y\|^2}{2\sigma^2} (1 - \sum_{i=1}^{k} R_i^2 + \sum_{i=1}^{k} (1 - t_i) R_i^2) \right] \, dt
\]

201
\[ \alpha \frac{1}{\sigma^{n+1}} \exp \left[ - \frac{||y||^2 (1 - \sum_{i=1}^{k} R_i^2)}{2\sigma^2} \right] \]

\[ \times \prod_{i=1}^{k} \left[ \int_0^1 (1 - t_i) \frac{a_p}{2} e^{-t_i} \exp \left( - \frac{||y||^2 R_i^2 (1 - t_i)}{2\sigma^2} \right) dt_i \right] \]

\[ \alpha \frac{1}{\sigma^{n+1}} \exp \left[ - \frac{(n - p - 1)\hat{\sigma}^2}{2\sigma^2} \right] \prod_{i=1}^{k} \left[ \int_0^{\infty} x_i \frac{a_p}{2} e^{-x_i} dx_i \right] \prod_{i=1}^{k} \frac{\sigma^{a_p/2} - 1}{\sigma^{a_p/2}}. \]

Now as \( N \to \infty, ||y|| \to \infty \) and \( R_1^2 \to 1 \) while \( R_i^2 \to 0 \) \( \forall \ i \neq 1 \) (by Lemma 3.4.1).

The expression \( \frac{||y||^2 R_i^2}{2\sigma^2} = \frac{y^T p_x_i y}{2\sigma^2} = (x_i \hat{\beta})^T (x_i \hat{\beta}) \) is constant for \( i \neq 1 \) and goes to \( \infty \) for \( i = 1 \). So only \( \frac{||y||^2 R_i^2}{2\sigma^2} \to \infty \), while the other integrals are over a finite unchanging domain.

\[ \lim_{N \to \infty} \pi(\sigma^2 \mid y) \propto \frac{1}{(\sigma^2)^{\frac{n+1+k}{2} + \frac{\sum_{i=1}^{k} p_i}{2}}} \exp \left[ - \frac{(n - p - 1)\hat{\sigma}^2}{2\sigma^2} \right] \prod_{i=2}^{k} \left[ \int_0^{\infty} x_i \frac{a_p}{2} e^{-x_i} dx_i \right] \]

\[ \alpha \frac{1}{(\sigma^2)^{\frac{n+1+k}{2} + \frac{\sum_{i=1}^{k} p_i}{2}}} \exp \left[ - \frac{(n - p - 1)\hat{\sigma}^2}{2\sigma^2} \right] \prod_{i=2}^{k} \left[ \int_0^{\infty} x_i \frac{a_p}{2} e^{-x_i} dx_i \right] \]

where \( \gamma(\cdot, \cdot) \) is the lower incomplete gamma function.

The normalizing constant in the density \( \pi(\sigma^2 \mid y) \) exists for any \( N \) since

\[ \frac{1}{(\sigma^2)^{n-1-k(a-2)-p/2 + 1}} \exp \left[ - \frac{(n - p - 1)\hat{\sigma}^2}{2\sigma^2} \right] \prod_{i=2}^{k} \left[ \int_0^{\infty} x_i \frac{a_p}{2} e^{-x_i} dx_i \right] \]

\[ \leq \frac{1}{(\sigma^2)^{n-1-k(a-2)-p/2 + 1}} \exp \left[ - \frac{(n - p - 1)\hat{\sigma}^2}{2\sigma^2} \right] \prod_{i=2}^{k} \left[ \int_0^{\infty} x_i \frac{a_p}{2} e^{-x_i} dx_i \right] \]

which is integrable as a function of \( \sigma^2 \) over \((0, \infty)\) if \( n > k(a - 2) + p + 1 \). The integral (over \( \sigma^2 \)) for the expression above is also clearly bounded away from zero as long as \( ||y||^2 R_i^2 \) does not equal zero or converge to zero for any \( i = 2, \ldots, k \). But for
any such \( i, \| y \|^2 R_i^2 = y^T P_x y \) is strictly greater than zero with probability one in every element of the sequence \( \{ \psi_N \} \) (and also in the limit). This guarantees that the normalizing constant in \( \pi(\sigma^2 | y) \) is finite and non-zero for all \( N \) and validates the existence of a proper limiting distribution for the sequence of posteriors of \( \sigma^2 \).

A.9 Proof of Corollary 3.4.2

Proceeding as in Corollary 3.4.1,

\[
\pi(\sigma^2 | y) \propto \frac{1}{\sigma^{n+1}} \int_{(0, \infty)^k} \prod_{i=1}^k (1 + g_i)^{-\frac{a + p_i}{2}} \exp \left[ -\frac{\| y \|^2}{2 \sigma^2} (1 - \sum_{i=1}^k g_i R_i^2) \right] \, dg.
\]

So, \( E(\sigma^2 | y) \)

\[
= \int_{(0,1)^k} \int_0^\infty \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{\| y \|^2}{2 \sigma^2} (1 - \sum_{i=1}^k t_i R_i^2) \right] \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \, d\sigma^2 \, dt
\]

\[
= \int_{(0,1)^k} \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \int_0^\infty \frac{1}{\sigma^{n+1}} \exp \left[ -\frac{\| y \|^2}{2 \sigma^2} (1 - \sum_{i=1}^k t_i R_i^2) \right] \, d\sigma^2 \, dt
\]

\[
= \frac{\| y \|^2}{n - 3} \int_{(0,1)^k} \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \left( 1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-3}{2}} \, dt.
\]

This leads to

\[
E(\sigma^2 | y) \leq \frac{\| y \|^2 \int_{(0,1)^k} \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \left( 1 - t_i R_i^2 \right)^{-\frac{n-3}{2}} \, dt}{n - 3 \int_{(0,1)^k} \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \left( 1 - t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt} \quad (A.1)
\]

\[
= \frac{\| y \|^2}{n - 3} \frac{2 F_1 \left( \frac{n-3}{2}, 1; \frac{a + p_i}{2}; R_i^2 \right)}{2 F_1 \left( \frac{n-1}{2}, 1; \frac{a + p_i}{2}; R_i^2 \right)}.
\]

To show that (A.1) holds, define the pdfs \( f_1 \) and \( f_2 \) as

\[
f_1(t_m) \propto \int_{(0,1)^{k-1}} \left[ \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \right] \left( 1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt_{-m}
\]

and \( f_2(t_m) \propto \int_{(0,1)^{k-1}} \left[ \prod_{i=1}^k (1 - t_i)^{-\frac{a + p_i}{2} - 2} \right] \left( 1 - t_1 R_1^2 \right)^{-\frac{n-1}{2}} \, dt_{-m}
\]
and use Lemma 3.4.2 to obtain the result

\[ E_{f_1}(1 - \sum_{i=1}^{k} t_i R_i^2) = 1 - \sum_{i=1}^{k} R_i^2 E_{f_1}(t_i) \leq 1 - \sum_{i=1}^{k} R_i^2 E_{f_2}(t_i) \leq 1 - R_1^2 E_{f_2}(t_1) = E_{f_2}(1 - t_1 R_1^2) \]

Note that for the ordinary hyper-\(g\) prior, the inequality is replaced by the equality

\[ E(\sigma^2 | y) = \frac{||y||^2}{n - 3} \frac{2F_1 \left( \frac{n-3}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)}{2F_1 \left( \frac{n-1}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)} \]

We shall use the following identity for Gaussian hypergeometric functions to simplify the RHS of the inequality (equality under the ordinary hyper-\(g\)):

\[ \lim_{z \to 1} (1 - z)^{a+b-c} 2F_1 (a, b; c; z) = \frac{\Gamma(a+b-c)\Gamma(c)}{\Gamma(a)\Gamma(b)} \]

when \(a + b - c > 0\).

As \(N \to \infty\), \(||y|| \to \infty\) and \(R_1^2 \to 1\). Thus,

\[
\lim_{N \to \infty} \frac{||y||^2}{n - 3} \frac{2F_1 \left( \frac{n-3}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)}{2F_1 \left( \frac{n-1}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)}
\]

\[= \lim_{N \to \infty} \frac{||y||^2 (1 - R_1^2)}{n - 3} \frac{(1 - R_1^2)^{\frac{a+2}{2}+1-\frac{a+p_1}{2}}}{(1 - R_1^2)^{\frac{n-1}{2}+1-\frac{a+p_1}{2}}} \frac{2F_1 \left( \frac{n-3}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)}{2F_1 \left( \frac{n-1}{2}, 1; \frac{a+p_1}{2}; R_1^2 \right)}
\]

\[= \frac{(n - 3)/2}{(n - 1 - a - p_1)/2} \times \lim_{N \to \infty} \frac{||y||^2 (1 - R_1^2)}{n - 3}, \text{ provided } n > a + p_1 + 1.
\]

Thus, \(\lim_{N \to \infty} E(\sigma^2 | y) \leq \lim_{N \to \infty} \frac{||y||^2 (1 - R_1^2)}{n - 1 - a - p_1}
\]

\[= \frac{1}{n - 1 - a - p_1} \left[ (n - p - 1)\bar{\sigma}^2 + \sum_{i=2}^{k} (X_i\hat{\beta}_i)^T (X_i\hat{\beta}_i) \right].
\]

**A.10 Proof of Theorem 3.5.1**

Information consistency under model \(\mathcal{M}_\gamma\) is equivalent to \(BF(\mathcal{M}_\gamma : \mathcal{M}_0) \to \infty\) as \(R_\gamma^2 \to 1\). We first establish the sufficiency of Condition (2). We know \(R_\gamma^2 = \sum_{j=1}^{k} R_j^2, \gamma \)

and Condition (2) of the theorem enforces \(R_i^2, \gamma \to 1\) for a given block \(i\), meaning all other \(R_j^2, \gamma \to 0, j \neq i\).
Then, \( BF(\mathcal{M}_\gamma : \mathcal{M}_0) \)
\[
= \left( \frac{a - 2}{2} \right)^{k_\gamma} \int_{(0,1)^{k_\gamma}} \left[ k_\gamma \prod_{j=1}^{k_\gamma} (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} \right] \left( 1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2 \right)^{-\frac{n-1}{2}} dt \\
\geq \left( \frac{a - 2}{2} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \left[ \int_0^1 (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} (1 - t_j R_{j,\gamma}^2)^{-\frac{n-1}{2}} dt_j \right] \text{ (by (3.12))}
\]

\[
\Rightarrow \lim_{R_{i,\gamma}^2 \to 1} BF(\mathcal{M}_\gamma : \mathcal{M}_0) \geq \lim_{R_{i,\gamma}^2 \to 1} \left( \frac{a - 2}{2} \right)^{k_\gamma} \int_0^1 (1 - t_i)^{-\frac{a + p_{i,\gamma}}{2} - 2} (1 - t_i R_{i,\gamma}^2)^{-\frac{n-1}{2}} dt_i \\
\times \prod_{j \neq i} \left[ \int_0^1 (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} dt_j \right].
\]

The first term on the RHS goes to \( \infty \) when \( n \geq a + p_{i,\gamma} - 1 \) while the rest of the terms converge to nonzero constants. Hence the Bayes factor also diverges in the limit as required to show information consistency.

The more general situation is represented through Condition (1) where the block structure does not play any role in driving consistency. For an arbitrary \( 0 < \eta < 1 \),

\[
BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \left( \frac{a - 2}{2} \right)^{k_\gamma} \int_{(0,1)^{k_\gamma}} \left[ k_\gamma \prod_{j=1}^{k_\gamma} (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} \right] \left( 1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2 \right)^{-\frac{n-1}{2}} dt \\
> \left( \frac{a - 2}{2} \right)^{k_\gamma} \int_{(1-\eta,1)^{k_\gamma}} \left[ \prod_{j=1}^{k_\gamma} (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} \right] \left( 1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2 \right)^{-\frac{n-1}{2}} dt \\
> \left( \frac{a - 2}{2} \right)^{k_\gamma} \int_{(1-\eta,1)^{k_\gamma}} \left[ \prod_{j=1}^{k_\gamma} (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} \right] \left( 1 - (1 - \eta) R_{\gamma}^2 \right)^{-\frac{n-1}{2}} dt.
\]

Therefore,

\[
\lim_{R_{i,\gamma}^2 \to 1} BF(\mathcal{M}_\gamma : \mathcal{M}_0) \geq \eta^{-\frac{n-1}{2}} \left( \frac{a - 2}{2} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \left[ \int_{1-\eta}^1 (1 - t_j)^{-\frac{a + p_{j,\gamma}}{2} - 2} dt_j \right] \\
\text{ (by Monotone Convergence Theorem)}
\]

\[
= \eta^{-\frac{n-1}{2}} \left( \frac{a - 2}{2} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \left[ \frac{2}{a + p_{j,\gamma} - 2} \eta^{-\frac{a + p_{j,\gamma}}{2} - 1} \right] \\
= \eta^{-\frac{n-1}{2}} k_\gamma + \frac{ak_\gamma}{2} + \frac{p_\gamma}{2} \prod_{j=1}^{k_\gamma} \left[ \frac{a - 2}{a + p_{j,\gamma} - 2} \right]
\]

205
where $p_\gamma = \sum_{j=1}^{k_\gamma} p_{i,\gamma}$ and the above inequality holds for any choice of $\eta \in (0, 1)$. When $n > k_\gamma(a - 2) + p_\gamma + 1$, the exponent of $\eta$ is negative and since the inequality is true for any such $\eta$, it holds for the limit $\eta \to 0$ as well, indicating that $\lim_{R^2_i \to 1} BF(M_\gamma : M_0) = \infty$.

### A.11 Proof of Theorem 3.5.2

Let $R^2_{i,T}$ and $p_{i,T}$ represent the component of $R^2$ and the number of predictors in the $i^{th}$ block of the true model $M_T$ while $R^2_{i,\gamma}$ and $p_{i,\gamma}$ denote the corresponding entities for model $M_\gamma$. Further assume $B_\gamma$ denotes the set of indices of the blocks within $M_\gamma$, and let $k_\gamma = |B_\gamma|$. Recall that $B_T$ and $k_T$ are the block indices and the number of blocks respectively in model $M_T$.

We shall use the following two lemmas in the proof of this theorem. The results from Lemma A.11.1 are slightly generalized versions of Lemmas B.2 and B.3 from Maruyama and George (2011) and can be proved in a similar way. Conditions 3.5.1 and 3.5.2 are used to prove Lemma A.11.1.

**Lemma A.11.1.** Let $R^2_{i,\gamma}$ and $R^2_{i,T}$ denote the $i^{th}$ component of $R^2$ under an arbitrary model $M_\gamma$ and the true model $M_T$.

