ZHANG, YAN. Bayesian Methods for High-dimensional Data. (Under the direction of Howard D. Bondell.)

High-dimensional data arises in multiple fields, such as genetics, finance, medicine, etc. A number of statistical methods have been developed to deal with variable selection in high-dimensional linear models. In this thesis, we propose several new Bayesian methods to address such problems. Chapter 1 presents a selective literature review of existing statistical methods with a main focus on Bayesian variable selection methodologies. A selective review of recommender systems are also given in this chapter.

Chapter 2 proposes a Bayesian variable selection method by incorporating global-local priors into the credible region selection framework, which separates model fitting and variable selection, to search for the sparsest solution within the joint posterior credible regions. The Dirichlet-Laplace (DL) prior is adapted to linear regression. Posterior consistency for the normal and DL priors is shown, along with variable selection consistency. We also introduce a new method to tune hyperparameters in prior distributions for linear regression. We propose to choose the hyperparameters to minimize a discrepancy between the induced distribution on R-square and a prespecified target distribution. Prior elicitation on R-square is more natural, particularly when there are a large number of predictor variables in which elicitation on that scale is not feasible. For a normal prior, these hyperparameters are available in closed form to minimize the Kullback-Leibler divergence between the distributions.

Chapter 3 proposes a new class of priors for linear regression, the R-square induced Dirichlet Decomposition (R2-D2) prior. The prior is induced by a Beta prior on the coefficient of determination, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. We demonstrate both theoretically and empirically the advantages of
the R2-D2 prior over a number of common shrinkage priors, including the Horseshoe, Horseshoe+, and Dirichlet-Laplace priors. The R2-D2 prior possesses the fastest concentration rate around zero and heaviest tails among these common shrinkage priors, which is established based on its marginal density, a Meijer G-function. We show that its Bayes estimator converges to the truth at a Kullback-Leibler super-efficient rate, attaining a sharper information theoretic bound than existing common shrinkage priors. We also demonstrate that the R2-D2 prior yields a consistent posterior. The R2-D2 prior permits straightforward Gibbs sampling and thus enjoys computational tractability. The proposed prior is further investigated in a mouse gene expression application.

Chapter 4 is motivated by a movie rating system. This chapter proposes a full Bayesian hybrid recommender system, using a shared-variable approach for matrix completion to coerce informative missingness into a regression-based model to improve prediction performance. We jointly model the movie rating and the probability of each movie being rated as a function of the expected rating. The proposed model is fitted through MCMC sampling. The computation is stable, accurate and easy to implement. We apply our model on a movie rating system, particularly the MovieLens data. Our method illustrates significant improvement in prediction by adding informative missingness into the model.
© Copyright 2016 by Yan Zhang

All Rights Reserved
Bayesian Methods for High-dimensional Data

by
Yan Zhang

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Statistics
Raleigh, North Carolina
2016

APPROVED BY:

__________________________  ____________________________
Brian J. Reich               Alyson G. Wilson

__________________________  ____________________________
Sankar Arumugam              Howard D. Bondell
Chair of Advisory Committee
DEDICATION

To my parents.
BIOGRAPHY

Yan (Dora) Zhang was born in the 1990s in a beautiful town of Shanxi Province, China. She was admitted to the School of Mathematical Sciences at Nankai University in 2007, and graduated in 2011 with a B.S. degree in Statistics. She then joined the Department of Statistics at North Carolina State University in August 2011. She got her M.S. degree in May 2013. Under the direction of Dr. Howard Bondell and Dr. Brian Reich, she will earn her Ph.D. degree in 2016. After completing her Ph.D., she will join the Department of Biostatistics at Johns Hopkins University as a Post-doc fellow in Dr. Nilanjan Chatterjee’s group.
ACKNOWLEDGEMENTS

To begin with, I would like to express my sincerest gratitude to my advisors, Dr. Howard Bondell and Dr. Brian Reich, for their insightful guidance of my research and great help with my future career development. I have learned a lot from them, not only on professional skills, but also on how to conduct research. Their profound knowledge, kindness, patience, enthusiasm, and noble morality of teaching tirelessly, have set me a good example to stay passionate and persistent, which also motivates my future career goals. I am proud and honored to be their student. Without them, I would not come so far.

I also enjoyed the advice and guidance from many other mentors, who will have a lasting influence on my academia life. Thank Dr. Alyson Wilson for her helpful questions and constant encouragement, as well as her big smile every time we met up. I have been fortunate to work with Dr. Sankar Arumugam for a time, whose passion, diligence and kindness have left me with a deep impression. My deep thankfulness also goes to Dr. Judy Wang, for her instructions and care about me at my early years of graduate life, whose professionalism and mildness have a far-reaching effect on me. I am also grateful to Dr. Lexin Li who gave me first glance at research, always with an open door when I needed help. Thank Dr. Curtis Storlie and Dr. Hua Zhou for their collaborative work for Chapter 4 of this thesis. My thanks extend to all faculty members in our department, for their efforts to offer us so many amazing courses and staying passionate in various fields of statistics, which greatly improves my professional skills and broadens my horizon. I also thank my teachers at Nankai University greatly for their insightful comments and suggestions.

My special thanks go to the DGPs of our department, who are always ready to help. I thank Dr. Monahan and Dr. Smith especially, who have given me warm regards when I was lonely. My thanks also go to all great staffs here, Terry, Chris, Alison, Lanakila, and others, for their
fantastic service to the department. I am also thankful to Katja, Mandy, Erika and Dave, my supervisors when I was an intern at GlaxoSmithKline. I thank Wen-Hua, Pat and Lorrain, my supervisors when I was a Research Scientist intern at Avaya lab. I thank them for helping me understand more of the applications of statistics in the real world.

Thanks to my good friend and roommate Ailin for her accompanying with me during these five years. I also benefited from occasional discussions with my talented classmates during my coursework and research, especially Shikai, Yichi and Chengchun. I enjoyed being here with my fellow students and our basketball team, such as Liwen, Anran, Wenhao and many others, as well as Elyna and her family, all of whom have helped enriching my life at Raleigh. No acknowledgment would be complete without my life-long friends in China, with whom I share my happiness and sorrow. Thank my best friend, ZZ, who always knows me, understands and supports my decision at each step.

I owe Mary a lot, who showed up at a key time point in my life, and helped me realize the truth about happiness. I am extraordinarily grateful to Lilun, for the sharing of his own experiences and assistance to me during my seeking of future careers.

Last but not the least, I would express my deepest gratitude to my family. To my parents, your unconditional love and support from the other side of the earth have always been the source of my strength. To my brother, your love and courage have been continuously inspiring me to overcome difficulties in my life. Also to my grandparents in Heaven, I miss you all the time.

These five years has been an amazing journey. I gained something, but I also lost something. However, when looking back on this journey, deep inside my heart, I am grateful for every experience I’ve had these years, pleasant or unpleasant, every person I’ve met with, every book I’ve read, and every trail I’ve run; all have made me the current one, a better me.
# TABLE OF CONTENTS

List of Tables .................................................................................. viii
List of Figures .................................................................................. ix

Chapter 1  Introduction .................................................................. 1
  1.1 Classical Variable Selection Methods .................................................. 2
  1.2 Bayesian Variable Selection Methods ................................................... 3
    1.2.1 Discrete Mixture Priors .............................................................. 3
    1.2.2 Continuous Shrinkage Priors ....................................................... 5
    1.2.3 Selection Based on Posteriors .................................................... 6
  1.3 Introduction to Recommender System ............................................... 7
  1.4 Outline .................................................................................. 10

Chapter 2  High-dimensional Variable Selection via Penalized Credible Regions with Global-Local Shrinkage Priors ............................................. 11
  2.1 Introduction ........................................................................ 11
  2.2 Background .......................................................................... 14
  2.3 Penalized Credible Regions with Global-Local Shrinkage Priors ............. 15
    2.3.1 Motivation ........................................................................ 15
    2.3.2 Dirichlet-Laplace Priors ......................................................... 16
    2.3.3 Computation of Posteriors ..................................................... 17
  2.4 Asymptotic Theory .................................................................... 18
    2.4.1 Posterior Consistency of the Normal and DL Priors .................... 19
    2.4.2 Selection Consistency of Penalized Credible Regions ................. 20
  2.5 Tuning Hyperparameters .............................................................. 21
  2.6 Simulation Results ................................................................... 24
    2.6.1 Comparisons of Different Priors ............................................ 24
    2.6.2 Additional Simulations on Hyperparameter Tuning .................... 29
  2.7 Real Data Analysis ................................................................... 32
  2.8 Discussion ........................................................................ 34
  2.9 Proofs ............................................................................. 34