(i) Then for $i \in B_T$,

$$R^2_{i,\gamma} \xrightarrow{P} \frac{\beta_{i,T}^T D_{i,T} \beta_{i,T} - V_{i,\gamma}}{\sigma^2 + \sum_{j \in B_T} \beta_{j,T}^T D_{j,T} \beta_{j,T} + \alpha_T^2}$$

$$R^2_{i,T} \xrightarrow{P} \frac{\beta_{i,T}^T D_{i,T} \beta_{i,T}}{\sigma^2 + \sum_{j \in B_T} \beta_{j,T}^T D_{j,T} \beta_{j,T} + \alpha_T^2}$$

where $D_{i,\gamma} = \lim_{n \to \infty} \frac{1}{n} X_{i,\gamma}^T X_{i,\gamma}$ is positive definite for all models $M_\gamma$ (and all blocks) and
\[ V_{i, \gamma} = \lim_{n \to \infty} \frac{1}{n} \beta_{i,T}^T X_{i,T} (I - P_{X_{i, \gamma}}) X_{i,T} \beta_{i,T} = \begin{cases} 0 & X_{i, \gamma} \supset X_{i,T} \\ > 0 & X_{i, \gamma} \not\supset X_{i,T} \end{cases} \]

(ii) When \( i \not\in B_T, R_{i, \gamma}^2 \to 0. \)

(iii) For \( i \in B_\gamma \setminus B_T, \) \( n R_{i, \gamma}^2 \to \chi^2_{p_{i, \gamma}} = O_p(1). \) (\( c \) is a constant)

(iv) For any model \( M_\gamma \supset M_T, \) \( \left( 1 - \sum_{j \in B_T} R_{j,T}^2 \right) \left( 1 - \sum_{j \in B_\gamma} R_{j,\gamma}^2 \right)^n = \left( \frac{1 - R_{T}^2}{1 - R_{\gamma}^2} \right)^n \) is bounded from above in probability.

**Lemma A.11.2.** Consider the function \( h(t) \) defined on \( (0, 1)^k \) as
\[ h(t) = \sum_{i \in I} b_i \log(1 - t_i) - m \log(1 - \sum_{i \in I} t_i r_i) \]
where each \( b_i > 0, r_i \geq 0, m > \sum_{i \in I} b_i \) and \( \sum_{i \in I} r_i < 1. \) Then the (unique) maximum of \( h(t) \) is attained at the point \( t = t^* \) in the interior of the set \( (0, 1)^k \) with \( t^*_i = 1 - \frac{b_i(1-r)}{r_i(m-b)} \) for all \( i \in I, \) where \( b = \sum_{i \in I} b_i \) and \( r = \sum_{i \in I} r_i. \)

If we denote the Hessian matrix as \( H(t) = \left( \frac{\partial^2 h(t)}{\partial t_i \partial t_j} \right), \) then
\[ H_{ij}(t^*) = \frac{(m - b)^2 r_i r_j}{m(1 - r)^2}, \quad i \neq j \]
and
\[ H_{ii}(t^*) = -\frac{(m - b)^2 r_i^2}{(1 - r)^2} \left[ \frac{1}{b_i} - \frac{1}{m} \right]. \]

The proof of Lemma A.11.2 is skipped for brevity. It is not difficult to check that the partial derivatives of \( h(t) \) attain a value of zero at \( t^*. \) The Hessian matrix is non-positive definite at \( t = t^*. \)

Returning to the proof of the theorem, first consider the case when the true model differs from the null (intercept only) model, i.e., \( M_T \neq M_0. \)

The Bayes factor comparing model \( M_\gamma \) to the true model \( M_T \) can be written as
\[
BF(M_\gamma : M_T) = \frac{BF(M_\gamma : M_0)}{BF(M_T : M_0)} = \left( \frac{a - 2}{2} \right)^{k_\gamma-k_T} \int \prod_{i \in B_\gamma} \left( 1 - t_i \right)^{-a + p_{i,\gamma}} - 2 \left( 1 - \sum_{i \in B_\gamma} t_i R_{i, \gamma}^2 \right)^{-\frac{a-1}{2}} dt \]
\[
\int \prod_{i \in B_T} \left( 1 - t_i \right)^{-a + p_{i,T}} - 2 \left( 1 - \sum_{i \in B_T} t_i R_{i,T}^2 \right)^{-\frac{a-1}{2}} dt \]

207
\[ (\frac{a-2}{2})^{k_\gamma-k_T} \int \frac{\exp(h_\gamma(t))}{\exp(h_T(t))} dt, \]  

(A.2)

We define \( h_\gamma(t) = \sum_{i \in B_\gamma} (\frac{a+p_i}{2} - 2) \log(1 - t_i) - n^{-1} \log(1 - \sum_{i \in B_\gamma} t_i R_{i,\gamma}^2) \) and \( h_T(t) \) is defined similarly. We wish to use Lemma A.11.2 on these two functions, and that requires the following conditions to hold:

(A) \( \frac{a+p_i}{2} > 2 \) for all \( i \in B_\gamma \) and \( \frac{a+p_i}{2} > 2 \) for all \( i \in B_T \).

(B) \( n > p + 1 + (a - 4) \times \min(k_\gamma, k_T) \).

It is easy to see that (B) is satisfied since \( n \geq p + 2 \) and \( 2 < a \leq 4 \). For (A) to hold, we must have \( a + p_i > 4 \) for all blocks in both models. If \( a > 3 \), this condition is automatically satisfied and we proceed with the proof assuming that \( a > 3 \) so that (A) is true. The proof for \( 2 < a \leq 3 \) is provided in Appendix A.16.

Using the multivariate generalization of the Laplace approximation, we can approximate (A.2) up to an \( O(\frac{1}{n}) \) term as \( BF(\mathcal{M}_\gamma : \mathcal{M}_T) \)

\[ \approx \left( \frac{a-2}{2} \right)^{k_\gamma-k_T} \frac{|H_T(\hat{t}_T)|^{1/2}}{|H_\gamma(\hat{t}_\gamma)|^{1/2}} \exp[\sum_{i \in B_\gamma} b_{i,\gamma} \log(1 - \hat{t}_{i,\gamma}) - m \log(1 - \sum_{i \in B_\gamma} \hat{t}_{i,\gamma} R_{i,\gamma}^2)] \]

where \( b_{i,j} = \frac{a+p_i}{2} - 2, m = \frac{n-1}{2} \) and \( H_j(\hat{t}_j) \) is the Hessian matrix of \( h_j(t) \) evaluated at the maximizer \( \hat{t}_j \) of \( h_j(t) \); \( j \in \{\gamma, T\} \).

Using Lemmas A.11.1 and A.11.2 we can show that \( |H_T(t_T)| = O(m^{2q_T}) \) and also \( |H_\gamma(\hat{t}_\gamma)| = O(m^{2q_\gamma}) \) for some \( 0 \leq q_\gamma \leq k_T \). This follows from the fact that \( (H_\gamma(t_\gamma))_{i,i} = O(m^2 R_{i,i,\gamma}^2) \) and so \( q_\gamma = k_\gamma - L_\gamma \), where \( L_\gamma \) is the number of components \( R_{i,i,\gamma}^2 \) going to zero in probability.
Then for large $m = \frac{n-1}{2}$ (or equivalently for large $n$), $BF(\mathcal{M}_\gamma : \mathcal{M}_T)$

$$
\approx O(m^{k_T-q_\gamma}) \exp \left[ -m \log \left( \frac{m(1-R^2_\gamma)}{m-b_\gamma} \right) + m \log \left( \frac{m(1-R^2_T)}{m-b_T} \right) 
+ \sum_{i \in B_\gamma} b_{i,\gamma} \log \left( \frac{b_{i,\gamma}(1-R^2_\gamma)}{R^2_{i,\gamma}(m-b_\gamma)} \right) - \sum_{i \in B_T} b_{i,T} \log \left( \frac{b_{i,T}(1-R^2_T)}{R^2_{i,T}(m-b_T)} \right) \right] \left( \frac{a-2}{2} \right)^{k_\gamma-k_T} 
= O(m^{k_T-q_\gamma}) \exp \left[ m \log \left( \frac{1-R^2_T}{1-R^2_\gamma} \right) + (b_T-b_\gamma) \log m - \sum_{i \in B_\gamma} b_{i,\gamma} \log(R^2_{i,\gamma}) + O(1) \right]
$$

where $R^2_i = \sum_{j \in B_j} R^2_{i,j}$ and $b_j = \sum_{i \in B_j} b_{i,j}$ for $j \in \{\gamma, T\}$.

Case 1: $\mathcal{M}_\gamma \not\supset \mathcal{M}_T$

From Lemma A.11.1 (i) and (ii), $R^2_\gamma < R^2_T$ in this case and hence $\log \left( \frac{1-R^2_\gamma}{1-R^2_T} \right)$ is negative. Again

$$
\sum_{i \in B_\gamma} b_{i,\gamma} \log(R^2_{i,\gamma}) = \sum_{i \in J} b_{i,\gamma} \log(R^2_{i,\gamma}) + \sum_{i \in J} b_{i,\gamma} \log(R^2_{i,\gamma}) = C + \sum_{i \in J} b_{i,\gamma} \log(mR^2_{i,\gamma}) - \sum_{i \in J} b_{i,\gamma} \log m,
$$

where $J \subseteq B_\gamma$ is the set of indices such that $R^2_{i,\gamma} \to 0$ for $i \in J$. Then

$$
\lim_{m \to \infty} BF(\mathcal{M}_\gamma : \mathcal{M}_T) = \lim_{m \to \infty} O(m^s) \cdot O(f^m) = 0
$$

where $0 < f < 1$ and $s$ is some real number which might be positive or negative depending on the block structures of models $\mathcal{M}_\gamma$ and $\mathcal{M}_T$. The limit is always zero regardless of the value of $s$ since the second term goes to zero at an exponential rate and the first term is either bounded in probability ($s = 0$) or goes to zero ($s < 0$) or to infinity ($s > 0$) at a polynomial rate.

Case 2: $\mathcal{M}_\gamma \supset \mathcal{M}_T$

Case 2A: $\mathcal{M}_\gamma$ has the same block structure as $\mathcal{M}_T$, i.e., $B_\gamma = B_T$, but has more predictors in at least one of the blocks.

In this case $R^2_\gamma \geq R^2_T$ but due to Lemma A.11.1 (iv), $\left( \frac{1-R^2_\gamma}{1-R^2_T} \right)^m$ is bounded in probability. Also Lemma A.11.1 (i) confirms that none of the $R^2_{i,\gamma}$ converge to zero in the limit and so $q_\gamma = k_\gamma = k_T$. 

209
Note that $b_\gamma = \sum_{i \in B_\gamma} b_{i,\gamma} = \sum_{i \in B_\gamma} \left[ \frac{a + p_{i,\gamma}}{2} - 2 \right] = \frac{p_\gamma}{2} - (4 - a) \frac{k_\gamma}{2}$ and similarly $b_T = \frac{p_T}{2} - (4 - a) \frac{k_T}{2}$. For the model $M_\gamma$, $b_T - b_\gamma = \frac{p_T - p_\gamma}{2} < 0$ and $k_T = k_\gamma$. Hence,

$$\lim_{m \to \infty} BF(M_\gamma : M_T) = \lim_{m \to \infty} O(1) \exp \left[ O(1) + \left( \frac{p_T - p_\gamma}{2} \right) \log m \right] = 0$$

Case 2B: $M_\gamma$ has more blocks than $M_T$, i.e., $B_\gamma \supset B_T$ and in addition has more predictors in at least one of the blocks common to both models.

As before $\left( \frac{1 - R^2_{i,\gamma}}{1 - R^2_{i,T}} \right)^m$ is bounded in probability and all $R^2_{i,\gamma} \to 0$ for $i \in B_\gamma \setminus B_T$ which implies that $q_\gamma = k_\gamma - (k_\gamma - k_T) = k_T$. Since $b_T - b_\gamma = \sum_{i \in B_T} \frac{p_{i,T}}{2} - (4 - a) \frac{k_T}{2} - \sum_{i \in B_\gamma} \frac{p_{i,\gamma}}{2} + (4 - a) \frac{k_\gamma}{2}$, we have

$$m \log \left( \frac{1 - R^2_T}{1 - R^2_{i,\gamma}} \right) + (b_T - b_\gamma) \log m - \sum_{i \in B_\gamma} b_{i,\gamma} \log(R^2_{i,\gamma})$$

$$= O(1) + \left[ \frac{1}{2} \sum_{i \in B_T} (p_{i,T} - p_{i,\gamma}) + \frac{1}{2} \sum_{i \in B_\gamma \setminus B_T} (0 - p_{i,\gamma}) - (4 - a) \frac{k_T - k_\gamma}{2} \right] \log m$$

$$- \sum_{i \in B_\gamma \setminus B_T} b_{i,\gamma} \log(mR^2_{i,\gamma}) + \sum_{i \in B_\gamma \setminus B_T} b_{i,\gamma} \log m$$

(since $mR^2_{i,\gamma} = O(1)$ by Lemma A.11.1 (iii))

$$= O(1) + \log m \left[ \frac{1}{2} \sum_{i \in B_T} (p_{i,T} - p_{i,\gamma}) - \frac{1}{2} \sum_{i \in B_\gamma \setminus B_T} p_{i,\gamma} + (a - 4) \frac{k_T - k_\gamma}{2} \right.$$

$$\left. + \frac{1}{2} \sum_{i \in B_\gamma \setminus B_T} p_{i,\gamma} + (a - 4) \frac{k_\gamma - k_T}{2} \right]$$

$$= O(1) + \log m \left[ \frac{1}{2} \sum_{i \in B_T} (p_{i,T} - p_{i,\gamma}) \right].$$

$$\Rightarrow \lim_{m \to \infty} BF(M_\gamma : M_T) = \lim_{m \to \infty} O(1) \exp \left[ O(1) + \frac{\sum_{i \in B_T} (p_{i,T} - p_{i,\gamma})}{2} \log m \right] = 0,$$

since $p_{i,\gamma} \geq p_{i,T}$ $\forall$ $i \in B_T$ with strict inequality for at least one $i$ ensures that the above sum is strictly negative.
Case 2C: $M_\gamma$ has more blocks than $M_T$, but has the exact same set of predictors in all the blocks common to both.

The Bayes factor in this case equals

$$BF(M_\gamma : M_T) \approx O(1) \exp \left[ O(1) + \sum_{i \in B_T} (p_{i,T} - p_{i,\gamma}) \log m \right]$$

$$= O(1) \exp [O(1) + 0] = O(1), \text{ for all } m$$

since $p_{i,\gamma} = p_{i,T} \forall i \in B_T$. The $O(1)$ term here is a combination of finite constants and random variables. As a consequence, the Bayes factor cannot equal 0 with probability 1. This is the only case where the Bayes factor of any arbitrary model $M_\gamma$ compared to the true model $M_T$ does not go to zero with increasing sample size, violating the principle of model selection consistency.

The case when $M_T = M_0$ is identical to Case 2C with $B_T = \phi$, and in this case too, the Bayes factor $BF(M_\gamma : M_0)$ will be not converge to zero in probability. Model selection consistency does not hold in this case.

A.12 Proof of Lemma 3.5.1

We follow the same notation as in Theorem 3.5.2. For any $i \in B_T$,

$$\int_{(0,1)^k_T} \frac{g_i}{1 + g_i} \pi(g \mid M_T, y)dg$$

$$= \int_{(0,1)^k_T} t_i \prod_{j=1}^{k_T} (1 - t_j)^{-\frac{a + p_{j,T}}{2} - 2} (1 - \sum_{j=1}^{k_T} t_j R_{j,T}^2)^{-\frac{n-1}{2}} dt$$

$$\geq \int_{(0,1)^k_T} t_i \prod_{j=1}^{k_T} (1 - t_j)^{-\frac{a + p_{j,T}}{2} - 2} (1 - t_i R_{i,T}^2)^{-\frac{n-1}{2}} dt \quad \text{(by Lemma 3.4.2)}$$

$$= \frac{\int_0^1 t_i (1 - t_i)^{-\frac{a + p_{i,T}}{2} - 2} (1 - t_i R_{i,T}^2)^{-\frac{n-1}{2}} dt_i}{\int_0^1 (1 - t_i)^{-\frac{a + p_{i,T}}{2} - 2} (1 - t_i R_{i,T}^2)^{-\frac{n-1}{2}} dt_i}.$$
Given a specific index $i$, define $m = \frac{n-1}{2}$, $b = \frac{a+\rho_iX}{2} - 2$ and $z = R_i^2$, where $0 < z < 1$ for all $n$ (since the predictor $X_{i,T}$ is part of the true model). Then,

$$E\left(\frac{g_i}{1+g_i} \mid \mathcal{M}_T, y\right) \geq \frac{\int_0^1 t(1-t)^b(1-tz)^{-m}dt}{\int_0^1 (1-t)^b(1-tz)^{-m}dt} = \sum_{k=0}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+1)(k+1)^{z^k}} \frac{1}{z^k}. \quad \text{(from Theorem 3.2.1)}$$

For an arbitrary $\eta > 0$, $\exists N_0$ (not depending on $m$) such that $\forall k > N_0$, $\frac{1}{1+z^{k+1}} > 1 - \eta$.

So, $E\left(\frac{g_i}{1+g_i} \mid \mathcal{M}_T, y\right) > \sum_{k=0}^{N_0} \frac{\Gamma(m+k)(k+1)}{\Gamma(b+k+3)} z^k + (1-\eta) \sum_{k=N_0+1}^{\infty} \frac{\Gamma(m+k)}{\Gamma(b+k+2)} z^k = \frac{q_1 + (1-\eta)T}{q_2 + T} = (1-\eta) + \frac{q_1 - (1-\eta)q_2}{q_2 + T}.$

To prove the lemma we have to show that for any $i \in B_T$, $\lim_{m \to \infty} E\left(\frac{g_i}{1+g_i} \mid \mathcal{M}_T, y\right) = 1$, which is equivalent to proving that $\lim_{m \to \infty} \frac{q_1 - (1-\eta)q_2}{q_2 + T} = 0$.

$$\frac{q_1 - (1-\eta)q_2}{q_2 + T} = \frac{q_2}{\Gamma(m+N_0+1)} \frac{[q_1 - (1-\eta)q_2]/\Gamma(m + N_0 + 1)} + \sum_{k=N_0+2}^{\infty} \frac{\Gamma(m+k)z^k}{\Gamma(b+k+2)\Gamma(m+N_0+1)}.$$

Now for any $0 < z < 1$, $\sum_{k=N_0+2}^{\infty} \frac{\Gamma(m+k)z^k}{\Gamma(b+k+2)\Gamma(m+N_0+1)} < \infty$ for any finite $m$. With a large enough $m$ it is possible to make $\frac{q_1 -(1-\eta)q_2}{\Gamma(m+N_0+1)}$ arbitrarily small while the denominator is a finite number exceeding $\frac{(\rho - \Phi)(N_0+1)}{\Gamma(b+N_0+3)}$, where $\Phi$ is some small positive number (smaller than $\rho$). Here $\rho$ can be defined as $\rho = \lim_{n \to \infty} R_i^2$ which must satisfy $0 < \rho < 1$ since the collection of predictors $X_{i,T}$ is part of the true model (see Lemma A.11.1 in Appendix A.11).

Thus, we can find a large enough number $M$ so that given any arbitrary $\eta > 0$ and $\delta > 0$, whenever $m \geq M$

$$E\left(\frac{g_i}{1+g_i} \mid \mathcal{M}_T, y\right) > 1 - \eta - \delta \quad \implies \lim_{n \to \infty} E\left(\frac{g_i}{1+g_i} \mid \mathcal{M}_T, y\right) = 1,$$

for any $i \in B_T$. 212
A.13 Proof of Theorem 3.6.3

As in the proof of Theorem 3.5.1, we start by proving the sufficiency of Condition (2). As before, \(R_\gamma^2 = \sum_{j=1}^{k_\gamma} R_{j,\gamma}^2\) together with Condition (2) implies \(R_{i,\gamma}^2 \to 1\) for a given block \(i\) and all other \(R_{j,\gamma}^2 \to 0, j \neq i\).

Then \(BF(M_\gamma : M_0) =\)

\[
\left(\frac{a - 2}{2n}\right)^{k_\gamma} \int_{(0,1)^{k_\gamma}} \left[ \prod_{j=1}^{k_\gamma} (1 - t_j)^{\frac{a + p_{j,\gamma}}{2} - 2} \left(1 - \frac{n - 1}{n} t_j\right)^{-\frac{a}{2}} \right] \times \left(1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2\right)^{-\frac{n-1}{2}} dt
\]

\[
\geq \left(\frac{a - 2}{2n}\right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \int_0^1 (1 - t_j)^{\frac{a + p_{j,\gamma}}{2} - 2} \left(1 - \frac{n - 1}{n} t_j\right)^{-\frac{a}{2}} (1 - t_j R_{j,\gamma}^2)^{-\frac{n-1}{2}} dt_j
\]

by applying (3.12).

So,

\[
\lim_{R_{i,\gamma}^2 \to 1} BF(M_\gamma : M_0) \geq \lim_{R_{i,\gamma}^2 \to 1} \left(\frac{a - 2}{2n}\right)^{k_\gamma} \int_0^1 (1 - t_i)^{\frac{a + p_{i,\gamma}}{2} - 2} \left(1 - \frac{n - 1}{n} t_i\right)^{-\frac{a}{2}} (1 - t_i R_{i,\gamma}^2)^{-\frac{n-1}{2}} dt_i
\]

\[
\times \prod_{j \neq i} \int_0^1 (1 - t_j)^{\frac{a + p_{j,\gamma}}{2} - 2} \left(1 - \frac{n - 1}{n} t_j\right)^{-\frac{a}{2}} dt_j
\]

\[
\geq \lim_{R_{i,\gamma}^2 \to 1} \left(\frac{a - 2}{2n}\right)^{k_\gamma} \int_0^1 (1 - t_i)^{\frac{a + p_{i,\gamma}}{2} - 2} (1 - t_i R_{i,\gamma}^2)^{-\frac{n-1}{2}} dt_i
\]

\[
\times \prod_{j \neq i} \int_0^1 (1 - t_j)^{\frac{a + p_{j,\gamma}}{2} - 2} dt_j
\], since \(1 - \frac{n - 1}{n} t_j \leq 1\) \(\forall j\).

The first term on the RHS goes to \(\infty\) when \(n \geq a + p_{i,\gamma} - 1\) while the rest of the terms converge to nonzero constants. Hence the Bayes factor also diverges in the limit as required to show information consistency.
Directing our attention to Condition (1), notice that for any arbitrary $0 < \eta < 1$,

$$BF(\mathcal{M}_\gamma : \mathcal{M}_0)$$

$$= \left( \frac{a - 2}{2n} \right)^{k_\gamma} \int_{(0,1)^{k_\gamma}} \left[ \prod_{j=1}^{k_\gamma} (1 - t_j)^{\frac{a+p_j+2}{2}} \left( 1 - \frac{n - 1}{n} t_j \right)^{-\frac{n-1}{2}} \right] \left( 1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2 \right)^{-\frac{n-1}{2}} \, dt$$

$$> \left( \frac{a - 2}{2n} \right)^{k_\gamma} \int_{(1-\eta,1)^{k_\gamma}} \left[ \prod_{j=1}^{k_\gamma} (1 - t_j)^{\frac{a+p_j+2}{2}} \left( 1 - \frac{n - 1}{n} t_j \right)^{-\frac{n-1}{2}} \right] \left( 1 - \sum_{j=1}^{k_\gamma} t_j R_{j,\gamma}^2 \right)^{-\frac{n-1}{2}} \, dt$$

Hence, $\lim_{R_{\gamma} \to 1} BF(\mathcal{M}_\gamma : \mathcal{M}_0)$

$$\geq \eta^{\frac{-n-1}{2}} \left( \frac{a - 2}{2n} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \int_{1-\eta}^{1} (1 - t_j)^{\frac{a+p_j+2}{2}} \left( 1 - \frac{n - 1}{n} t_j \right)^{-\frac{n-1}{2}} \, dt_j$$

(by Monotone Convergence Theorem)

$$\geq \eta^{\frac{-n-1}{2}} \left( \frac{a - 2}{2n} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \int_{1-\eta}^{1} (1 - t_j)^{\frac{a+p_j+2}{2}} \, dt_j \left( \text{as } (1 - \frac{n - 1}{n} t_j) \leq 1 \right)$$

$$= \eta^{\frac{-n-1}{2}} \left( \frac{a - 2}{2n} \right)^{k_\gamma} \prod_{j=1}^{k_\gamma} \frac{2}{a + p_{j,\gamma} - 2 \eta^{\frac{a+p_j+2}{2}} - 1}$$

$$= \eta^{\frac{-n-1}{2} - k_\gamma + \frac{a k_\gamma + p_\gamma}{2}} \prod_{j=1}^{k_\gamma} \left[ \frac{a - 2}{n(a + p_{j,\gamma} - 2)} \right]$$

where $p_\gamma = \sum_{j=1}^{k_\gamma} p_{j,\gamma}$ and the above inequality holds for any choice of $\eta \in (0, 1)$. When $n > k_\gamma (a - 2) + p_\gamma + 1$, the exponent of $\eta$ is negative and since the inequality is true for any such $\eta$, it holds for the limit $\eta \to 0$ as well indicating that $\lim_{R_{\gamma} \to 1} BF(\mathcal{M}_\gamma : \mathcal{M}_0) = \infty$.

**A.14 Proof of Theorem 3.6.4**

We prove that condition (3.13) holds. That is, we show that for the true model $\mathcal{M}_T$ and any model $\mathcal{M}_\gamma \neq \mathcal{M}_T$, $BF(\mathcal{M}_\gamma : \mathcal{M}_T) \to 0$. 

214
First assume that $\mathcal{M}_T \neq \mathcal{M}_0$, the intercept only model. Then

$$BF(\mathcal{M}_\gamma : \mathcal{M}_T) = \left( \frac{a - 2}{2n} \right)^{k_\gamma - k_T} \times \frac{\int \prod_{i \in B_T} \left[ (1 - t_i) \frac{a + p_{1,i} - 2}{2} (1 - \frac{n - 1}{n} t_i)^{a/2} \right] (1 - \sum_{i \in B_T} t_i R_{i,\gamma}^2 )^{-\frac{n-1}{2} - a/2} \, dt}{\int \prod_{i \in B_T} \left[ (1 - t_i) \frac{a + p_{1,i} - 2}{2} (1 - \frac{n - 1}{n} t_i)^{a/2} \right] (1 - \sum_{i \in B_T} t_i R_{i,\gamma}^2 )^{-\frac{n-1}{2} - a/2} \, dt}. \quad (A.3)$$

A variation of the Laplace approximation can be used to approximate (A.3) up to an $O(\frac{1}{n})$ term which involves evaluation of the integrands at the maximizers of $h_\gamma(t) = \prod_{i \in B_T} \left[ (1 - t_i) \frac{a + p_{1,i} - 2}{2} (1 - \frac{n - 1}{n} t_i)^{a/2} \right] (1 - \sum_{i \in B_T} t_i R_{i,\gamma}^2 )^{-\frac{n-1}{2} - a/2}$. Without loss of generality, we assume that $a + p_{1,\gamma} > 4$ and $a + p_{1,T} > 4 \forall i$; if that is not the case we can modify the proof as in Appendix A.16 and still obtain the result.

Thus, just as in Theorem 3.5.2, we have for large $n$

$$BF(\mathcal{M}_\gamma : \mathcal{M}_T) \approx \left( \frac{a - 2}{2n} \right)^{k_\gamma - k_T} \frac{\exp \left[ \sum_{i \in B_T} b_{i,\gamma} \log(1 - \hat{t}_{i,\gamma}) - m \log \left( 1 - \sum_{i \in B_T} \hat{t}_{i,\gamma} R_{i,\gamma}^2 \right) \right]}{\exp \left[ \sum_{i \in B_T} b_{i,T} \log(1 - \hat{t}_{i,T}) - m \log \left( 1 - \sum_{i \in B_T} \hat{t}_{i,T} R_{i,T}^2 \right) \right]} \times \frac{|H_T(\hat{t}_T)|^{1/2} \prod_{i \in B_T} (1 - \frac{n - 1}{n} \hat{t}_{i,\gamma})^{-a/2}}{|H_\gamma(\hat{t}_\gamma)|^{1/2} \prod_{i \in B_T} (1 - \frac{n - 1}{n} t_{i,T})^{-a/2}}$$

where $b_{i,j} = \frac{a + p_{1,j}}{2} - 2$, $m = \frac{n - 1}{2}$ and $H_j(\hat{t}_j)$ is the Hessian matrix of $h_j(t)$ evaluated at the maximizer $\hat{t}_j$ of $h_j(t)$; $j \in \{\gamma, T\}$.

This means that $\lim_{m \to \infty} BF_{BHg/n}(\mathcal{M}_\gamma : \mathcal{M}_T)$

$$= \lim_{m \to \infty} (2m + 1)^{k_T - k_\gamma} BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T) \frac{\prod_{i \in B_T} (1 - \frac{n - 1}{n} \hat{t}_{i,\gamma})^{-a/2}}{\prod_{i \in B_T} (1 - \frac{n - 1}{n} \hat{t}_{i,T})^{-a/2}}$$

where the subscripts denote the corresponding Bayes factors for the blockwise hyper-$g$ and blockwise hyper-$g/n$ priors.
Since \(1 - \frac{n-1}{n} \hat{t}_{i, \gamma} = \frac{1}{2m+1} - \frac{2m}{2m+1} \frac{b_{i, \gamma}(1-R_{i, \gamma}^2)}{R_{i, \gamma}(m-b_{i, \gamma})}\), note that both \((1 - \frac{n-1}{n} \hat{t}_{i, \gamma})^{-a/2}\) and \((1 - \frac{n-1}{n} \hat{t}_{i, \gamma})^{-a/2}\) are \(O(m^{a/2})\) terms for \(i \in B_T\) and whenever \(R_{i, \gamma}^2 \to 0\), they correspond to \(O(1)\) terms.