Chapter 3  High Dimensional Linear Regression via the R2-D2 Shrinkage Prior ..................................................... 50
  3.1 Introduction ........................................................................ 50
  3.2 A New Class of Global-Local Shrinkage Priors ................................... 53
    3.2.1 Motivation ........................................................................ 53
    3.2.2 The R2-D2 Prior ................................................................. 54
    3.2.3 Posterior Computation ......................................................... 58
  3.3 Theoretical Properties ................................................................. 59
3.3.1 Marginal Density of the R2-D2 Prior ........................................... 59
3.3.2 Asymptotic Tail Behaviors ....................................................... 60
3.3.3 Concentration Properties ....................................................... 62
3.3.4 Predictive Efficiency ............................................................. 63
3.3.5 Posterior Consistency ............................................................ 65
3.4 Simulation Study ................................................................. 66
3.5 PCR Data Analysis ............................................................... 69
3.6 Discussion ................................................................. 74
3.7 Proofs ................................................................. 75

Chapter 4 A Regression-based Method for Matrix Completion with Informative Missingness: Application to Movie Rating Data ............ 87
4.1 Introduction ............................................................... 87
4.2 Model Details ............................................................. 90
  4.2.1 Notations ............................................................. 90
  4.2.2 Hybrid Model ........................................................... 90
  4.2.3 Hybrid+Info Model ..................................................... 92
  4.2.4 Special Cases ............................................................ 93
4.3 Model Fitting ............................................................. 94
  4.3.1 Prior Specification ....................................................... 94
  4.3.2 Matrix Representation .................................................. 95
  4.3.3 MCMC Details .......................................................... 95
  4.3.4 Prediction .............................................................. 96
4.4 Analysis of MovieLens Data ................................................... 97
4.5 Discussion and Future Work ................................................ 99
4.6 Appendix: Computational Details ......................................... 100

References ............................................................... 104
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Mean area under the ROC Curve and the PRC curve for ( p = 50, n = 60 ), based on 200 datasets with standard errors in parentheses.</td>
<td>26</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Mean area under the ROC Curve and the PRC curve for ( p = 500, n = 60 ), based on 200 datasets with standard errors in parentheses.</td>
<td>28</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Mean area under the ROC Curve and the PRC curve for ( p = 1000, n = 60 ), based on 200 datasets with standard errors in parentheses.</td>
<td>28</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Average squared error for the posterior mean, given Dirichlet-Laplace prior with ( a ) fixed at ( \frac{1}{2}, \frac{1}{n} ) and ( \frac{1}{p} ), based on 200 datasets with standard errors in parentheses.</td>
<td>31</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Theoretic ( \gamma ) in the normal prior (2.3) based on Theorem 5, together with mean of the derived and tuned ( \gamma ) through methods proposed in Section 2.5, based on 200 datasets with standard errors in parentheses.</td>
<td>32</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>Mean squared prediction error and model size, with standard errors in parenthesis, based on 100 random splits of the real data.</td>
<td>33</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Asymptotic properties for Horseshoe, Horseshoe+, R2-D2 and Dirichlet-Laplace priors as discussed in Section 3.3.</td>
<td>61</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>CPU time (in hours) for a single dataset to generate 10,000 posterior samples with ( n = 60 ) and varying ( p ) for different shrinkage priors.</td>
<td>67</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>The average of the sum of squared error (SSE) for ( \beta_j ) is given separately for ( \beta_j = 0,</td>
<td>\beta_j</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>The average of the sum of squared error (SSE) for ( \beta_j ) is given separately for ( \beta_j = 0,</td>
<td>\beta_j</td>
</tr>
<tr>
<td>Table 3.5</td>
<td>The average of the sum of squared error (SSE) for ( \beta_j ) is given separately for ( \beta_j = 0,</td>
<td>\beta_j</td>
</tr>
<tr>
<td>Table 3.6</td>
<td>Mean squared prediction error, with standard errors in parenthesis, based on 100 random splits of the real data.</td>
<td>72</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Special cases of informative missingness joint model in (4.2) and (4.3).</td>
<td>94</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Test set RMSE of the MovieLens-100K.</td>
<td>98</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 50$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.</td>
<td>27</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 500$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 1000$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.</td>
<td>30</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Marginal density of the R2-D2, DL, Horseshoe prior and Cauchy distribution. In all cases, the hyperparameters are selected to ensure the interquartile range is 1.</td>
<td>60</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>The agreement of the variable selection between two different priors, based on the induced ordering through the magnitude of posterior t-statistics for PEPCK ($n = 60$ and $p = 1000$). The point $(x, y)$ on the line means in the first ordered $x$-length variable list by using the prior on $x$-axis, there are $y$ variables matching with such list by using the prior on $y$-axis.</td>
<td>73</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Plots of the MovieLens dataset. The left panel plots the number of movies rated by each user against the user’s mean ranking; the right panel plots the number of times each movie is rated against its mean rating.</td>
<td>89</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The multiple linear regression model is probably the most widely used statistical model. Consider the linear regression model,

$$Y_i = x_i^T \beta + \varepsilon_i, \ i = 1, \cdots, n,$$

where $Y_i$ is the $i$th response, $x_i$ is the $p$-dimensional vector of covariates related to the $i$th observation, $\beta = (\beta_1, \cdots, \beta_p)^T$ is the coefficient vector, and the $\varepsilon_i$’s are the error terms assumed to be normal and independent with $\text{E}(\varepsilon_i) = 0$ and $\text{Var}(\varepsilon_i) = \sigma^2$. Often, only a small subset of the covariates actually have a substantial effect on the response, while the effect of most predictors is negligible. Thus model selection is necessary to reduce the large number of candidate predictors. In high-dimensional settings, particularly when the number of predictors $p$ is much larger than the number of observations $n$, it is more challenging to conduct variable selection. High-dimensional data arises in various application fields, such as genetics, finance, medicine, etc. Therefore, a number of statistical methods have been developed to deal with variable selection in high-dimensional linear models. In this chapter, we present a selective review of the
developments in variable selection from both non-Bayesian and Bayesian viewpoint, with the major focus on recently developed Bayesian methods.

1.1 Classical Variable Selection Methods

In frequentist variable selection frameworks, the commonly used methods are backward, forward and stepwise selection, through which predictors are added or deleted from the model at each step based on some testing rule. In addition, information criterion such as Mallow’s $C_p$ (Mallows, 1973), AIC (Akaike, 1974), and BIC (Schwarz et al., 1978) have been proposed, offering a trade-off between model complexity and goodness-of-fit of the model. When the number of covariates is larger than the sample size, maximum likelihood estimation breaks down, and classic model selection criteria tend to over fit the model, motivating a rich variety of variable selection alternatives for high-dimensional data. Wang et al. (2009) introduced the modified BIC (mBIC), and Chen and Chen (2008, 2012) proposed a family of extended Bayesian information criteria (EBIC). On the other hand, there is an increasing literature on regularization methods. Classic penalized regression models include ridge regression (Hoerl and Kennard, 1970), the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), the elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), the group Lasso (Yuan and Lin, 2006), fused Lasso (Tibshirani et al., 2005), the Dantzig selector (Candes and Tao, 2007), and octagonal shrinkage and clustering algorithm for regression (OSCAR; Bondell and Reich, 2008), bridge regression (Park and Yoon, 2011), etc. Optimization algorithms related to these estimators are well developed, and most of them are extremely fast, thus making penalized regression approaches popular in high-dimensional settings.

However, there are also some limitations of the regularized methods. (i) Penalized regres-
sion methods require tuning parameters. Cross-validation is commonly used to select the tuning parameter, but this can be computationally expensive. (ii) Regularized methods solve optimization problems and only provide a point estimate of the regression coefficients. To obtain the confidence interval or the \( p \)-value of hypothesis tests, some other methods such as the bootstrap (Efron and Efron, 1982), the covariance test statistic (Lockhart et al., 2014), or valid post-selection inference (Berk et al., 2013) are required. (iii) It is challenging for penalized regression to incorporate external information or hierarchical data into the model. These limitations of regularization methods have motivated a number of Bayesian variable selection methods.

1.2 Bayesian Variable Selection Methods

In Bayesian framework, the variable selection problem is actually equivalent to parameter estimation. Unlike frequentist methods, which are trying to search for a single optimal model, Bayesian analysis tries to estimate the posterior probability of all considered models. There are two main categories of Bayesian priors: discrete mixtures and continuous shrinkage priors. The first approach specifies a discrete prior on the subset of coefficients included in the model through the introduction of latent indicator variables, and then a continuous prior for the coefficients together with a prior on the inclusion probability of indicators are given. The second approach models \( \beta \) with absolutely continuous shrinkage priors centered at zero and with heavy tails.

1.2.1 Discrete Mixture Priors

Discrete mixture priors, also referred to as spike and slab priors, or shrinking and diffusing priors, include a point mass at zero and a continuous prior elsewhere. For the regression coef-
ficients $\beta_j$, the point-mass mixture prior is

$$\beta_j \sim (1 - I_j)\delta_0 + I_j g_j, \quad j = 1, \ldots, p,$$

where $I_j$ is the indicator variable denoting whether $\beta_j$ is zero or not, $\delta_0$ is a Dirac delta measure placing mass on zero (the “spike”), and $g_j$ is an absolutely continuous density on $\mathbb{R}$ (the “slab”). For example, in Mitchell and Beauchamp (1988), $g_j$ is a diffuse uniform distribution. The prior for the zero components can also be a continuous density instead of a fixed value, then the prior can be represented as

$$\beta_j \sim (1 - I_j)g_{1j} + I_j g_{2j},$$

where $g_{1j}$ and $g_{2j}$ are different continuous densities. For example, George and McCulloch (1993) introduces stochastic search variable selection (SSVS), using a two-component Gaussian mixture priors with zero means, one with a fixed small variance as the spike prior, and the other with a larger variance as the slab prior. Ishwaran and Rao (2005) uses Gaussian mixtures with continuous bimodal priors to model hyperparameters. Narisetty et al. (2014) also works with mixture normal priors, but with hyperparameters depending on the sample size to achieve appropriate shrinkage.

Discrete mixture priors allow for sparsity in $\beta$ and simultaneously accommodate large coefficients. This model is therefore the “gold-standard” for variable selection in high-dimensional data. Despite their intuitive and appealing representations, discrete mixture priors pose several difficulties. The major difficulty is the computational challenge of estimating posterior probability spread across the $2^p$ models spanned by the inclusion indicators. Enumeration is impossible, especially when $p$ is large, and so SSVS (George and McCulloch, 1993) is required, with the expectation that promising models have a higher probability of being sampled. However, there is a risk of missing important models by using SSVS. Garcia-Donato and
Martinez-Beneito (2013) has noted this and provided a review of the literature considering this issue.

1.2.2 Continuous Shrinkage Priors

Computational issues with discrete mixture priors have motivated a number of continuous shrinkage priors. Instead of including an indicator variable in the model, continuous shrinkage priors place a prior on \( \beta_j \) directly to approximate the spike and slab priors. Polson and Scott (2010) noted that all the existing continuous shrinkage priors can be essentially written as special cases of the global-local (GL) scale mixture Gaussian family, i.e., for \( j = 1, \cdots, p \),

\[
\beta_j | \phi_j, \omega \sim N(0, \omega \phi_j), \phi_j \sim g_1, \omega \sim g_2,
\]

(1.1)

where \( \omega \) represents the global shrinkage and \( \phi_j \)'s are the local variance components. With carefully chosen \( g_1 \) and \( g_2 \), these shrinkage priors have an infinite spike at zero and very heavy tails, and the marginal density of \( \beta_j \) in (1.1) can nicely approximate the discrete mixture priors. Therefore, the resulting hierarchical models can effectively shrink small coefficients and reliably estimate the coefficients of important variables simultaneously.

Classic examples of GL shrinkage priors include Bayesian regularization priors. Actually, most frequentist penalization approaches have a corresponding Bayesian interpretation, in which the frequentist point estimator is the posterior mode under a Bayesian shrinkage prior. For example, the famous Lasso (Tibshirani, 1996) estimator is equivalent as the posterior mode under a double exponential prior; while the ridge regression estimator (Hoerl and Kennard, 1970) is equivalent to the posterior mode under a normal prior. Recent Bayesian regularization methods include Bayesian Lasso (Park and Casella, 2008; Hans, 2009), Bayesian Ridge (Polson et al., 2013), Bayesian adaptive Lasso (Leng et al., 2014), Bayesian elastic net
Compared to the frequentist penalized methods, Bayesian regularization methods have several advantages. (i) Bayesian methods offer interval estimates along with the point estimator from posterior distributions. (ii) They provide flexibility to estimate the tuning parameter jointly with the regression coefficients. (iii) They allow easy incorporation of external information or hierarchical modeling into Bayesian regularization framework. (iv) Computation is straightforward through Gibbs sampling.

In addition to Bayesian regularization shrinkage priors, there is emerging recent literature on the GL priors. Carvalho et al. (2009, 2010) described the Horseshoe prior, a mixture of the Gaussian and half-Cauchy distributions, which is free of hyperparameters and has Cauchy-like tails. Griffin et al. (2010) considered a class of normal-gamma priors. Polson and Scott (2009) introduced a class of hypergeometric-beta GL priors. Armagan et al. (2011) proposed a prior mixed with Gaussian and generalized beta distributions, encompassing many priors including the Horseshoe as a special case. Armagan et al. (2013a) proposed a generalized double Pareto prior, with a Laplace spike at zero and Student's $t$-like heavy tails. Bhattacharya et al. (2015) proposed a new class of Dirichlet-Laplace (DL) priors, which has optimal posterior concentration and leads to efficient posterior computation. Bhadra et al. (2015) introduced the Horseshoe+ prior for ultra-sparse signal detection, which is an extension of Horseshoe prior, but with faster concentration rate at zero. Each GL prior has its own advantages and disadvantages; in Chapter 3 of this thesis, we will provide a theoretic framework for comparing the GL priors along with our new GL prior.

### 1.2.3 Selection Based on Posteriors

While continuous shrinkage priors exhibit desirable theoretical, computational and empirical properties, they also have their own challenges. Since the posterior probability mass on zero is
always zero, unlike the discrete mixture priors which directly generate sparse estimates, shrinkage priors require additional steps to go from the continuous posterior distribution to a sparse estimate. There are several methods to deal with this. The most common method is to threshold to decide which predictor to be included. For example, Carvalho et al. (2010) described a simple rule for Horseshoe prior that yields a sparse estimate. In addition to thresholding, there are two major approaches: penalized variable selection based on posterior credible regions, and decoupling shrinkage and selection. Bondell and Reich (2012) proposed the penalized credible region variable selection method, which fits the full model under a continuous shrinkage prior, and then selects the sparsest solution within the posterior credible region. Hahn and Carvalho (2015) proposed the decoupling shrinkage and selection method, which uses a loss function combining a posterior summarizer with an explicit parsimony penalty to induce sparse estimator.