So for large \(m\),

\[
BF_{BHg/n}(\mathcal{M}_\gamma : \mathcal{M}_T) = (2m + 1)^{k_T - k_\gamma} BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T) \frac{O(m^{a_k})}{O(m^{ak_T2})} = O(m^{k_T - k_\gamma + \frac{a}{2}(q_\gamma - k_T)}) BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T)
\]

where \(q_\gamma\) is defined as in Theorem 3.5.2.

**Case 1:** \(\mathcal{M}_\gamma \not\supset \mathcal{M}_T\)

Following the steps of the proof of Theorem 3.5.2 we know that \(BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T)\) goes to zero as \(m \to \infty\) \((n \to \infty)\) at an exponential rate and hence \(BF_{BHg/n}(\mathcal{M}_\gamma : \mathcal{M}_T) \to 0\) for all such models \(\mathcal{M}_\gamma\), regardless of the values of \(k_T, k_\gamma\) and \(q_\gamma\).

**Case 2:** \(\mathcal{M}_\gamma \supset \mathcal{M}_T\)

For Cases 2A and 2B in Theorem 3.5.2, \(BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T) \to 0\) at a polynomial rate. Since \(q_\gamma = k_T\) in both cases and \(k_\gamma \geq k_T\), \(BF_{BHg/n}(\mathcal{M}_\gamma : \mathcal{M}_T)\) must go to zero in probability as well.

For Case 2C of Theorem 3.5.2, \(BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T) = O(1)\), but as in Case 2A and 2B, \(q_\gamma = k_T\) along with \(k_\gamma > k_T\) so that

\[
\lim_{m \to \infty} BF_{BHg/n}(\mathcal{M}_\gamma : \mathcal{M}_T) = \lim_{m \to \infty} O(m^{k_T - k_\gamma}) BF_{BHg}(\mathcal{M}_\gamma : \mathcal{M}_T) = \lim_{m \to \infty} O(m^{k_T - k_\gamma})O(1) = 0.
\]

This establishes relation (3.13) for the block hyper-\(g/n\) prior implying that this prior is model selection consistent for \(\mathcal{M}_T \neq \mathcal{M}_0\).
Notice that when $\mathcal{M}_T = \mathcal{M}_0$, all models $\mathcal{M}_\gamma$ belong to Case 2C from Theorem 3.5.2 with the minor difference that $B_T = \phi$ in this case. Keeping this in mind, the rest of the proof should be the same as the one above.

Both Conditions 3.5.1 and 3.5.2 are essential in this proof since we borrowed results from Lemma A.11.1 that depend on these assumptions.

A.15 Proof of Theorem 3.6.5

Since the block hyper-$g/n$ prior leads to consistent model selection, $\pi(\mathcal{M}_T \mid y) \xrightarrow{P} 1$ as $n \to \infty$. Following the proof of Theorem 3.5.3 it is sufficient to prove the condition $\lim_{n \to \infty} E \left( \frac{g_i}{1+g_i} \mid \mathcal{M}_T, y \right) = 1 \ \forall \ i \in B_T$, to claim prediction consistency. As the least squares estimates for the coefficients in the true model $\mathcal{M}_T$ converge to the true regression coefficients in probability and the posterior probability for the true model tends to 1, prediction consistency is immediate once we prove the preceding condition.

Given any $i \in B_T$, $E \left( \frac{g_i}{1+g_i} \mid \mathcal{M}_T, y \right)$

\[
\begin{align*}
&= \int_{(0,1)^k_T} t_i \prod_{j=1}^{k_T} \left[ (1 - t_j)^{\frac{a + p_j.T}{2}} - 2(1 - \frac{n-1}{n} t_j)^{-a/2} \right] \left( 1 - \sum_{j=1}^{k_T} t_j R_{j,T}^2 \right)^{-\frac{n-1}{2}} dt \\
&\geq \int_{(0,1)^k_T} t_i \prod_{j=1}^{k_T} \left[ (1 - t_j)^{\frac{a + p_j.T}{2}} - 2(1 - \frac{n-1}{n} t_j)^{-a/2} \right] \left( 1 - t_i R_{i,T}^2 \right)^{-\frac{n-1}{2}} dt \\
&\geq \int_0^1 t_i \left[ (1 - t_i)^{\frac{a + p_i.T}{2}} - 2(1 - \frac{n-1}{n} t_i)^{-a/2} \right] \left( 1 - t_i R_{i,T}^2 \right)^{-\frac{n-1}{2}} dt \\
&\geq \int_0^1 t_i \left( 1 - t_i \right)^{\frac{a + p_i.T}{2}} - 2(1 - t_i R_{i,T}^2)^{-\frac{n-1}{2}} dt. \quad \text{(due to (3.15))}
\end{align*}
\]
Hence, \( \lim_{n \to \infty} E \left( \frac{g_i}{1 + g_i} \mid \mathcal{M}_T, y \right) \)
\[ \geq \lim_{n \to \infty} \int_0^1 t_i(1 - t_i) \frac{a + p_i \gamma}{2} - 2(1 - t_i R_{i,T}^2)^{-\frac{n-1}{2}} \, dt \]
\[ = 1. \text{ (by Lemma 3.5.1)} \]

So \( \lim_{n \to \infty} E \left( \frac{g_i}{1 + g_i} \mid \mathcal{M}_T, y \right) = 1 \forall i \in B_T \) and consistency for the block hyper-\( g/n \) predictions follows.

**A.16 Proof of Theorem 3.5.2 when \( 2 < a \leq 3 \)**

We rely on Conditions 3.5.1 and 3.5.2 in this proof through the use of Lemma A.11.1. When \( \mathcal{M}_\gamma \neq \mathcal{M}_0 \), as in the proof of Theorem 3.5.2 for \( 3 < a \leq 4 \),

\[
BF(\mathcal{M}_\gamma : \mathcal{M}_T) = \left( \frac{a - 2}{2} \right)^{k_\gamma - k_T} \frac{\prod_{i \in B_T} (1 - t_i)^{a + p_i \gamma - 2} (1 - \sum_{i \in B_T} t_i R_{i,T}^2)^{-\frac{n-1}{2}} \, dt}{\prod_{i \in B_T} (1 - t_i)^{a + p_i \gamma - 2} (1 - \sum_{i \in B_T} t_i R_{i,T}^2)^{-\frac{n-1}{2}} \, dt}
\]

\[ = \left( \frac{a - 2}{2} \right)^{k_\gamma - k_T} \frac{\prod_{i \in J_T} \left( 1 - t_i \right)^{-\frac{1}{2}} h_\gamma(t) \, dt}{\prod_{i \in J_T} \left( 1 - t_i \right)^{-\frac{1}{2}} h_\gamma(t) \, dt}
\]

where \( h_j(t) = \prod_{i \in J_j} \left( 1 - t_i \right)^{a + p_i \gamma - 2} \frac{\prod_{i \in B_T \setminus J_j} \left( 1 - t_i \right)^{a + p_i \gamma - 2} (1 - \sum_{i \in B_T \setminus J_j} t_i R_{i,j}^2)^{-\frac{n-1}{2}}}{\prod_{i \in B_T \setminus J_j} \left( 1 - t_i \right)^{a + p_i \gamma - 2}} \]

\[ \prod_{i \in B_T} \left( 1 - t_i \right)^{-\frac{1}{2}} \sum_{i \in B_T} t_i R_{i,j}^2 \]

with \( p_{i,j} = p_{i,j} + 1 \) when \( i \in J_j \) and \( p_{i,j} = p_{i,j} \) otherwise. Here both \( J_\gamma \subseteq B_\gamma \) and \( J_T \subseteq B_T \) are the block indices corresponding to \( a + p_{i,j} \leq 4 \) which happens only when \( 2 < a \leq 3 \) and the related \( p_{i,j} = 1; j \in \{\gamma, T\} \).

It is clear that \( b_{i,j}^* = \frac{a + p_{i,j}^*}{2} - 2 > 0 \forall i \), as in the situation described in Appendix A.11. Using a variation of the Laplace approximation,

\[
\lim_{m \to \infty} BF(\mathcal{M}_\gamma : \mathcal{M}_T) = \left( \frac{a - 2}{2} \right)^{k_\gamma - k_T} \lim_{m \to \infty} \frac{|H_T(\hat{t}_{i,\gamma})|^{1/2}}{|H_\gamma(\hat{t}_{i,\gamma})|^{1/2}} \prod_{i \in J_T} \left( 1 - \hat{t}_{i,T} \right)^{-1/2}
\]

\[
\exp \left[ \sum_{i \in B_T} b_{i,\gamma}^* \log(1 - \hat{t}_{i,\gamma}) - m \log(1 - \sum_{i \in B_T} \hat{t}_{i,\gamma} R_{i,\gamma}^2) \right] \times \exp \left[ \sum_{i \in B_T} b_{i,T}^* \log(1 - \hat{t}_{i,T}) - m \log(1 - \sum_{i \in B_T} \hat{t}_{i,T} R_{i,T}^2) \right]
\]

(A.4)

218
with the same definitions for $H_j(t_j)$, $\hat{t}_{i,j}$ and $m$ as in Appendix A.11.

If we can show that (A.4) behaves exactly like the Laplace approximation to (A.2) for any model $M_\gamma \neq M_T$, the proof is complete. Observe that the difference between the approximations to (A.2) and (A.4) comes just from the extra term

$$\prod_{i \in J_\gamma} (1 - \hat{t}_{i,\gamma}) - \frac{1}{2} \prod_{i \in J_T} (1 - \hat{t}_{i,T})$$

which equals

$$\prod_{i \in J_\gamma} \left[ \frac{b_{i,\gamma}(1-R_{i,\gamma}^2)}{R_{i,\gamma}^2(m-b_{i,\gamma})} \right]^{-1/2} \prod_{i \in J_T} \left[ \frac{b_{i,T}(1-R_{i,T}^2)}{R_{i,T}^2(m-b_{i,T})} \right]^{-1/2} = \left[ \prod_{i \in J_\gamma} O(mR_{i,\gamma}^2) / \prod_{i \in J_T} O(mR_{i,T}^2) \right]^{1/2}.$$  \hspace{1cm} (A.5)

Case 1: $M_\gamma \not\supset M_T$

In this case, the remaining term in the Laplace approximation to the integral goes to zero in probability at an exponential rate (see Appendix A.11) while the extra part will either be bounded, go to infinity or go to zero in probability at a polynomial rate. For all models $M_\gamma$ belonging to Case 1, (A.4) equals zero.

Case 2: $M_\gamma \supset M_T$

Case 2A: For Case 2A of Theorem 3.5.2 $B_\gamma = B_T$ and $M_T \subset M_\gamma$ which means that whenever $p_{i,\gamma} = 1$, we must have $p_{i,T} = 1$ whereas it is possible to have $p_{i,T} = 1$ along with $p_{i,\gamma} > 1$. This indicates that $|J_T| \geq |J_\gamma|$ for this group of models. Also note that all $R_{i,j}^2 \overset{p}{\to} C$ for some non-zero constant $C$ (by Lemma A.11.1 (i) from Appendix A.11). Hence (A.5) reduces to $O(m^{\frac{|J_\gamma| - |J_T|}{2}})$ while the other part of the integral is of the order $O(m^{\frac{|p_{T} - p_\gamma|}{2}}) = O(m^{\frac{|p_{T} - p_\gamma| + |J_T| - |J_\gamma|}{2}})$. This implies that (A.4) equals $\lim_{m \to \infty} O(m^{p_{T} - p_\gamma}) = 0$, as in Case 2A from Appendix A.11.

Case 2B: For Case 2B of Theorem 3.5.2 $B_\gamma \supset B_T$ and $M_T \subset M_\gamma$. As before for $i \in B_T \cap B_\gamma = B_T$, whenever $p_{i,\gamma} = 1$, we must have $p_{i,T} = 1$ whereas it is possible to have $p_{i,T} = 1$ along with $p_{i,\gamma} > 1$. This translates to the inequality
\(|J_γ \cap B_T| \leq |J_T|\). Whereas for \(i \in B_γ \setminus B_T\), it might happen that \(p_{i,γ} = 1\), but the corresponding \(p_{i,T} = 0\) since the block does not exist. However, \(R_{i,γ}^2 \to 0\) and \(nR_{i,γ}^2 \overset{d}{\to} cχ^2_{p_{i,γ}}\) for \(i \in B_γ \setminus B_T\) (by Lemma A.11.1 (ii) and (iii)) so that \(mR_{i,γ}^2 = O(1)\) for all \(i \in J_γ \setminus B_T\). Thus (A.5) becomes \(O(m^{\frac{|J_γ \cap B_T| - |J_T|}{2}})\) and the other part of the integral is \(O\left(m^{\frac{1}{2} \sum_{i \in B_T} |p_{i,T} - p_{i,γ}| + \frac{1}{2} (|J_T| - |J_γ \cap B_T|)}\right)\). (A.4) is zero since it reduces to \(\lim_{m \to \infty} O\left(m^{\frac{1}{2} \sum_{i \in B_T} |p_{i,T} - p_{i,γ}|}\right) = 0\) and we have the same result as in Case 2B from Appendix A.11.

Case 2C: For Case 2C of Theorem 3.5.2 we know that there is a one to one equivalence between \(p_{i,T} = 1\) and \(p_{i,γ} = 1\) for \(i \in B_T\) since the predictors are exactly the same in all blocks common to \(\mathcal{M}_γ\) and \(\mathcal{M}_T\). This means that \(|J_γ \cap B_T| = |J_T|\), while as before, \(mR_{i,γ}^2 = O(1)\) for all \(i \in J_γ \setminus B_T\). Now (A.5) is of the order \(O(m^{\frac{|J_γ \cap B_T| - |J_T|}{2}}) = O(1)\) and the other term in the Laplace approximation is also \(O(1)\), leading to the same conclusion as in Case 2C from Appendix A.11 that the limit of the Bayes factor does not equal zero with probability 1.

The situation when \(\mathcal{M}_γ = \mathcal{M}_0\) is similar to Case 2C here and in Theorem 3.5.2 with \(B_T = \phi\). By the same reasoning, the Bayes factor does not converge to zero in probability.

A.17 Description of the Ozone Data

The ozone data set comprises of the 10 variables described below:

- **upo3**: Daily ozone concentration (maximum 1-hour average, parts per million) at Upland, CA.
- **vdht**: Vanderburg 500-millibar-pressure height (m).
- **wdsp**: Wind speed (mph) at Los Angeles International Airport (LAX).
**hmdt**: Humidity (%) at LAX.

**sbtp**: Sandburg Air Force Base temperature (°F).

**ibht**: Inversion base height at LAX.

**dgpg**: Daggett pressure gradient (mmHg) from LAX to Daggett, CA.

**ibtp**: Inversion base temperature at LAX (°F).

**vsty**: Visibility (miles) at LAX.

**day**: Calendar day, an integer between 1 and 365.
Appendix B: Appendix for Chapter 4

B.1 Proof of Lemma 4.3.2

We will prove the lemma for the simplest case when \( k = 2 \). Generalization of this result to \( k > 2 \) is immediate and has been discussed at the end.