1.3 Introduction to Recommender System

In addition to our work on variable selection, we also consider a problem in matrix completion. In the literature, there are three main approaches in recommendation systems: content-based filtering, collaborative filtering, and hybrid recommender systems. Content-based filtering uses the profile of the user’s past preferences and item features to make predictions. It characterizes each user uniquely with no need to match his or her likes and dislikes with other users (Mooney and Roy, 2000). The user profile may correspond to user’s earlier interests, age, gender, occupation, education, or address, and the item profile may concern genre, producer, subject word, actor, or director (Debnath et al., 2008). Content-based algorithms try to recommend items that are similar to the user’s previously “liked” items. Pandora Radio (http://www.pandora.com/) is a popular example of an algorithm that recommends
songs to its customers similar to their previous “likes”. The advantage of a content-based algorithm is that it can recommend previously unrated items to users based on their unique interests. The disadvantage of content-based filtering is that it only recommends items highly correlated the user’s profile, so the user experiences only a very narrow distribution of items; second, the user’s own rating profile is the only factor to influence recommendations, thus requiring quite a lot of user participation, and new users will get inaccurate recommendations (Balabanović and Shoham, 1997). For a more detailed introduction to content-based methods, one can refer to Pazzani and Billsus (2007), Sharanand and Maes (1995), Pazzani and Billsus (1997), Mooney et al. (1998), Basu et al. (1998), Adomavicius and Tuzhilin (2005), Cheng et al. (2010), Naveed et al. (2011), etc.

Instead of basing recommendations on the similarity of items, collaborative filtering makes recommendations based on the similarity of user with other collected users. The famous Netflix Prize (Bennett and Lanning, 2007) motivates lots of algorithms within this framework. Unlike the content-based filtering, since other users’ preferences affect the recommendation, it is possible for a user to receive items totally different from his or her past experiences. The disadvantages of collaborative filtering are as follows. First, it requires a sufficient number of user ratings to start in order to measure the similarities among users. If an item has not been rated by many users, then it is unlikely to be recommended even if this item might have high ratings based on the small number of ratings. Second, recommendation will be poor for a user with unique tastes, as there are no similar users in the database. Therefore, collaborative filtering performs well when there is rich user data and sparse content information. According to Breese et al. (1998), there are two categories of collaborative filtering algorithms. The first is memory-based methods, which makes prediction for the queried user and item using the entire rest of the rating matrix. Related literature includes Altman (1992), Resnick et al. (1994), Sharanand and Maes (1995), Nakamura and Abe (1998), Sarwar et al. (2001), Debnath et al. (2008), etc.
The other category is the model-based methods, which fit a parameterized model to the entire rating matrix, and then make predictions based on the learned model. Related methods can be found in Breese et al. (1998), Billsus and Pazzani (1998), Ungar and Foster (1998), O'Connor and Herlocker (1999), Getoor and Sahami (1999), Goldberg et al. (2001), Sarwar et al. (2002), Hofmann (2003), Marlin (2003), Pavlov and Pennock (2002), Xue et al. (2005), Bayesian classifier Miyahara and Pazzani (2000), Vucetic and Obradovic (2005), Agarwal and Chen (2009), etc. Recently developed methods involve matrix factorization by learning a low-rank model, such as Hill et al. (1995), Shardanand and Maes (1995), singular value decomposition (Billsus and Pazzani, 1998), non-negative matrix factorization (Lee and Seung, 1999), convex optimization (Candès and Recht, 2009), Candés and Plan (2010), Recht (2011), Cai et al. (2010), Bayesian Probabilistic Matrix Factorization (Salakhutdinov and Mnih, 2008), non-linear probabilistic matrix factorization models (Lawrence and Urtasun, 2009), maximum-margin matrix factorization (Srebro et al., 2004; Rennie and Srebro, 2005), nonparametric matrix factorization (Yu et al., 2009), etc. For a more detailed review of classic and state-of-the-art collaborative filtering algorithms, one may refer to Lee et al. (2012).

The two methods above perform unsatisfactorily without sufficient participation of users (Adomavicius and Tuzhilin, 2005). The hybrid approach avoids the shortcomings by combining the two methods, using both the rating matrix and user and item features for prediction. Related methods include Fab’s system (Balabanović and Shoham, 1997), Basu et al. (1998), Condliff et al. (1999), Ansari et al. (2000), Pennock et al. (2000), Popescu et al. (2001), Schein et al. (2002), Melville et al. (2002), Agarwal and Chen (2009), etc.
1.4 Outline

In this dissertation, we focus on Bayesian methods for high-dimensional problems. The rest of the thesis is outlined as follows. Chapter 2 extends the Bayesian credible region variable selection method proposed by Bondell and Reich (2012) by incorporating global-local priors. In this chapter, we also introduce a new method to tune the hyperparameters in prior distributions for linear regression. We propose to choose the hyperparameters to minimize a discrepancy between the induced distribution on R-square and a prespecified target distribution. Chapter 3 proposes a new class of global-local shrinkage priors for linear regression, the R-square induced Dirichlet Decomposition (R2-D2) prior. The prior is induced by a Beta prior on the coefficient of determination, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. We demonstrate both theoretically and empirically the advantages of the R2-D2 prior over a number of common shrinkage priors, including the Horseshoe, Horseshoe+, and Dirichlet-Laplace priors.

Chapter 4 is mostly self-contained. Driven by a movie rating system, this chapter proposes a full Bayesian hybrid recommender system, using a shared-variable approach for matrix completion to coerce informative missingness into a regression-based model to improve prediction performance. We apply our model on a movie rating system, particularly the MovieLens data. Our method illustrates significant improvement in prediction by adding informative missingness into the model.
Chapter 2

High-dimensional Variable Selection via Penalized Credible Regions with Global-Local Shrinkage Priors

2.1 Introduction

High dimensional data has become increasingly common in all fields. Linear regression is a standard and intuitive way to model dependency in high dimensional data. Consider the linear regression model:

\[ Y = X\beta + \varepsilon \]  

(2.1)

where \( X \) is the \( n \times p \) high-dimensional set of covariates, \( Y \) is the \( n \) scalar responses, \( \beta = (\beta_1, \cdots, \beta_p) \) is the \( p \)-dimensional coefficient vector, and \( \varepsilon \) is the error term assumed to have \( E(\varepsilon) = 0 \) and \( \text{Var}(\varepsilon) = \sigma^2 I_n \). Ordinary least squares is not feasible when the number of predictors \( p \) is larger than the sample size \( n \). Variable selection is necessary to reduce the large number
of candidate predictors. The classical variable selection methods include subset selection, criteria such as AIC (Akaike, 1998) and BIC (Schwarz et al., 1978), and penalized methods such as the least absolute shrinkage and selection operator (Lasso; Tibshirani, 1996), smoothly clipped absolute deviation (Fan and Li, 2001, SCAD;), the elastic net (Zou and Hastie, 2005), adaptive Lasso (Zou, 2006), the Dantzig selector (Candes and Tao, 2007), and octagonal shrinkage and clustering algorithm for regression (OSCAR; Bondell and Reich, 2008).

In the Bayesian framework, approaches for variable selection include: stochastic search variable selection (SSVS) (George and McCulloch, 1993), Bayesian regularization (Park and Casella, 2008; Li et al., 2010; Polson et al., 2013; Leng et al., 2014), empirical Bayes variable selection (George and Foster, 2000), spike and slab variable selection (Ishwaran and Rao, 2005), and global-local (GL) shrinkage priors. Those traditional Bayesian methods conduct variable selection either relying on the calculation of posterior inclusion probabilities for each predictor or each possible model, or a choice of posterior threshold.

Typical global-local shrinkage priors are represented as the class of global-local scale mixtures of normals (Polson and Scott, 2010),

\[
\beta_j \sim N(0, w\xi_j), \quad \xi_j \sim \pi(\xi_j), \quad (w, \sigma^2) \sim \pi(w, \sigma^2),
\]

where \(w\) controls the global shrinkage towards the origin, while \(\xi_j\) allows local deviations of shrinkage. Various options of shrinkage priors for \(\beta\), include normal-gamma (Griffin et al., 2010), Horseshoe prior (Carvalho et al., 2009, 2010), generalized double Pareto prior (Armagan et al., 2013a), Dirichlet-Laplace (DL) priors (Bhattacharya et al., 2015), and others that can be represented as (2.2). The GL shrinkage priors usually shrink small coefficients greatly due to a tight peak at zero, and rarely shrink large coefficients due to the heavy tails. It has been shown that GL shrinkage priors have improved posterior concentrations (Bhattacharya
et al., 2015). However, the shrinkage prior itself would not lead to variable selection, and to go further, some rules need to be set on the posteriors.