The following identity is an essential step of the proof:

\[
\lim_{z \to 1} (1 - z)^{a + b - c} \binom{2}{a, b, c}(z) = \frac{\Gamma(a + b - c)\Gamma(c)}{\Gamma(a)\Gamma(b)}, \text{ when } \Re(c - a - b) < 0. \quad (B.1)
\]

We want to show that if \( \eta_1 + \eta_2 = M \) and \( \lim_{N \to \infty} \eta_i > 0 \), \( \lim_{N \to \infty} L = 1 \), \( \lim_{N \to \infty} M = 1 \) and \( \lim_{N \to \infty} \frac{1 - L}{M} = D \), then as \( N \to \infty \)

\[
I = \int_{(0, 1)^2} \prod_{i=1}^{2} (1 - t_i)^{a + p_i - 2}(1 - t_1\eta_1 - t_2\eta_2)^{-m} dt \to 0.
\]

The denominator of \( I \),

\[
\text{Denom} = \frac{2}{a + p_1 - 2} \left[ \int_0^1 (1 - t_2)^{a + p_2 - 2}(1 - t_2L)^{-m} dt \right]_2
\]

\[
= \frac{4}{(a + p_1 - 2)(a + p_2 - 2)} \binom{2}{m, 1, \frac{a + p_2}{2}; L}.
\]

Numerator of \( I \),

\[
\text{Num} \leq \int_0^1 \int_0^1 (1 - t_2)^{a + p_2 - 2}(1 - t_1\eta_1 - t_2\eta_2)^{-m} dt_1 dt_2
\]

(since \( (1 - t_1)^{a + p_1 - 2} \leq 1 \) when \( a \geq 3 \))
\[ \begin{align*}
&= \frac{1}{(m-1) \eta_1} \left[ \int_0^1 (1-t_2)^{a+p_2 \over 2-2} (1-t_2 \eta_2 - \eta_1)^{1-m} dt_2 \\
&\quad \quad - \int_0^1 (1-t_2)^{a+p_2 \over 2-2} (1-t_2 \eta_2)^{1-m} dt_2 \right] \\
&= \frac{(1-\eta_1)^{1-m}}{(m-1) \eta_1} \left[ \int_0^1 (1-t_2)^{a+p_2 \over 2-2} (1-t_2 \eta_2) \left(1 - \frac{\eta_2}{1-\eta_1}\right)^{1-m} dt_2 \\
&\quad \quad - \frac{2}{(a+p_2 - 2)(m-1) \eta_1} 2F_1(m-1, 1, a + p_2 \over 2; \eta_2) \right].
\end{align*} \]

So, \( \text{Num} \leq C_1 \times 2F_1(m - 1, 1, \frac{a + p_2}{2}; \frac{\eta_2}{1-\eta_1}) - C_2 \), where \( C_1 \) and \( C_2 \) are finite non-zero constants with the property that \( \lim_{N \to \infty} C_1 \) and \( \lim_{N \to \infty} C_2 \) are also finite and non-zero.

Then, \( I = \frac{\text{Num}}{\text{Denom}} \leq \frac{C_1 2F_1(m - 1, 1, \frac{a + p_2}{2}; \frac{\eta_2}{1-\eta_1}) - C_2}{C_3 2F_1(m, 1, \frac{a + p_2}{2}; L)}. \)

Applying (B.1), the following relations hold under additional restrictions on \( m \):

\[ \lim_{L \to 1} (1-L)^{m+\frac{a+p_2}{2}} 2F_1(m, 1, \frac{a + p_2}{2}; L) = \frac{\Gamma(\frac{a+p_2}{2}) \Gamma(m + 1 - \frac{a+p_2}{2})}{\Gamma(m) \Gamma(1)}, \]

\[ \lim_{\eta_1 \to 1} (1 - \frac{\eta_2}{1-\eta_1})^{m-\frac{a+p_2}{2}} 2F_1(m - 1, 1, \frac{a + p_2}{2}; \frac{\eta_2}{1-\eta_1}) = \frac{\Gamma(\frac{a+p_2}{2}) \Gamma(m - \frac{a+p_2}{2})}{\Gamma(m-1) \Gamma(1)}. \]

The second relation can be further simplified as

\[ \lim_{N \to \infty} (1-\eta_1)^{-m+\frac{a+p_2}{2}} (1-\eta_1 - \eta_2)^{m-\frac{a+p_2}{2}} 2F_1(m - 1, 1, \frac{a + p_2}{2}; \frac{\eta_2}{1-\eta_1}) = \frac{\Gamma(\frac{a+p_2}{2}) \Gamma(m - \frac{a+p_2}{2})}{\Gamma(m-1) \Gamma(1)}. \]

since \( M \to 1 \) and \( 0 < \lim_{N \to \infty} \eta_1 < 1 \).

Hence,

\[ \lim_{N \to \infty} I \leq \lim_{N \to \infty} \frac{C_1 2F_1(m - 1, 1, \frac{a + p_2}{2}; \frac{\eta_2}{1-\eta_1}) - C_2}{C_3 2F_1(m, 1, \frac{a + p_2}{2}; L)}. \]
done similarly applying \([\text{B.1}}\) and the following inequality of non-negative integrals, we must have 
\[
\lim_{N \to \infty} \frac{C_1^* (1 - M)^{\frac{a+p_2}{2} - m} - C_2}{C_3^* (1 - L)^{\frac{a+p_2}{2} - m - 1}} = \lim_{N \to \infty} \frac{C_1^* (1 - M)^{\frac{a+p_2}{2} - m} (1 - L)^{m + 1 - \frac{a+p_2}{2}} - C_2 (1 - L)^{m + 1 - \frac{a+p_2}{2}}}{C_3^*} = \frac{-C_2 \lim_{N \to \infty} (1 - L)^{m + 1 - \frac{a+p_2}{2}} + C_1^* \lim_{N \to \infty} (1 - L)^{\frac{1-L}{1-M}} C_1^*}{C_3^*} = \frac{0 + C_1^* \times D^{m - \frac{a+p_2}{2}} \times 0}{C_3^*} = 0
\]
where \(C_1^*, C_3^*\) are finite non-zero constants provided that \(m > \frac{a+p_2}{2}\). When \(k = 2\), the requirement regarding the minimum value of \(m\) is precisely the condition stated in the lemma which ensures \(m > \frac{a+p_2}{2} + k - 2 = \frac{a+p_2}{2}\). So \(\lim_{N \to \infty} I \leq 0\) and \(I\) being a ratio of non-negative integrals, we must have \(\lim_{N \to \infty} I = 0\).

This completes the proof of the lemma for \(k = 2\). The proof for \(k > 2\) can be done similarly applying \([\text{B.1}}\) and the following inequality
\[
I = \int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{\frac{a+p_i}{2} - 2} (1 - t_1 \eta_1 - t_2 \eta_2 - \ldots - t_k \eta_k)^{-m} dt \\
\int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{\frac{a+p_i}{2} - 2} (1 - t_j L)^{-m} dt \\
\leq \frac{\int_{(0,1)^k} (1 - t_j)^{\frac{a+p_i}{2} - 2} (1 - \sum_{i=1}^{k} t_i \eta_i)^{-m} dt}{\int_{(0,1)^k} \prod_{i=1}^{k} (1 - t_i)^{\frac{a+p_i}{2} - 2} (1 - t_j L)^{-m} dt}
\]
since \((1 - t_i)^{\frac{a+p_i}{2}} \leq 1\) for any \(i\) when \(a \geq 3\). The only difference in the extension to \(k > 2\) is that the required condition in the general case is \(m > \frac{a+p_i}{2} + k - 2\).

**B.2 Proof of Lemma [4.3.3]**

In the sequence \(\{\Psi_N\}\), 
\[
\mathbf{y}^T \mathbf{y} = \sum_{i=1}^{k} \mathbf{y}^T P_i \mathbf{y} + (n - p - 1) \hat{\sigma}^2 = \sum_{i=1}^{k} (Q_i \hat{\tau}_i)^T (Q_i \hat{\tau}_i) + (n - p - 1) \hat{\sigma}^2, \quad \text{and} \quad \hat{\sigma}^2 \text{ is fixed for all } N.
\]
Because of blockwise orthogonality of the design \(Q\) and invariance of least squares estimates to linear transformations, given an index \(i\), 
\[
O\left((Q_i \hat{\tau}_i)^T (Q_i \hat{\tau}_i)\right) = O(||\tau_i||^2) = O(||\beta_j||^2) \text{ for some index } j.
\]

224
When Condition 4.3.1 holds, it is evident that for \( i = l + 1, l + 2, \ldots, k \), \( R_i^2 = \frac{O(1)}{O(||\beta_s||^2)} \to 0 \), since \( ||\beta_s||^2 \to \infty \). This proves (i).

When \( i = 1, 2, \ldots, l \), we can show that \( R_i^2 = \frac{O(||\beta_s||^2)}{O(||\beta_s||^2)} \to \eta_i \) for constants \( 0 < \eta_i < 1 \) for all \( i \). This proves (ii).

It is also easy to prove that \( 1 - \sum_{i=1}^{l} R_i^2 = \frac{O(1)}{O(||\beta_s||^2)} \to 0 \), implying that \( \sum_{i=1}^{l} R_i^2 \to 1 \) as stated in (iii).

To prove (iv) when a BOB hyper-\( g \) prior is specified on the sequence \( \{\Psi_N\} \), one needs to show that for some finite \( C \geq 1 \),

\[
\lim_{N \to \infty} \frac{\int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{k} t_i R_i^2\right)^{-m} dt}{\int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-m} dt} = C
\]

where \( b_i = \frac{a + p_i}{2} - 2 \geq 0 \) (since \( a \geq 3 \)) and \( m = \frac{n-1}{2} \).

Since \( R_i^2 \geq 0 \) for each \( i \),

\[
\int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{k} t_i R_i^2\right)^{-m} dt \geq \int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-m} dt
\]

\[
\lim_{N \to \infty} \int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{k} t_i R_i^2\right)^{-m} dt \geq \lim_{N \to \infty} \int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-m} dt
\]

which indicates that \( C \geq 1 \), but it still needs to be shown that \( C \) is finite. For the remaining part of the proof we need the following proposition which has been derived in Appendix B.3.
Proposition B.2.1. Let \( f_1(t) \) and \( f_2(t) \) denote the properly normalized densities on \((0,1)^k\) with \( m > 0 \) and \( b_i \geq 0 \:\forall\: i:\)

\[
f_1(t) \propto \prod_{i=1}^{k} (1 - t_i)^{b_i} (1 - \sum_{i \in I} t_i R_i^2)^{-m}
\]

\[
f_2(t) \propto (1 - \sum_{i \in I} t_i R_i^2)^{-m}
\]

where \( I \subseteq \{1,2,..,k\} \) and \( R_i^2 > 0 \:\forall\: i \in I.\)

Assume that \( \sum_{i=1}^{k} R_i^2 < 1 \) so that both the densities defined above are proper. If \( h(t) = \left(\frac{1 - \sum_{i \in I} t_i R_i^2}{1 - \sum_{i=1}^{k} t_i R_i^2}\right)^m \), then \( E_{f_2}(h(t)) \geq E_{f_1}(h(t)), \) i.e.,

\[
\frac{\int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i \in I} t_i R_i^2\right)^{-m} dt}{\int \prod_{i=1}^{k} (1 - t_i)^{b_i} \left(1 - \sum_{i \in I} t_i R_i^2\right)^{-m} dt} \leq \frac{\int \left(1 - \sum_{i=1}^{k} t_i R_i^2\right)^{-m} dt}{\int \left(1 - \sum_{i \in I} t_i R_i^2\right)^{-m} dt}.
\]

Applying the result from Proposition B.2.1 in this problem, we have

\[
\frac{\int \prod_{i=1}^{k} (1 - t_i)^{a+\frac{p_i}{2}} \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}{\int \prod_{i=1}^{k} (1 - t_i)^{a+\frac{p_i}{2}} \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt} \leq \frac{\int \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}{\int \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt},
\]

and so

\[
\lim_{N \to \infty} \frac{\int \prod_{i=1}^{k} (1 - t_i)^{a+\frac{p_i}{2}} \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}{\int \prod_{i=1}^{k} (1 - t_i)^{a+\frac{p_i}{2}} \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt} \leq \lim_{N \to \infty} \frac{\int \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}{\int \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}.
\]

Observe that the conditions essential for using the proposition are satisfied since

\( \frac{n-1}{2} = m \geq 0 \) and \( b_i = \frac{a+p_i}{2} - 2 \geq 0 \) (because \( a \geq 3 \)).

We simply need to prove that \( \lim_{N \to \infty} \frac{\int \left(1 - \sum_{i=1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt}{\int \left(1 - \sum_{i = 1}^{l} t_i R_i^2\right)^{-\frac{n-1}{2}} dt} = C \), for some finite constant \( C \) which must be bigger than or equal to 1 as reasoned earlier. Notice that
when \( m > k \) or equivalently \( n > 2k + 1 \),

\[
\int_{(0,1)^k} \left(1 - \sum_{i=1}^k t_i z_i \right)^{-m} \, dt = \left[ (m - 1)(m - 2) \ldots (m - k) \right]^{-1} \left[ (1 - \sum_{i=1}^k z_i)^{k-m} - \sum_{i_1, \ldots, i_{k-1}} (1 - \sum_{j=1}^{k-1} z_{i_j})^{k-m} + \sum_{i_1, \ldots, i_{k-2}} (1 - \sum_{j=1}^{k-2} z_{i_j})^{k-m} - \ldots + (-1)^k 1 \right]. \tag{B.2}
\]

So,

\[
\lim_{N \to \infty} \frac{\int \left(1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt}{\int \left(1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt} = \lim_{N \to \infty} \frac{1}{\prod_{i=l+1}^k (m - i) R_i^2} \times \frac{E_N}{E_D}
\]

where \( E_N \) and \( E_D \) contain the remaining set of terms when (B.2) is applied to the numerator and denominator respectively.

For each \( i = l + 1, \ldots, k \), \( R_i^2 = \frac{O(1)}{O(||\beta_i||^2)} \), and \( E_N \) consists of terms which are either of the form \( \left( \frac{O(1)}{O(||\beta_i||^2)} \right)^k \) or \( \left( \frac{O(||\beta_i||^2)}{O(||\beta_i||^2)} \right)^k \) or simply \( O(1) \). Similarly \( E_D \) consists of terms which are either \( \left( \frac{O(1)}{O(||\beta_i||^2)} \right)^l \) or \( \left( \frac{O(||\beta_i||^2)}{O(||\beta_i||^2)} \right)^l \) or \( O(1) \). So,

\[
\lim_{N \to \infty} \frac{\int \left(1 - \sum_{i=1}^k t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt}{\int \left(1 - \sum_{i=1}^l t_i R_i^2 \right)^{-\frac{n-1}{2}} \, dt} = \left[ O(||\beta_i||^2) \right]^{k-l} \frac{O(||\beta_i||^2)^{m-k}}{O(||\beta_i||^2)^{m-l}} = O(1).
\]

Thus the limit equals a finite constant \( C \geq 1 \) as stated in the lemma.

**B.3 Proof of Proposition B.2.1**

First note that \( h(t) = \left( \frac{1 - \sum_{i \in I} t_i R_i^2}{1 - \sum_{i \in I} t_i R_i^2} \right)^m \) is non-decreasing (strictly increasing if all \( R_i^2 > 0 \)) in each of the individual coordinates \( t_i, i = 1, 2, \ldots, k \), when we keep the remaining coordinates \( t_{-i} \) fixed. A popular result on multivariate likelihood ratio order (MLRO) that appears in, for example [Karlin and Rinott (1980)], suggests that a random variable \( U \) with density \( f_1 \) is stochastically less than or equal
to another random variable V with density \( f_2 \) if the following sufficient condition holds

\[
f_1(x)f_2(y) \leq f_1(x \land y)f_2(x \lor y) \quad \forall \ x, y.
\]

Here we want to show that for any \( x, y \)

\[
\frac{1}{C_1C_2} \prod_{i=1}^{k} (1 - x_i)^{b_i} (1 - \sum_{i \in I} x_i R_i^2)^{-m} (1 - \sum_{i \in I} y_i R_i^2)^{-m} \leq \frac{1}{C_1C_2} \prod_{i=1}^{k} (1 - (x_i \land y_i))^b_i (1 - \sum_{i \in I} (x_i \land y_i) R_i^2)^{-m} (1 - \sum_{i \in I} (x_i \lor y_i) R_i^2)^{-m} \quad (B.3)
\]

where \( C_1, C_2 \) are normalizing constants. Then \( f_2(t) >_{st} f_1(t) \) and hence \( E_{f_2}(h(t)) \geq E_{f_1}(h(t)) \) for any coordinate-wise non-decreasing function \( h(t) \).

To show \((B.3)\), note that \((1 - (x_i \land y_i))^b_i \geq (1 - x_i)^b_i \land x \) whenever \( b_i \geq 0 \) and it is sufficient to show that \((1 - \sum_{i \in I} x_i R_i^2)(1 - \sum_{i \in I} y_i R_i^2) \geq (1 - \sum_{i \in I} (x_i \land y_i) R_i^2)(1 - \sum_{i \in I} (x_i \lor y_i) R_i^2) \) \( \forall \ x, y \). If we denote \( x_i \land y_i \) by \( m_i \) and \( x_i \lor y_i \) by \( M_i \) \( \forall i \), it suffices to prove that for all \( l \in \mathbb{N} \)

\[
(1 - \sum_{i=1}^{l} x_i)(1 - \sum_{i=1}^{l} y_i) \geq (1 - \sum_{i=1}^{l} m_i)(1 - \sum_{i=1}^{l} M_i).
\]

The above relation is trivially true for \( l = 2 \). Assume that it is true for all \( l = 2, 3, \ldots, q \).

If we can prove that the relation is true for \( l = q + 1 \), then the proof is done (by induction). Without loss of generality, assume that \( x_{q+1} < y_{q+1} \). Therefore,

\[
(1 - \sum_{i=1}^{q+1} x_i)(1 - \sum_{i=1}^{q+1} y_i) \\
= (1 - \sum_{i=1}^{q} x_i - x_{q+1})(1 - \sum_{i=1}^{q} y_i - y_{q+1}) \\
= (1 - \sum_{i=1}^{q} x_i)(1 - \sum_{i=1}^{q} y_i) - x_{q+1}(1 - \sum_{i=1}^{q} y_i) - y_{q+1}(1 - \sum_{i=1}^{q} x_i) + x_{q+1}y_{q+1} \\
\geq (1 - \sum_{i=1}^{q} m_i)(1 - \sum_{i=1}^{q} M_i) + m_{q+1}M_{q+1} - m_{q+1}(1 - \sum_{i=1}^{q} y_i) - M_{q+1}(1 - \sum_{i=1}^{q} x_i).
\]
Showing that the last display is greater than \((1 - \sum_{i=1}^{q+1} m_i)(1 - \sum_{i=1}^{q+1} M_i) = (1 - \sum_{i=1}^{q} m_i)(1 - \sum_{i=1}^{q} M_i) + m_{q+1}M_{q+1} - m_{q+1}(1 - \sum_{i=1}^{q} M_i) - M_{q+1}(1 - \sum_{i=1}^{q} m_i)\) is equivalent to proving that

\[
M_{q+1} \sum_{i=1}^{q} x_i + m_{q+1} \sum_{i=1}^{q} y_i \geq M_{q+1} \sum_{i=1}^{q} m_i + m_{q+1} \sum_{i=1}^{q} M_i
\]

\[\Leftrightarrow M_{q+1} \left( \sum_{i=1}^{q} (x_i - m_i) \right) \geq m_{q+1} \left( \sum_{i=1}^{q} (M_i - y_i) \right)\]

which is true since \(\sum_{i=1}^{q} (x_i - m_i) = \sum_{i=1}^{q} (M_i - y_i) = \sum_{i:x_i > y_i} (x_i - y_i)\).

This completes the proof of the Proposition.

### B.4 Proof of Theorem 4.3.6

For any \(i \in \{1, 2, \ldots, l\}\),

\[
E \left( \frac{g_i}{1 + g_i} \mid y \right) = \frac{\int_{(0,1)^k} t_1 \prod_{j=1}^k (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^k t_j R_j^2)^{-\frac{n-1}{2}} dt}
\]

\[
\geq \frac{\int_{(0,1)^k} t_1 \prod_{j=1}^k (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^k} \prod_{j=1}^k (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}} dt}
\]

\[
= \frac{\int_{(0,1)^l} t_1 \prod_{j=1}^l (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^l} \prod_{j=1}^l (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}} dt}
\]

using a minor variation of Lemma 3.4.2.

Due to Condition 4.3.1 and Lemma 4.3.1 \(\sum_{j=1}^l R_j^2 \to 1\) as \(N \to \infty\). Define \(t_{1,l} = (t_1, \ldots, t_l)\) and let \(h(t_{1,l}) \propto \prod_{j=1}^l (1 - t_j)^{a_i + p_j - 2}(1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}}\). Then \(E \left( \frac{g_i}{1 + g_i} \mid y \right) \geq E_h(t_i)\) and if we show that \(\lim_{N \to \infty} E_h(t_i) = 1\) for \(i = 1, 2, \ldots, l\), then \(\lim_{N \to \infty} E \left( \frac{g_i}{1 + g_i} \mid y \right) = 1\) as well for any such \(i\).

We argue that for any \(0 < \eta < 1\), \(\lim_{N \to \infty} P(t_{1,l} > (1 - \eta)1) = 1\). Then

\[
E_h(t_i) = E_h(t_i \mid t_{1,l} > (1 - \eta)1) P(t_{1,l} > (1 - \eta)1) + E_h \left( t_i \mid [t_{1,l} > (1 - \eta)1]^c \right) P \left( [t_{1,l} > (1 - \eta)1]^c \right)
\]

229
\[ E_h(t_i \mid t_{1:l} > (1 - \eta) \mathbf{1}) P(t_{1:l} > (1 - \eta) \mathbf{1}) > (1 - \eta) P(t_{1:l} > (1 - \eta) \mathbf{1}) \]

\[ \implies \lim_{N \to \infty} E_h(t_i) > (1 - \eta) \lim_{N \to \infty} P(t_{1:l} > (1 - \eta) \mathbf{1}) = 1 - \eta \]

for any arbitrary \(0 < \eta < 1\). This can only happen if \(\lim_{N \to \infty} E_h(t_i) = 1\) and our first claim follows.

We still need to prove that \(\lim_{N \to \infty} P(t_{1:l} > (1 - \eta) \mathbf{1}) = 1\) for any \(0 < \eta < 1\). First define \(S_d = \{t_{1:l} : t_i > 1 - \eta, \text{ for } d \text{ indices } i \text{ and } t_i \leq 1 - \eta, \text{ for } (l - d) \text{ indices } i\} \) for every \(d = 0, 1, \ldots, l - 1\). If the notation \(\sum_d\) is used to denote the sum over the \((l)\) unique elements contained in set \(S_d\), then for any \(d \in \{0, 1, \ldots, l - 1\},\)

\[
\frac{P(t_{1:l} > (1 - \eta) \mathbf{1})}{P(t_{1:l} \in S_d)} = \frac{\int_{(1 - \eta, 1)} \prod_{j=1}^l (1 - t_j)^{-a_p + \bar{p}_j} - 2 \left(1 - \sum_{j=1}^l t_j R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}{
\sum_d \int_{(0,1 - \eta)} \prod_{j=1}^l (1 - t_j)^{-a_p + \bar{p}_j} - 2 \left(1 - \sum_{j=1}^l t_j R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}
\]

\[
> \frac{\int_{(1 - \eta, 1 - \eta^c)} \prod_{j=1}^l (1 - t_j)^{-a_p + \bar{p}_j} - 2 \left(1 - \sum_{j=1}^l t_j R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}{
\sum_d \int_{(0,1 - \eta)} \prod_{j=1}^l (1 - t_j)^{-a_p + \bar{p}_j} - 2 \left(1 - \sum_{j=1}^l t_j R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}\]

\[
> \frac{\sum_{(l)_d} \prod_{j=1}^l (\eta^{c+1})^{-a_p + \bar{p}_j} - 2 \left(1 - (1 - \eta^c) \sum_{j=1}^l R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}{
\sum_{(l)_d} \prod_{j=1}^l (\eta^c)^{-a_p + \bar{p}_j} - 2 \left(1 - (1 - \eta) \sum_{j=1}^l t_j R_j^2\right)^{-\frac{n-1}{2}} dt_{1:l}}\]

where each \(I_{d,w} = \{i : t_i > 1 - \eta\}\) and \(|I_{d,w}| = d\). Obviously, there are \((l)\) many choices of \(I_{d,w}\). The last step holds because \((1 - t_j)^{-a_p + \bar{p}_j} - 2\) is non-increasing in \(t_j\) whenever \(a \geq 3\) and \((1 - \sum_{j=1}^l t_j R_j^2)^{-\frac{n-1}{2}}\) is non-decreasing in each \(t_j\).

But we know that as \(N \to \infty\), \(\sum_{j=1}^l R_j^2 \to 1\) and \(\sum_{j \in I_{d,w}} R_j^2 \to M_{d,w}\), for some \(0 \leq M_{d,w} < 1\). Note that \(M_{d,w} = 0\) is feasible only when \(d = 0\), in all other cases the
inequalities are strict. This means that as \( N \to \infty \), \( \sum_{j \not \in I_{d,w}} R_j^2 = \sum_{i=1}^{t} R_i^2 - \sum_{j \in I_{d,w}} R_j^2 \to 1 - M_{d,w} \).

Hence, given any \( 0 < \eta < 1 \) and any \( d = 0, 1, \ldots, l - 1 \)

\[
\lim_{N \to \infty} \frac{P(t_{1;l} > (1 - \eta)1)}{P(t_{1;l} \in S_d)} \geq \frac{\eta^c \sum_{j=1}^{l} \left( \frac{a + p_j - 2}{2} \right) \eta^{-c \frac{n-1}{2}} (\eta^c - \eta^{c+1})}{\sum_{u=1}^{l} \left( (1 - M_{d,w}) \eta^{-\frac{n-1}{2}} (1 - \eta)l - \eta^d \right)}
\]

\[
= \frac{(1 - \eta)^d \eta^{-\frac{n-1}{2}} \eta^{l(a - 2) + \sum_{j=1}^{l} p_j - n + 1}}{\sum_{u=1}^{l} \left( (1 - M_{d,w}) \eta^{-\frac{n-1}{2}} \eta^{n - \sum_{j \not \in I_{d,w}} p_j \eta^{l - \frac{n-1}{2}}} \right)}
\]

The last relation is true for any \( c \geq 2 \) and hence, also for the limit as \( c \to \infty \). Since \( 0 < \eta < 1 \), \( \lim_{c \to \infty} \eta^2 \left[ (a - 2) + \sum_{j=1}^{l} p_j - n + 1 \right] = \infty \) whenever \( n > l(a - 2) + \sum_{j=1}^{l} p_j + 1 \). This proves that \( \lim_{N \to \infty} \frac{P(t_{1;l} > (1 - \eta)1)}{P(t_{1;l} \in S_d)} = \infty \) for any \( \eta \) and all possible values of \( d \). Notice that

\[
\lim_{N \to \infty} \frac{1 - P(t_{1;l} > (1 - \eta)1)}{P(t_{1;l} > (1 - \eta)1)} = \lim_{N \to \infty} \frac{P(t_{1;l} \in \bigcup_{d=0}^{l-1} S_d)}{P(t_{1;l} > (1 - \eta)1)} = \lim_{N \to \infty} \frac{\sum_{d=0}^{l-1} P(t_{1;l} \in S_d)}{P(t_{1;l} > (1 - \eta)1)} = \sum_{d=0}^{l-1} \lim_{N \to \infty} \frac{P(t_{1;l} \in S_d)}{P(t_{1;l} > (1 - \eta)1)} = 0,
\]

which is possible if and only if \( \lim_{N \to \infty} P(t_{1;l} > (1 - \eta)1) = 1 \).

The next part of the proof is borrowed from the proof of Theorem 3.4.1 in Appendix A.7. Using the same steps we can show that for any \( i = l + 1, l + 2, \ldots, k \),

\[
\frac{2}{a + p_i} \leq E \left( \frac{g_i}{1 + g_i} \mid y \right) \leq \frac{2}{a + p_i} \frac{2 F_1 \left( \frac{n-1}{2}, 2; \frac{a + p_i}{2} + 1; \kappa_i \right)}{2 F_1 \left( \frac{n-1}{2}, 1; \frac{a + p_i}{2}; \kappa_i \right)}
\]

231
where \( \kappa_i = \frac{R_i^2}{1 - \sum_{j \neq i} R_j^2} = \frac{y^T P_i y}{(n-p-1)\sigma^2 + y^T P_i y} \). Observe that for all \( i > l \), \( \kappa_i \) is fixed and bounded away from 1 in the sequence \( \{ \Psi_N \} \). This signifies that
\[
\frac{2}{a + p_i} \leq \lim_{N \to \infty} E \left( \frac{g_i}{1 + g_i} \mid y \right) < 1 \quad \forall \ i > l.
\]

### B.5 Proof of Lemma 4.5.3

As before, the Bayes factor for a model \( M_l \) can be written as
\[
BF(M_l : M_0) = \left( \frac{a - 2}{2} \right)^k \int_{(0,1)^k} \prod_{i=1}^k (1 - t_i)^{a+p_i-2} \left( 1 - \sum_{i=1}^k t_i R_{i,l}^2 \right)^{-\frac{n-1}{2}} \ dt
\]
which simplifies under the conditions \( a = 3 \) and \( p_i = 1 \) \( \forall i \) in the lemma to
\[
BF(M_l : M_0) = \left( \frac{1}{2} \right)^p \int_{(0,1)^p} \left( 1 - \sum_{i=1}^p t_i R_{i,l}^2 \right)^{-\frac{n-1}{2}} \ dt.
\]

We want to show that
\[
\lim_{N \to \infty} \int_{(0,1)^p} \left( 1 - \sum_{i=1}^p t_i R_{i,l}^2 \right)^{-\frac{n-1}{2}} \ dt = 0 \quad (\text{B.4})
\]
where \( R_{i,1}^2 \) and \( R_{i,l}^2 \) denote the components of \( R^2 \) in models \( M_1 \) and \( M_l \) respectively.