Bondell and Reich (2012) proposed a Bayesian variable selection method only based on posterior credible regions. However, the implementation and results of that paper depended on the use of conjugate normal priors. Due to the improved concentration, incorporating the global-local shrinkage priors into this framework can perform better, both in theory and practice. We show posterior consistency for the DL prior in this regression setting, along with selection consistency.

Another difficulty in high dimensional data is the choice of hyperparameters, which can highly affect the results. In this chapter, we also propose an intuitive default method to tune the hyperparameters in the prior distributions. By minimizing a discrepancy between the induced distribution of $R^2$ from the prior and the desired distribution (Beta distribution by default), one gets a default choice of hyperparameter value. For the choice of normal priors, the hyperparameter that minimizes the Kullback-Leibler (KL) divergence between the distributions is shown to have a closed form solution.

The remainder of the chapter is organized as follows. Section 2.2 reviews the penalized credible region variable selection method. Section 2.3 details the proposed method which combines shrinkage priors and penalized credible region variable selection. Section 2.4 presents the posterior consistency for the choice of shrinkage priors, as well as the asymptotic behavior of the selection consistency for diverging $p$. Section 2.5 discusses a default method to tune the hyperparameters in the prior distributions based on the induced prior distribution on $R^2$. Section 2.6 reports the simulation results, and Section 2.7 gives the analysis of a real-time PCR dataset. All proofs are given in Section 2.9.
2.2 Background

Bondell and Reich (2012) proposed a penalized regression method based on Bayesian credible regions. First, the full model is fit using all predictors with a continuous prior. Then based on the posterior distribution, a sequence of joint credible regions are constructed, within which, one searches for the sparsest solution. The choice of a conjugate normal prior of

$$\beta | \sigma^2, \gamma \sim N(0, \sigma^2 / \gamma I_p)$$  \hspace{1cm} (2.3)

is used, where $\sigma^2$ is the error variance term as in (2.1), and $\gamma$ is the ratio of prior precision to error precision. The variance, $\sigma^2$, is often given a diffuse inverse Gamma prior, while $\gamma$ is the hyperparameter which is either chosen to be fixed or given a Gamma hyperprior.

The credible region is to find $\tilde{\beta}$, such that

$$\tilde{\beta} = \arg\min_{\beta} \|\beta\|_0 \text{ subject to } \beta \in C_{\alpha},$$  \hspace{1cm} (2.4)

where $\|\beta\|_0$ is the $L_0$ norm of $\beta$, i.e., the number of nonzero elements, and $C_{\alpha}$ is the $(1 - \alpha) \times 100\%$ posterior credible regions based on the particular prior distributions. The use of elliptical posterior credible regions yields the form $C_{\alpha} = \{\beta : (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) \leq c_{\alpha}\}$, for some nonnegative $c_{\alpha}$, where $\hat{\beta}$ and $\Sigma$ are the posterior mean and covariance respectively. Then by replacing the $L_0$ penalization in (2.4) with a smooth homotopy between $L_0$ and $L_1$ proposed by Lv and Fan (2009) and linear approximation, the optimization problem in (2.4) becomes

$$\tilde{\beta} = \arg\min_{\beta} (\beta - \hat{\beta})^T \Sigma^{-1} (\beta - \hat{\beta}) + \lambda_{\alpha} \sum_{j=1}^{p} |\hat{\beta}_j| - 2 |\beta_j|,$$  \hspace{1cm} (2.5)

where there exists a one-to-one correspondence between $c_{\alpha}$ and $\lambda_{\alpha}$. The sequence of solutions
to (2.5) can be directly accomplished by plugging in the posterior mean and covariance and using the LARS algorithm (Efron et al., 2004).

2.3 Penalized Credible Regions with Global-Local Shrinkage Priors

2.3.1 Motivation

Global-local shrinkage priors produce a posterior distribution with good empirical and theoretical properties. Compared to the usual normal prior, GL priors concentrate more along the regions with zero parameters. The penalized credible region variable selection approach separates model fitting and variable selection. So it seems natural to fit the model under a GL shrinkage prior, and then conduct variable selection through the penalized credible region method. The motivation is that within the same credible region level, GL shrinkage priors would lead to more concentrated posteriors, thus having better performance for variable selection, by finding sparse solutions more easily.

Although GL shrinkage priors would not lead to elliptical posterior distributions, valid credible regions can still be constructed using elliptical contours. These would no longer be the high density regions, but would remain valid regions. Elliptical contours would also be reasonable approximations to the high density regions, at least around the largest mode. Thus, the penalized credible region selection method can be feasibly performed by plugging the posterior mean and covariance matrix into the optimization algorithm (2.5). So given any GL prior, once MCMC steps produce the posterior samples, the sample mean, $\hat{\beta}$, and sample covariance, $\Sigma$, would hence be obtained, then variable selection can be performed through the penalized credible region method. In this chapter, we modify the Dirichlet-Laplace (DL) prior to implement
in the regression setting. We also consider the Laplace prior, also referred as Bayesian Lasso, described in Park and Casella (2008) and Hans (2010), as

\[ \beta_j \sim \text{DE}(\sigma/\lambda) \ (j = 1 \cdots, p), \quad (2.6) \]

where \( \lambda \) is the Lasso parameter, controlling the global shrinkage.

### 2.3.2 Dirichlet-Laplace Priors

For the normal mean model, Bhattacharya et al. (2015) proposed a new class of Dirichlet-Laplace (DL) shrinkage priors, possessing the optimal posterior concentration property. We construct the generalization of the DL priors for the linear regression model. The proposed hierarchical DL prior is as follows: for \( j = 1, \cdots, p, \)

\[
\begin{align*}
\beta_j | \sigma, \phi_j, \tau & \sim \text{DE}(\sigma \phi_j \tau), \\
(\phi_1, \cdots, \phi_p) & \sim \text{Dir}(a, \cdots, a), \\
\tau & \sim \text{Ga}(pa, 1/2).
\end{align*}
\quad (2.7)
\]

where \( \text{DE}(b) \) denotes a zero mean Laplace kernel with density \( f(y) = (2b)^{-1} \exp\{-|y|/b\} \) for \( y \in \mathbb{R} \), \( \text{Dir}(a, \cdots, a) \) is the Dirichlet distribution with concentration vector \((a, \cdots, a)\), and \( \text{Ga}(pa, 1/2) \) denotes a Gamma distribution with shape \( pa \) and rate \( 1/2 \). Here, small values of \( a \) would lead most of \((\phi_1, \cdots, \phi_p)\) to be close to zero and only few of them nonzero; while large values allow less singularity at zero, thus controlling the sparsity of regression coefficients. The \( \phi_j \)'s are the local scales, allowing deviations in the degree of shrinkage. As pointed out in Bhattacharya et al. (2015), \( \tau \) controls global shrinkage towards the origin and to some extent determines the tail behaviors of the marginal distribution of \( \beta_j \)'s. We also assume a common
prior on the variance term \( \sigma^2 \), \( \text{IG}(a_1, b_1) \), the inverse Gamma distribution with shape \( a_1 \) and scale \( b_1 \).

### 2.3.3 Computation of Posteriors

For posterior computation, the Gibbs sampling steps proposed in Bhattacharya et al. (2015) can be modified to accommodate the linear regression model. The DL prior (2.7) can be equivalently denoted as

\[
\beta_j | \sigma^2, \phi_j, \psi_j, \tau \sim N(0, \sigma^2 \psi^2 \phi^2 \tau^2),
\]

\[
\psi_j \sim \text{Exp}(1/2),
\]

\[
(\phi_1, \cdots, \phi_p) \sim \text{Dir}(a, \cdots, a),
\]

\[
\tau \sim \text{Ga}(pa, 1/2),
\]

where \( \text{Exp}(\cdot) \) is the usual exponential distribution. Note that DL prior is also a global-local shrinkage prior as it is a particular form of (2.2). Gibbs sampling steps would be obtained based on (2.8). Since \( \pi(\psi, \phi, \tau | \beta, \sigma^2) = \pi(\psi | \phi, \tau, \beta, \sigma^2) \pi(\tau | \phi, \beta, \sigma^2) \pi(\phi | \beta, \sigma^2) \), and the joint posterior of \( (\psi, \phi, \tau) \) is independent of \( y \) conditionally on \( \beta \) and \( \sigma^2 \), so the steps to draw posteriors steps are as follows: (i) \( \sigma^2 | \beta, \psi, \phi, \tau, y \), (ii) \( \beta | \psi, \phi, \tau, \sigma^2, y \), (iii) \( \psi | \phi, \tau, \beta, \sigma^2 \), (iv) \( \tau | \phi, \beta, \sigma^2 \), (v) \( \phi | \beta, \sigma^2 \). The derivation is similar as in Bhattacharya et al. (2015), hence omitted here.

The parameterization of the three-parameter generalized inverse Gaussian (giG) distribution, \( Y \sim \text{giG}(\chi, \rho, \lambda_0) \), means the density of \( Y \) is \( f(y) \propto y^{\lambda_0-1} \exp\{-0.5(\rho y + \chi/y)\} \) for \( y > 0 \). Then the summary of the Gibbs sampling steps are as below:

(i) Sample \( \sigma^2 | \beta, \psi, \phi, \tau, y \). Draw \( \sigma^2 \) from an inverse Gamma distribution, \( \text{IG}(a_1 + (n + p)/2, b_1 + (\beta^T S^{-1} \beta + (Y - X\beta)^T (Y - X\beta))/2) \), where \( S = \text{diag}(\psi_1 \phi_1^2 \tau^2, \cdots, \psi_p \phi_p^2 \tau^2) \).
(ii) Sample $\beta | \psi, \phi, \tau, \sigma^2, y$. Draw $\beta$ from a $N(\mu, \sigma^2 V)$, where $V = (X^T X + S^{-1})^{-1}$ with the same $S$ as above, and $\mu = V X^T Y = (X^T X + S^{-1})^{-1} (X^T Y)$.

(iii) Sample $\psi_j | \phi, \tau, \beta, \sigma^2$. First draw $\psi_j^{-1} | \phi_j, \tau, \beta, \sigma^2, j = 1, \cdots, p$, independently from the distribution InvGaussian$(\mu_j = \sigma \phi_j \tau / |\beta_j|, \lambda_0 = 1)$, where InvGaussian$(\mu, \lambda_0)$ denotes the inverse Gaussian with density $f(y) = \sqrt{\lambda_0 / (2 \pi y^3)} \exp\{ -\lambda_0 (y - \mu)^2 / (2\mu^2 y) \}$ for $y > 0$. Then take the reciprocal to get the draws of $\psi_j (j = 1, \cdots, p)$.

(iv) Sample $\tau | \phi, \beta, \sigma^2$. Draw $\tau$ from a giG$(\chi = 2 \sum_{j=1}^p |\beta_j| / (\phi_j \sigma), \rho = 1, \lambda_0 = pa - p)$.

(v) Sample $\phi_j | \beta, \sigma^2$. Draw $T_1, \cdots, T_p$ independently with $T_j \sim$ giG$(\chi = 2 |\beta_j| / \sigma, \rho = 1, \lambda_0 = a - 1)$, then set $\phi_j = T_j / T$ where $T = \sum_{j=1}^p T_j$.

2.4 Asymptotic Theory

In this section, we first study the posterior properties of the normal and DL prior, when both $n$ and $p_n$ go to infinity, and further investigate the selection consistency of the penalized variable selection method. Assume the true regression parameter is $\beta_0^n$, and the estimated regression parameter is $\hat{\beta}_n$. Denote the true set of non-zero coefficients is $A_0^n = \{ j : \beta_{nj}^0 \neq 0, j = 1, \cdots, p_n \}$, and the estimated set of non-zero coefficients is $A_n = \{ j : \beta_{nj} \neq 0, j = 1, \cdots, p_n \}$. Also let $q_n = |A_0^n|$ denote the number of predictors with nonzero true coefficients.

As $n \to \infty$, consider the sequence of credible sets of the form $\{ \beta_n : (\beta_n - \hat{\beta}_n)^T \Sigma_n^{-1} (\beta_n - \hat{\beta}_n) \leq c_n \}$, where $\hat{\beta}_n$ and $\Sigma_n$ are the posterior mean and covariance matrix respectively, and $c_n$ is a sequence of non-negative constants. Let $\Gamma_n$ denote the $p_n \times p_n$ matrix whose columns are eigenvectors of $X^T X / n$ ordered by decreasing eigenvalues, i.e., $d_1 \geq d_2 \geq \cdots \geq d_{p_n} \geq 0$. Then $X^T X / n = \Gamma_n D_n \Gamma_n^T$ where $D_n = \text{diag}\{d_1, \cdots, d_{p_n}\}$.

Assume the following regularity conditions throughout.
The error terms $\varepsilon_i$, $i = 1, \cdots, n$, are independent and identically distributed (i.i.d.) with mean zero and finite variance $\sigma^2$;

(A2) $p_n = o(n)$;

(A3) $0 < d_{\min} < \liminf_{n \to \infty} d_{p_n} \leq \limsup_{n \to \infty} d_{1} < d_{\max} < \infty$, where $d_{\min}$ and $d_{\max}$ are fixed;

(A4) $\limsup_{j=1,\cdots,p_n} |\beta_{nj}| < \infty$.

2.4.1 Posterior Consistency of the Normal and DL Priors

Armagan et al. (2013b) investigates the asymptotic behavior of posterior distributions of regression coefficients in the linear regression model (2.1) as $p$ grows with $n$. They prove the posterior consistency for a variety of priors, including the Laplace prior, Student’s $t$ prior, generalized double Pareto prior, and the Horseshoe-like priors. By definition, posterior consistency implies that the posterior distribution of $\beta_n$ converges in probability to $\beta_0$, i.e., for any $\epsilon > 0$, $P(\beta_n : ||\beta_n - \beta_0|| > \epsilon | Y) \to 0$ as $p_n, n \to \infty$. In this section, we show that the normal and Dirichlet-Laplace prior also yield consistent posteriors. However, the DL prior can obtain posterior consistency under weaker conditions on the signal.

**Theorem 1.** Under assumptions (A1)-(A4), if $q_n = o\{n^{1-\rho}/(p_n(\log n)^2)\}$ for $\rho \in (0, 1)$, and $\sqrt{\sigma^2/\gamma_n} = C/\sqrt{p_n n^{\rho/2} \log n}$ for finite $C > 0$, the normal prior (2.3) yields a consistent posterior.

**Theorem 2.** Under assumptions (A1)-(A4), if $q_n = o(n/ \log n)$, and $a = C/(p_n n^{\rho} \log n)$ for any finite $\rho > 0$ and finite $C > 0$, the Dirichlet-Laplace prior (2.7) yields a consistent posterior.

Note that the difference in the above two theorems is the number of nonzero components, i.e., $q_n$. As $n/ \log n > n^{1-\rho}/(p_n(\log n)^2)$, the Dirichlet-Laplace prior leads to posterior con-
consistency in a much broader domain, compared to the normal prior as well as compared to the Laplace prior whose posterior consistency is given in Theorem 2 in Armagan et al. (2013b). This strengthens the justification for replacing the normal prior with the DL prior theoretically.

2.4.2 Selection Consistency of Penalized Credible Regions

Bondell and Reich (2012) has shown that when $p$ is fixed and $\beta$ is given the normal prior in (2.3), the penalized credible region method is consistent in variable selection. In this chapter, we show that posterior consistency of a global-local shrinkage prior also yields consistency in variable selection under the case of $p_n \to \infty$.

Theorem 3. Under assumptions (A1) - (A4), given the normal prior in (2.3), if $c_n \to \infty$, $c_n/(np_n) \to 0$, and the prior precision, $\gamma_n = o(n)$, then the penalized credible region method is consistent in variable selection, i.e. $P(A_n = A_{0n}) \to 1$.