Assume that in model \( M_l \), \( X_{l_1} \) and \( X_{l_2} \) are two blocks with arbitrary indices \( 1 \leq l_1 < l_2 \leq p \) that have interchanged positions in the order of block orthogonalization. This is the simplest case where only two blocks are not in the ‘correct’ order and we will prove that (B.4) is true for such a model \( M_l \). The result for more complicated cases where many blocks are wrongly ordered follows similarly.

Observe that \( R_{i,1}^2 = O(|\beta_i|^2) \) for all \( i = 1, 2, \ldots, p \). On the other hand \( R_{i,l}^2 = O(|\beta_i|^2) \) for \( i \in T = \{ l_1 + 1, \ldots, l_2 \} \) and \( R_{i,l}^2 = O(|\beta_i|^2) \) for all other \( i \not\in T \). Hence by applying (B.2) (which requires \( n > 2p + 1 \)), the LHS of (B.4) reduces to
\[
\lim_{N \to \infty} \int_{(0,1)^p} \left( 1 - \sum_{i=1}^p t_i R_{i,l}^2 \right)^{-\frac{n-1}{2}} \ dt = \lim_{N \to \infty} \prod_{i=1}^p R_{i,1}^2 \times \lim_{N \to \infty} E_1
\]

where each term in $E_l$ and $E_1$ is of the form $\left[ \frac{O(|\beta_i|^2)}{O(|\beta_1|^2)} \right]^{p-(n-1)/2}$ or $\left[ \frac{O(|\beta_i|^2)}{O(|\beta_1|^2)} \right]^{p-(n-1)/2}$ or $O(1)$. The terms of the order $\left[ \frac{O(|\beta_i|^2)}{O(|\beta_1|^2)} \right]^{p-(n-1)/2}$ will converge to a finite constant, while terms of the order $\left[ \frac{O(|\beta_i|^2)}{O(|\beta_1|^2)} \right]^{p-(n-1)/2}$ for $i > 1$. will diverge to infinity as $N \to \infty$. However, the rate of divergence is the same in both $E_l$ and $E_1$ due to which $\lim_{N \to \infty} E_l/E_1$ is a finite non-zero constant.

Hence

$$
\lim_{N \to \infty} \frac{\int_{(0,1)^p} \left(1 - \sum_{i=1}^p t_i R_{i,t}^2 \right)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^p} \left(1 - \sum_{i=1}^p t_i R_{i,t}^2 \right)^{-\frac{n-1}{2}} dt} = \lim_{N \to \infty} \prod_{t \in T} \frac{\prod_{i=1}^p O(|\beta_i|^2) O(|\beta_1|^2)}{\prod_{i=1}^p O(|\beta_i|^2) O(|\beta_1|^2)} \times C
$$

$$
= C \times \lim_{N \to \infty} \prod_{t \in T} \frac{\prod_{i=1}^p O(|\beta_i|^2) O(|\beta_1|^2)}{\prod_{i=1}^p O(|\beta_i|^2) O(|\beta_1|^2)} \times \lim_{N \to \infty} \prod_{i \in T} \frac{O(|\beta_i|^2) O(|\beta_1|^2)}{O(|\beta_i|^2) O(|\beta_1|^2)}
$$

$$
= C^* \lim_{N \to \infty} \frac{\prod_{i=l_1+1}^{l_2} O(|\beta_i|^2)}{\prod_{i=l_1+1}^{l_2} O(|\beta_1|^2)} = 0
$$

where $C$ and $C^*$ are finite and non-zero constants. The last expression goes to zero since $\lim_{N \to \infty} \frac{|\beta_i|^2}{|\beta_1|^2} = 0$, $\forall i > l_1$, in the sequence $\{\Psi_N\}$ described in (4.19) when $p_i = 1 \ \forall \ i$. This proves (B.4).

**B.6 Proof of Lemma 4.5.4**

The proof is closely related to the proof of Lemma 4.5.3 in Appendix B.5. Since all the blocks are of size one, the number of blocks is equal to the number of predictors in both models $M'$ and $M^*$. Let $p^*$ and $p'$ be the number of blocks (predictors) in the models $M^*$ and $M'$ respectively. Since $M^* \in S$, we must have $p^* = p$ or $p - 1$. Following the proof of Lemma 4.5.3 if $n > 2p + 1$,

$$
\lim_{N \to \infty} BF(M' : M^*)
$$

$$
= \lim_{N \to \infty} \frac{\int_{(0,1)^p} \left(1 - \sum_{i=1}^{p'} t_i [R_i]_2^2 \right)^{-\frac{n-1}{2}} dt}{\int_{(0,1)^p} \left(1 - \sum_{i=1}^{p^*} t_i [R_i]_2^2 \right)^{-\frac{n-1}{2}} dt} = \lim_{N \to \infty} \frac{\prod_{i=1}^{p'} [R_i]_2^2}{\prod_{i=1}^{p^*} [R_i]_2^2} \times \frac{E'}{E^*} \quad (B.5)
$$

233
where \( E' \) and \( E^* \) are derived using (B.2).

The combination of all the terms in \( E^* \) lead to \( E^* = (O(|\beta_1|^2))^{(n-1)/2-p^*} \). The order of \( E' \) is decided by the blocks present in the model \( M' \). If \( X_1 \) appears in \( M' \), then \( E' = \left( \frac{O(|\beta_1|^2)}{O(|\beta_1|^2)} \right)^{(n-1)/2-p'} \) for the smallest index \( l \) missing from \( M' \) (if all \( p \) predictors are present in model \( M' \), take \( l = p \)), otherwise \( E' = O(1) \). Recall that for each \( i \), \( |R_i|^2 = O(|\beta_i|^2)/O(|\beta_1|^2) \) and \( |R_i|^2 = O(|\beta_j|^2)/O(|\beta_1|^2) \), for some \( j = 1, \ldots, p \), which depends on the index \( i \) and the permutation of the blocks in \( M' \). This implies that \( E^*/\prod_{i=1}^{p^*} |R_i|^2 = \frac{O(|\beta_1|^2)^{(n-1)/2}}{\prod_{i=1}^{p} O(|\beta_i|^2)} \).

Case 1: \( X_1 \) is included in model \( M' \)

In this case, \( E' / \prod_{i=1}^{p^*} |R_i|^2 = \left( \frac{O(|\beta_1|^2)}{O(|\beta_1|^2)} \right)^{(n-1)/2-p'} \prod_{j \in T} O(|\beta_j|^2) \), where \( T \) is an index set with \( p' \) elements determined by the model \( M' \) and the order of the covariate blocks within \( M' \). Some of the elements in \( T \) might be repeated, and \( l \geq 2, l \notin T \) because \( X_l \) does not belong to model \( M' \). Again, from the definition of \( l = \min\{ j : X_j \notin M' \} \), one of the following two scenarios must be true:

(i) all of the indices \( \{1 : l - 1\} = \{1, 2, \ldots, l - 1\} \) belong to the set \( T \).

(ii) some (not all) of the indices \( \{1 : l - 1\} \) belong to the set \( T \), and there are at least \( l - 1 \) indices in \( T \) that are smaller than \( l \). Index 1 appears at least once among these \( l - 1 \) elements.

Hence, (B.5) reduces to the limit of the product of three parts

\[
\lim_{N \to \infty} \left( \frac{O(|\beta_1|^2)}{O(|\beta_1|^2)} \right)^{(n-1)/2} \frac{\prod_{j=1}^{p^*} O(|\beta_j|^2)}{O(|\beta_1|^2)^{(n-1)/2-p'} \prod_{j \in T \cap \{ 1 : l - 1 \}} O(|\beta_j|^2)} \prod_{j \in T \cap \{ 1 : l - 1 \}} O(|\beta_j|^2).
\]

The first part is an \( O(1) \) term, while the third part is either \( O(1) \) or goes to zero as \( N \to \infty \) depending on the elements in \( T \). (B.5) equals zero due to the second and third parts of the product when \( n \) is large enough, since \( \beta_1 \) diverges faster than all \( \beta_j \), \( j = l + 1, \ldots, p^* \), in the sequence \( \{ \Psi_N \} \). The choice \( n > 2(p+p')-1 \) always works,
but in general a smaller value of \( n \) is sufficient for the result to be true, depending on \( l \) and the permutation of the blocks in \( M' \).

**Case 2:** \( X_1 \) is not included in model \( M' \)

Then \( E'/\prod_{i=1}^{p'} |R'_i|^2 = \frac{O(1)O(|\beta_1|^2)p'}{\prod_{j \in T} O(|\beta_j|^2)} \) for an index set \( T \) of size \( p' \) which does not include the element 1 and (B.3) equals

\[
\lim_{N \to \infty} \frac{O(1)O(|\beta_1|^2)p' \prod_{j=1}^{p'-1} O(|\beta_j|^2)}{\prod_{j \in T} O(|\beta_j|^2)} O(|\beta_1|^2)^{(n-1)/2} = 0
\]

when \( n \) is sufficiently large using an argument similar to the one in Case 1.

### B.7 Description of the Boston Housing Data

The Boston housing data set consists of the following 14 variables:

- **crim**: per capita crime rate by town.
- **zn**: proportion of residential land zoned for lots over 25,000 sq.ft.
- **indus**: proportion of non-retail business acres per town.
- **chas**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
- **nox**: nitric oxides concentration (parts per 10 million).
- **rm**: average number of rooms per dwelling.
- **age**: proportion of owner-occupied units built prior to 1940.
- **dis**: weighted distances to five Boston employment centres.
- **rad**: index of accessibility to radial highways.
- **tax**: full-value property-tax rate per $10,000.
- **ptratio**: pupil-teacher ratio by town.
- **b**: \( 1000(Bk - 0.63)^2 \) where \( Bk \) is the proportion of blacks by town.
- **lstat**: % lower status of the population.
- **medv**: Median value of owner-occupied homes in $1000’s.
C.1 Derivation of $\pi(\beta \mid \sigma_\epsilon^2)$ in (5.4)

We first write $\beta$ in terms of its hyperspherical coordinates as in Section 5.4.2:

\[
\begin{align*}
\beta_1 &= \delta \cos(\phi_1) \\
\beta_2 &= \delta \sin(\phi_1) \cos(\phi_2) \\
&\vdots \\
\beta_{p-1} &= \delta \sin(\phi_1) \cdots \sin(\phi_{p-2}) \cos(\phi_{p-1}) \\
\beta_p &= \delta \sin(\phi_1) \cdots \sin(\phi_{p-2}) \sin(\phi_{p-1}),
\end{align*}
\]

where $0 \leq \phi_j \leq \pi$, for $j = 1, \ldots, p-2$, $0 \leq \phi_{p-1} < 2\pi$ and $\delta^2 = \sum_{j=1}^{p} \beta_j^2 \geq 0$. For a given $\delta^2$, the regression coefficients $\beta$ are constrained to lie on the $(p-1)$-dimensional surface of a $p$-dimensional hypersphere of radius $\delta$. The assumption that, given $\delta$, probability is uniformly distributed on this surface results in the following Jacobian for the transformation:

\[
\delta^{p-1} \sin^{p-2}(\phi_1) \sin^{p-3}(\phi_2) \cdots \sin(\phi_{p-2}) \, d\delta \, d\phi_1 \cdots d\phi_{p-1}.
\]

Therefore, for a fixed $\delta$, to obtain a uniform distribution on the surface of the hypersphere we must have

\[
\pi(\phi_1, \ldots, \phi_{p-1}) \propto \sin^{p-2}(\phi_1) \sin^{p-3}(\phi_2) \cdots \sin(\phi_{p-2}).
\]
Now,  
\[ \int_0^{\pi} \sin^n(x) \, dx = \frac{\sqrt{\pi} \Gamma \left( \frac{n+1}{2} \right)}{\Gamma \left( \frac{n}{2} + 1 \right)} \]
for \( n > -1 \), so the marginal density function for \( \phi_j, \, j = 1, \ldots, p-2 \) is 
\[ \pi(\phi_j) = \frac{\Gamma \left( \frac{p-j-1}{2} + 1 \right)}{\sqrt{\pi} \Gamma \left( \frac{p-j}{2} \right)} \sin^{p-j-1}(\phi_j), \quad 0 \leq \phi_j \leq \pi, \]
and \( \pi(\phi_{p-1}) = (2\pi)^{-1}, \quad 0 \leq \phi_{p-1} < 2\pi. \)

The properly normalized joint density function for all the \( \phi_j \) is therefore 
\[ \pi(\phi_1, \ldots, \phi_{p-1}) = \frac{1}{2\pi} \prod_{n=1}^{p-2} \sin^n(\phi_{p-1-n}) \frac{\Gamma \left( \frac{n+1}{2} \right)}{\Gamma \left( \frac{n}{2} + 1 \right)} \]
\[ = \frac{\Gamma(p/2)}{2\pi^{p/2}} \prod_{n=1}^{p-2} \sin^n(\phi_{p-1-n}). \tag{C.1} \]

Recalling that the density function for \( \delta \) given \( \sigma^2 \) is 
\[ \pi(\delta \mid \sigma^2) = \frac{2\sigma^2b}{2\pi} \frac{\Gamma(p/2)}{B(a,b)} \delta^{2a-1}(\sigma^2 + \delta^2)^{-(a+b)}, \]
the joint density function for \( \delta \) and \( \phi = (\phi_1, \ldots, \phi_{p-1})^T \) is 
\[ \pi(\phi, \delta \mid \sigma^2) = p_{\delta}(\delta)p_{\phi}(\phi_1, \ldots, \phi_{p-1}) \]
\[ = \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{\sigma^2b}{B(a,b)} \delta^{2a-1}(\sigma^2 + \delta^2)^{-(a+b)} \prod_{n=1}^{p-2} \sin^n(\phi_{p-1-n}), \]
for \( \delta \geq 0, \, 0 \leq \phi_j \leq \pi \) for \( j = 1, \ldots, p-2 \) and \( 0 \leq \phi_{p-1} < 2\pi. \)

Using the density \( \pi(\phi, \delta \mid \sigma^2) \), we can transform \((\phi_1, \ldots, \phi_{p-1}, \delta) \rightarrow (\beta_1, \ldots, \beta_{p-1}, \delta) \) to obtain a prior density function for \( \beta_1, \ldots, \beta_{p-1} \) given \( \delta \). This is a many-to-one mapping: both \( \phi_{p-1} \) and \( |2\pi - \phi_{p-1}| \) correspond to the same value of \( \beta_{p-1} \) because \( \phi_{p-1} \) range between 0 and 2\( \pi \). We will adjust for this by multiplying the resulting density function by 2. The Jacobian matrix for this transformation is
\[
\begin{pmatrix}
-\delta \sin(\phi_1) & 0 & \cdots & \cos(\phi_1) \\
\delta \cos(\phi_1) \cos(\phi_2) & -\delta \sin(\phi_1) \sin(\phi_2) & \cdots & \sin(\phi_1) \cos(\phi_2) \\
\vdots & \vdots & \ddots & \vdots \\
\delta \cos(\phi_1) \cdots \sin(\phi_{p-2}) \cos(\phi_{p-1}) & \cdots & \sin(\phi_1) \cdots \cos(\phi_{p-1}) & 0 \\
0 & 0 & \cdots & 1
\end{pmatrix},
\]
the absolute value of the determinant of which is
\[
\delta^{p-1} \sin^{p-1}(\phi_1) \sin^{p-2}(\phi_2) \cdots |\sin(\phi_{p-1})|.
\]

The joint density function of \( \beta_{-p} = (\beta_1, \ldots, \beta_{p-1})^T \) and \( \delta \) given \( \sigma^2 \) is therefore
\[
\pi(\beta_{-p}, \delta | \sigma^2) = \frac{2 \Gamma(p/2) \sigma^2}{\pi^{p/2}} \frac{\delta^{2a-1}(\sigma^2 + \delta^2)^{-(a+b)}}{BE(a,b)} \prod_{n=1}^{p-2} \sin^n(\phi_{p-1-n}) |\delta^{p-1} \sin(\phi_{p-1})| \prod_{j=1}^{p-2} \sin^{p-j}(\phi_j)
\]
\[
= \frac{\Gamma(p/2) \sigma^2}{\pi^{p/2} \delta^{2a-p}(\sigma^2 + \delta^2)^{-(a+b)}} \prod_{j=1}^{p-1} \{\sin(\phi_j)\}^{-1} \left( \frac{\delta}{\sqrt{\delta^2 - \sum_{j=1}^{p-1} \beta_j^2}} \right),
\]
since \(|\beta_p| = \sqrt{\delta^2 - \sum_{j=1}^{p-1} \beta_j^2} \).

Now consider the transformation from \((\beta_1, \ldots, \beta_{p-1}, \delta) \rightarrow (\beta_1, \ldots, \beta_{p-1}, \beta_p)\), where \( \beta_p = \pm \sqrt{\delta^2 - \sum_{j=1}^{p-1} \beta_j^2} \). This is a one-to-many mapping where \( \beta_p \) can be positive or negative for a fixed \( \delta \). We adjust for this in the density function calculation by dividing by two. The Jacobian matrix is
\[
\begin{pmatrix}
I_{p-1} \\
\mathbf{u}^T
\end{pmatrix}
\begin{pmatrix}
0_{(p-1) \times 1} \\
\frac{1}{2} (\sum_{j=1}^{p} \beta_j^2)^{-1/2} \delta
\end{pmatrix},
\]
for some \((p-1)\)-dimensional vector \( \mathbf{u} \). The absolute value of the determinant of the Jacobian matrix is \( \sqrt{\beta_p^2 / \mathbf{u}^T \beta} \).
The joint density function for $\beta$ given $\sigma^2_\epsilon$ is therefore
\[
\pi(\beta \mid \sigma^2_\epsilon) = \frac{1}{2} \pi(\beta_p, \delta \mid \sigma^2_\epsilon) \sqrt{\frac{\beta_p^2}{\beta^T \beta}}
\]
\[
= \frac{\Gamma(p/2)}{\pi^{p/2}} \frac{\sigma^{2b}_\epsilon}{B(a, b)} \delta^{2-a-p}(\sigma^2_\epsilon + \delta^2)^{-(a+b)} \left( \delta / \sqrt{\delta^2 - \sum_{j=1}^{p-1} \beta_j^2} \right) \sqrt{\frac{\beta_p^2}{\beta^T \beta}}
\]
\[
= \frac{\Gamma(p/2)}{B(a, b)} \frac{\sigma^{2b}_\epsilon}{\pi^{p/2}} (\beta^T \beta)^{a-p/2}(\sigma^2_\epsilon + \beta^T \beta)^{-(a+b)}, \ \beta \in \mathbb{R}^p.
\]

C.2 Derivation of the Scale Mixture Representation of $\beta$

We know that
\[
\pi(\beta \mid \sigma^2_\epsilon) = \frac{\Gamma(p/2)}{B(a, b)} \frac{\sigma^{2b}_\epsilon}{\pi^{p/2}} (\beta^T \beta)^{a-p/2}(\sigma^2_\epsilon + \beta^T \beta)^{-(a+b)}, \ \beta \in \mathbb{R}^p
\]
which we wish to express in the form
\[
\beta \mid v, \sigma^2_\epsilon \sim N(0, v\sigma^2_\epsilon I_p)
\]
\[
v \sim G_v(\cdot)
\]
for some distribution $G_v$ with density $g_v$ such that
\[
\pi(\beta \mid \sigma^2_\epsilon) = \int_0^\infty \frac{1}{(2\pi v\sigma^2_\epsilon)^{p/2}} \exp \left[ -\frac{1}{2v\sigma^2_\epsilon} \beta^T \beta \right] g_v(v)dv. \quad (C.2)
\]

The following method for deriving the mixing density using inverse Laplace transforms for function convolutions was suggested by Jim Berger (personal communication, October 20, 2012). Consider the function $H(s) = F(s)G(s) = \frac{1}{s^{c+(\alpha_\epsilon+\beta_\epsilon)}}$, where $F(s) = \frac{1}{s^c}$ and $G(s) = \frac{1}{s^{d+(\alpha_\epsilon+\beta_\epsilon)}}$ for some $c, d > 0$. The inverse Laplace transforms for $F$ and $G$ are respectively $f(t) = \frac{t^{c-1}}{\Gamma(c)}$ and $g(t) = \frac{t^{d-1}}{\Gamma(d)}e^{-t\sigma^2_\epsilon}$ so that $F(s) = \int_0^\infty f(t)e^{-st}dt$ and $G(s) = \int_0^\infty g(t)e^{-st}dt$. Due to the convolution theorem for Laplace transforms,
\[
H(s) = F(s)G(s) = \int_0^\infty \left( \int_0^t f(t-x)g(x)dx \right) e^{-st}dt
\]
\[
= \int_0^\infty \left( \int_0^t (t-x)^{c-1} x^{d-1} \frac{1}{\Gamma(c)} \frac{1}{\Gamma(d)} e^{-x\sigma^2_\epsilon} \right) e^{-st}dx, \quad 239
\]
Comparing the last expression with (C.2), it is clear that constants in the last result allows us to express (5.4) as

\[
\pi(\beta \mid \sigma^2) = \frac{\Gamma(p/2)\sigma^2 e^{2b/2}}{\Gamma(a)\Gamma(b)\Gamma(p/2 - a)} \int_0^\infty \left[ \int_0^t (t - x)^{p/2 - a - 1}x^{a+b-1}e^{-x\sigma^2_\epsilon}dx \right] \exp\left(-\beta^T\beta\right)dt \\
= \frac{\Gamma(p/2)\sigma^2 e^{2b/2}}{\Gamma(a)\Gamma(b)\Gamma(p/2 - a)} \int_0^\infty \left[ \int_0^{1/2v\sigma^2_\epsilon} \left( \frac{1}{2v\sigma^2_\epsilon} - x \right)^{p/2 - a - 1}x^{a+b-1}e^{-x\sigma^2_\epsilon}dx \right] \\
\times \frac{1}{2v\sigma^2_\epsilon} \exp\left(-\frac{\beta^T\beta}{2v\sigma^2_\epsilon}\right) dv
\]

Comparing the last expression with (C.2), it is clear that

\[
g_v(v) = \frac{\Gamma(p/2)\sigma^2 e^{2b-p-2p/2-1}}{\Gamma(a)\Gamma(b)\Gamma(p/2 - a)} v^{p/2-2} \int_0^{1/2v\sigma^2_\epsilon} \left( \frac{1}{2v\sigma^2_\epsilon} - x \right)^{p/2 - a - 1}x^{a+b-1}e^{-x\sigma^2_\epsilon}dx \\
= \frac{\Gamma(p/2)2^{-b}}{\Gamma(a)\Gamma(b)\Gamma(p/2 - a)} v^{-(b+1)} \int_0^1 (1 - y)^{p/2 - a - 1}e^{-y/2v}y^{a+b-1}dy
\]

(substituting \( y = 2v\sigma^2_\epsilon x \))

as \( _1F_1(A; B; z) = \int_0^1 e^{zt} t^{A-1}(1 - t)^{B-A-1} dt \) when \( B > A > 0 \).

### C.3 Log-concavity of the Full Conditional of \( 1/\sigma^2_\epsilon \)

To simplify notation, we refer to \( \sigma_\epsilon \) simply as \( \sigma \) in the proof. We know that

\[
\pi(\sigma^2 \mid \delta^2, \mathbf{y}, \phi) \propto (n-1)\sigma^2 + \delta^2)^{-a+b - 2a_1 - n - 2} \times \exp\left(-\frac{1}{\sigma^2}\left[\frac{1}{b_1} + \frac{1}{2}(\mathbf{y} - \mathbf{V}\tau)^T(\mathbf{y} - \mathbf{V}\tau)\right]\right).
\]

Letting \( v = \frac{1}{\sigma^2} \), the Jacobian for the transformation is \( -\frac{1}{v^2} \). If we define \( SS = (\mathbf{y} - \mathbf{V}\tau)^T(\mathbf{y} - \mathbf{V}\tau) \), then
\[ \pi(v \mid \delta^2, y, \phi) \propto \left( \frac{n-1}{v} + \delta^2 \right)^{-a+b} v^{a_1 + \frac{n}{2} + 1 - b} \exp \left( -v \left[ \frac{1}{b_1} + \frac{SS_1}{2} \right] \right) \left| \frac{1}{v} \right| \]
\[ \propto \left( \frac{n-1}{v} + \delta^2 \right)^{-a+b} v^{a_1 + \frac{n}{2} - b - 1} \exp \left( -v \left[ \frac{1}{b_1} + \frac{SS_1}{2} \right] \right). \]

We want to show that \( P = \log \pi(v \mid \text{rest}) \) is concave, where

\[ P = -A \log \left( \frac{n-1}{v} + \delta^2 \right) + B \log v - Cv - \log D, \]

\( A = (a+b), B = \frac{n}{2} + a_1 - b - 1, C = \frac{1}{b_1} + \frac{SS}{2}, \) and \( D \) is the normalizing constant for the density \( \pi(v \mid \delta^2, y, \phi) \). When \( n > 2(b+1) - 2a_1 \), all of the constants \( A, B, C \) and \( D \) are positive since \( a, b, a_1, b_1 > 0 \) and \( SS \geq 0 \). Now,

\[ \frac{dP}{dv} = \frac{B}{v} - C - \frac{A}{\frac{n-1}{v} + \delta^2} \left[ - (n-1) \frac{1}{v^2} \right] = \frac{B}{v} - C + \frac{(n-1)A}{(n-1)v + \delta^2 v^2} \]

\[ \Rightarrow \frac{d^2P}{dv^2} = \frac{-B}{v^2} - 0 - \frac{A(n-1)}{\left[ (n-1)v + \delta^2 v^2 \right]^2} (n-1 + 2v\delta^2) \]
\[ = \frac{-B}{v^2} - \frac{(n-1)A(n-1 + 2v\delta^2)}{\left[ (n-1)v + \delta^2 v^2 \right]^2}. \]

Because \( A, B, v > 0 \), \( \frac{d^2P}{dv^2} \) is strictly less than zero, implying that the full conditional of \( 1/\sigma_e^2 \) is log-concave.

**C.4 MH Step for the joint update of \((M, \phi_{-1})\)**

The dimension matching problem can be resolved with a joint proposal for \( M \) and \( \phi_{-1} \). Factor the joint density as

\[ \pi(M, \phi_{-1} \mid y, \delta^2, \phi_1, \sigma_e^2, \gamma) \propto \pi(\phi_{-1} \mid M, y, \delta^2, \phi_1, \sigma_e^2, \gamma) \pi(M \mid y, \delta^2, \phi_1, \sigma_e^2, \gamma). \]

Then the acceptance probability for a proposed update \((M^*, \phi_{-1}^*)\) from the current iteration \((M, \phi_{-1})\) is given by \( \alpha = \min \{1, \alpha_r\} \), where
\[
\alpha_r = \frac{\pi(M^*, \phi_{-1}^* \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma) q((M^*, \phi_{-1}^*) \rightarrow (M, \phi_{-1}))}{\pi(M, \phi_{-1} \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma) q((M, \phi_{-1}) \rightarrow (M^*, \phi_{-1}^*))} = \frac{\pi(\phi_{-1}^* \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma) \pi(M^* \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{\pi(\phi_{-1} \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma) \pi(M^* \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)} \frac{p(M^*) p(\phi_{-1} \mid M)}{p(M^* \mid M) p(\phi_{-1}^* \mid M^*)}.
\]

The proposal \( q(\cdot) \) can be decomposed as \( q((M, \phi_{-1}) \rightarrow (M^*, \phi_{-1}^*)) = p(M^* \mid M) p(\phi_{-1}^* \mid M^*) \). \( p(M^* \mid M) \) is chosen to be symmetric and the proposed model \( M^* \) has one fewer predictor or one more predictor compared to the current model \( M \). Given the proposed model and hence its dimension, each of the new angles in \( \phi_{-1}^* \) is drawn independently from a Uniform(0, \( \pi \)) distribution. Furthermore, the ratio \( \frac{p(M \mid M^*)}{p(M^* \mid M)} = 1 \) due to the symmetric proposal distribution. Thus,

\[
\alpha_r = \frac{\pi(\phi_{-1}^* \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{\pi(\phi_{-1} \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)} \frac{p(\phi_{-1} \mid M)}{p(\phi_{-1}^* \mid M^*)} \frac{p(M^*) \pi(\phi_{-1}^* \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{p(M)} \frac{\pi(M \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{\pi(\phi_{-1} \mid M, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)} = \frac{p(y \mid M^*, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{p(y \mid M, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)} \frac{p(\phi_{-1} \mid M)}{p(\phi_{-1}^* \mid M^*)} \frac{\pi(M^* \mid y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{\pi(M)} \frac{\pi(\phi_{-1}^* \mid M^*, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}{\pi(\phi_{-1} \mid M, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma)}.
\]

Due to the reparameterization and the structure of the spherical prior when conditioned on \( \phi_1 \), the density \( \pi(\phi_{-1} \mid M, y, \delta^2, \phi_1, \sigma_\epsilon^2, \gamma) \) is the same as the prior density for \( \phi_{-1} \) given a particular model \( M \). It can be shown easily that the prior density of \( \phi_{-1} \) is

\[
\pi(\phi_{-1} \mid M) = \frac{\Gamma((|M|-1)/2)}{\pi |M|-1/2} \prod_{n=1}^{|M|-3} \sin^n(\phi_{|M|-1-n})
\]

where \( |M| \) denotes the size of a model \( M \).
Bibliography