The proof is given in the Appendix. The selection consistency allows us to expect that the true model is contained in the credible regions with high probability, when the number of predictors increases together with the sample size. Such selection consistency is obtained under the normal prior. However, as reviewed in Section 2.1, since the GL shrinkage priors can be expressed as a scale mixture of normals, as long as the posterior distribution of the precision is $o(n)$ with probability 1, then the result can be directly applied to the GL shrinkage prior.

Theorem 4. Under assumptions (A1) - (A4), given any global-local shrinkage prior represented as (2.2), when the conditions of posterior consistency are satisfied, then the posterior distribution of the precision is $o(n)$ with probability 1 as $n \to \infty$. Furthermore, if $c_n \to \infty$ and $c_n/(np_n) \to 0$, then the penalized credible region method with the particular shrinkage prior is consistent in variable selection, i.e. $P(A_n = A_{0n}) \to 1$. 
So given the conditions of posterior consistency for the global-local shrinkage prior, we automatically get the selection consistency of the credible region method. For example, for the DL prior in (2.7), we have the following result.

**Corollary 1.** Under assumptions (A1) - (A4), given the DL prior in (2.7), if \( q_n = o(n / \log n) \), \( a = C/(p_n n^\rho \log n) \) for any finite \( \rho > 0 \) and finite \( C > 0 \), \( c_n \to \infty \) and \( c_n/(np_n) \to 0 \), then the penalized credible region method is consistent in variable selection, i.e. \( P(A_n = A_n^0) \to 1 \).

### 2.5 Tuning Hyperparameters

The value of hyperparameters in the prior distribution plays an important role in the posteriors. For example, in the normal prior (2.3), \( \gamma \) is the hyperparameter, whose value controls the degree of shrinkage. This is often chosen to be fixed at a “large” value or given a hyperprior. However, the choice of the “large” value affects the results, as does the choice of hyperprior such as a gamma prior, particularly in the high dimensional case. Also, in the DL prior (2.7), the choice of \( a \) is critical. If \( a \) is too small, then the DL prior would shrink each dimension of \( \beta \) towards zero; while, if \( a \) is too large, there would be no strong concentration around the origin. Instead of fixing \( a \), a discrete uniform prior can be given on \( a \) supported on some interval (for example, \([1/\max(n, p), 1/2]\)), with several support points on the interval. However, introducing the hyperprior for the hyperparameters will not only arise new values to tune, but also increase the complexity of the MCMC sampling. In practice, although the specification of a \( p \)-dimensional prior on \( \beta \) may be difficult, some prior information on a univariate function may be easier. The motivation is to incorporate such prior information of the one-dimensional function into the priors on the \( p \)-dimensional \( \beta \).

In this chapter, we propose an intuitive way to tune the values of hyperparameters, by incorporating a prior on \( R^2 \) (the coefficient of determination). Practically, a scientist may have
information on $R^2$ from previous experiments, and this can be coerced into say a Beta$(a, b)$ distribution. Without any prior information for $R^2$, a uniform prior, Beta$(1, 1)$, may be used as default. In this way, tuning hyperparameters is equivalent to searching for the hyperparameter which leads to the induced distribution of $R^2$ closest to the desired distribution.

For the linear regression model (2.1), the population form can be represented as $y = x^T \beta + \varepsilon$, with $x$ independent of $\varepsilon$. Let $\sigma_y^2$ be the marginal variance of $y$ and $\sigma^2$ be the variance of the random error term. The definition of the POPULATION $R^2$ is given by:

$$pop R^2 = 1 - \frac{\sigma^2}{\sigma_y^2},$$

which is the proportion of the variation of $y$ in the population explained by the independent variables. Furthermore, for fixed $\beta$, it follows that $\sigma_y^2 = \beta^T \text{Cov}(x) \beta + \sigma^2$. Assume $E(x) = 0$, then we can estimate Cov$(x)$ by $X^T X/n$. So $R^2$ as a function of $\beta$ and $\sigma^2$ is given by $R^2 = 1 - \sigma^2 / (\beta^T X^T X \beta / n + \sigma^2)$. Given that the form of prior distributions considered includes $\sigma$ in the scale, it follows that $\beta = \sigma \eta$ for $\eta$ having the distribution of the prior fixed with $\sigma^2 = 1$.

Hence

$$R^2 = 1 - \frac{1}{1 + \eta^T X^T X \eta / n}.$$  \hspace{1cm} (2.9)

For a specified prior on $\eta$, the induced distribution of $R^2$ can be derived based on (2.9). Then the hyperparameters which yield the induced distribution of $R^2$ closest to the desired distribution is the tuned value.

In practice, one can consider a grid of possible values of the hyperparameters. For each value, draw a vector $\eta$. This is converted to a draw of $R^2$. Given this hyperparameter, a comparison between the sample of $R^2$ and the desired distribution is performed, for example, a
Kolmogorov-Smirnov (KS) test. The best fit is then chosen. The whole tuning process only involves the prior distributions, no MCMC sampling, thus avoiding comprehensive computing.

However, given a specific prior for $\beta$, based on (2.9), the exact induced distribution of $R^2$ can be derived, which relies on the value of hyperparameters. By minimizing the Kullback-Liebler directed divergence between such distribution and the desired distribution (Beta distribution by default), a default hyperparameter value can be found. For continuous random variables with density function $f_1$ and $f_2$, the KL divergence is defined as

$$D(f_1|f_2) = \int_{-\infty}^{\infty} f_1(x) \log(f_1(x)/f_2(x)) \, dx.$$ 

For the choice of normal priors, the following theorem shows that there is a closed form solution for the hyperparameter to minimize the KL divergence for large $p$.

**Theorem 5.** For the normal prior in (2.3), to minimize the KL directed divergence between the induced distribution of $R^2$ and the Beta$(a, b)$ distribution, as $p \to \infty$, the hyperparameter, $\gamma$, is chosen to be

$$\gamma = (A + \sqrt{B})^{1/3} + (A - \sqrt{B})^{1/3} - P/3,$$

where $P = (2a - b) \sum_{j=1}^{p} d_j / a$, $Q = 2(a + b) \sum_{j=1}^{p} d_j^2 / a + (a - 2b)(\sum_{j=1}^{p} d_j)^2 / a$, $R = -b(\sum_{j=1}^{p} d_j)^3 / a$, $C = P^2 / 9 - Q/3$, $A = PQ / 6 - P^3 / 27 - R / 2$, $B = A^2 - C^3 \geq 0$, and $d_1, \cdots, d_p$ denote the eigenvalues of $X^T X / n$.

In theory, for other continuous priors, one can derive the optimal hyperparameters similarly. However, sometimes the calculation can be quite complex. In this case, the simulation-based approach discussed earlier can be implemented. However, since GL priors can be represented as mixture normal priors (see Section 2.1), by matching its prior precision with that of the normal prior, the derived default solution as shown in Theorem 5 can offer an intuitive idea for the hyperparameter values in the GL shrinkage priors.
2.6 Simulation Results

2.6.1 Comparisons of Different Priors

To compare the performance of the penalized credible region variable selection method using different shrinkage priors, including the normal prior (2.3), Laplace prior (2.6), and DL prior (2.7), a simulation study is conducted.

We use a similar simulation setup as in Bondell and Reich (2012). In each setting, 200 datasets are simulated from the linear model (2.1) with $\sigma^2 = 1$, sample size $n = 60$, and the number of predictors $p$ varying in $\{50, 500, 1000\}$. To represent different correlation settings, $X_{ij}$ are generated from standard normal distribution, and the correlation between $x_{ij1}$ and $x_{ij2}$ is $\rho^{\|j_1-j_2\|}$, with $\rho = 0.5$ and 0.9. The true coefficient $\beta$ is $(0^T_{p-40}, B_1^T, 0_{20}^T, B_2^T, 0^T_{p-40})^T$ for $p \in \{50, 500, 1000\}$ in which $0_k$ represents the $k$-dimensional zero vector, $B_1$ and $B_2$ are both 5-dimensional vector generated component-wise and uniform from $(0, 1)$. For each case of shrinkage prior, the posterior mean and covariance can be obtained from the Gibbs samplers, and then plugged into the optimization algorithm (2.5) of the penalized credible region method to implement the variable selection.

For each method, the induced ordering of the predictors are created. We consider the resulting model at each ordering step to measure the performance. For each step on the ordering, true positives (TP) are defined as those selected variables which also appear in the true model. False positives (FP) are those selected variables which also do not appear in the true model. True negatives (TN) correspond to those not selected variables which are not in the true model. False negatives (FN) refer to variables which are not selected in the model, but indeed are in the true model. The Receiver-Operating Characteristic (ROC) curve plots the false positive rate (FPR or 1-Specificity) on the x-axis and the true positive rate (TPR or Sensitivity) on the y-axis, where FPR is the fraction of FP’s of the fitted model in the total number of irrelevant variables.
in the true model, and TPR is the fraction of TP’s of the fitted model in the total number of important variables in the true model. The Precision-Recall (PRC) curve plots the precision on the y-axis, and the Recall (or TPR or Sensitivity) on x-axis, where precision is the ratio of true positives to the total declared positive number.

The compared credible set methods are listed as below:

- **Method “Normal_hyper”**, refers to the normal prior, with “non-informative” hyperparameters, i.e., $N(0, \sigma^2_b)$ is the prior for $\beta$, and $IG(0.001, 0.001)$ prior is given for $\sigma^2_b$.

- **Method “Normal_tune”**, refers to the normal prior (2.3), where $\gamma$ is tuned through the $R^2$ method introduced in Section 2.5, with a target of uniform distribution.

- **Method “Laplace_hyper”**, means Laplace prior (2.6), with $\lambda$ given a $Ga(1, 1)$ prior.

- **Method “Laplace_tune”**, means Laplace prior (2.6), and $\lambda$ is tuned through the $R^2$ method introduced in Section 2.5, with a target of uniform distribution.

- **Method “DL_hyper”** is the DL prior (2.7), in which $a$ is given a discrete uniform prior supported on the interval $[1/\max(n, p), 1/2]$ with 1000 support points in this interval.

- **Method “DL_tune”** is the DL prior (2.7), in which $a$ is tuned through the $R^2$ method introduced in Section 2.5, with a target of uniform distribution.

In all above cases, the variance term $\sigma^2$ is given an $IG(0.001, 0.001)$ prior. In addition, we show the results from using the Lasso (Tibshirani, 1996) fit via the LARS algorithm (Efron et al., 2004).

For the above priors (normal, Laplace and DL), we ran the MCMC chain (Gibbs sampling) for 15,000 iterations, with the first 5,000 for burn-in. Posterior mean and covariance were calculated based on the 10,000 samples, which were then plugged into the penalized credible
Table 2.1: Mean area under the ROC Curve and the PRC curve for \( p = 50, n = 60 \), based on 200 datasets with standard errors in parentheses.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho = 0.5 )</td>
<td>( \rho = 0.9 )</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.900 (0.0047)</td>
<td>0.815 (0.0052)</td>
</tr>
<tr>
<td>Normal_hyper</td>
<td>0.909 (0.0048)</td>
<td>0.899 (0.0041)</td>
</tr>
<tr>
<td>Normal_tune</td>
<td>0.949 (0.0037)</td>
<td>0.978 (0.0020)</td>
</tr>
<tr>
<td>Laplace_hyper</td>
<td>0.890 (0.0049)</td>
<td>0.859 (0.0052)</td>
</tr>
<tr>
<td>Laplace_tune</td>
<td>0.942 (0.0040)</td>
<td>0.976 (0.0020)</td>
</tr>
<tr>
<td>DL_hyper</td>
<td>0.917 (0.0044)</td>
<td>0.908 (0.0044)</td>
</tr>
<tr>
<td>DL_tune</td>
<td>0.939 (0.0039)</td>
<td>0.945 (0.0032)</td>
</tr>
</tbody>
</table>

interval optimization algorithm (2.5), to conduct variable selection. Table 2.1 gives the mean and standard error for the area under the ROC and PRC curve for \( p = 50 \) with \( \rho \in \{0.5, 0.9\} \). In addition, Figure 2.1 plots the mean ROC and PRC curves of the 200 datasets for the selected above methods to compare. Table 2.2 and Figure 2.2 give the results for the \( p = 500 \) case. Table 2.3 and Figure 2.3 show the results for the \( p = 1000 \) case. Since the Lasso estimator can select at most \( \min\{n, p\} \) predictors, when \( p = 500 \) or \( 1000 \), the ROC and PRC curves cannot be fully constructed. So the area under the curves cannot be compared directly for Lasso with other methods, which are omitted in Table 2.2 and 2.3, but partial ROC and PRC curves can still be plotted, which are shown in Figure 2.2 and 2.3.

On the one hand, in terms of whether given a hyperprior for the hyperparameter or tuning hyperparameters through the \( R^2 \) method proposed in Section 2.5 would lead to better posterior performance, one might compare each “*.hyper” and “*.tune” pair in Table 2.1, 2.2 and 2.3. In general, for all three priors, the tuning method leads to significantly better posterior performance than the hyperprior method in all simulation setups.

On the other hand, in terms of comparing performance of different priors applied on the
penalized credible region variable selection, combining both the tables and figures, we have the following findings. When considering the Precision-Recall in particular, the DL and Laplace priors outperform the normal prior and Lasso. This is particularly true if the hyperparameters in them are tuned via a uniform distribution on $R^2$. We note that when there are only a few true
Table 2.2: Mean area under the ROC Curve and the PRC curve for $p = 500$, $n = 60$, based on 200 datasets with standard errors in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 0.5$</td>
<td>$\rho = 0.9$</td>
</tr>
<tr>
<td>Lasso</td>
<td>- -</td>
<td>0.550 (0.0087)</td>
</tr>
<tr>
<td>Normal_hyper</td>
<td>0.948 (0.0031)</td>
<td>0.990 (0.0013)</td>
</tr>
<tr>
<td>Normal_tune</td>
<td>0.950 (0.0029)</td>
<td>0.992 (0.0007)</td>
</tr>
<tr>
<td>Laplace_hyper</td>
<td>0.937 (0.0030)</td>
<td>0.969 (0.0020)</td>
</tr>
<tr>
<td>Laplace_tune</td>
<td>0.959 (0.0027)</td>
<td>0.995 (0.0004)</td>
</tr>
<tr>
<td>DL_hyper</td>
<td>0.927 (0.0038)</td>
<td>0.908 (0.0047)</td>
</tr>
<tr>
<td>DL_tune</td>
<td>0.949 (0.0027)</td>
<td>0.970 (0.0025)</td>
</tr>
</tbody>
</table>

Table 2.3: Mean area under the ROC Curve and the PRC curve for $p = 1000$, $n = 60$, based on 200 datasets with standard errors in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho = 0.5$</td>
<td>$\rho = 0.9$</td>
</tr>
<tr>
<td>Lasso</td>
<td>- -</td>
<td>0.507 (0.0093)</td>
</tr>
<tr>
<td>Normal_hyper</td>
<td>0.942 (0.0039)</td>
<td>0.992 (0.0018)</td>
</tr>
<tr>
<td>Normal_tune</td>
<td>0.943 (0.0039)</td>
<td>0.991 (0.0018)</td>
</tr>
<tr>
<td>Laplace_hyper</td>
<td>0.914 (0.0041)</td>
<td>0.968 (0.0021)</td>
</tr>
<tr>
<td>Laplace_tune</td>
<td>0.951 (0.0038)</td>
<td>0.994 (0.0012)</td>
</tr>
<tr>
<td>DL_hyper</td>
<td>0.931 (0.0040)</td>
<td>0.943 (0.0034)</td>
</tr>
<tr>
<td>DL_tune</td>
<td>0.925 (0.0045)</td>
<td>0.967 (0.0025)</td>
</tr>
</tbody>
</table>

and many unimportant variables, the Precision-Recall curve is a more appropriate measure than the ROC curve. For example, when $p = 1000$, in both $\rho = 0.5$ and 0.9 cases, in Figure 2.3, the PRC curve shows that the DL prior is significantly better than the normal prior; the ROC curve of the normal prior goes higher when FPR (or 1-Specificity) is large, however, when FPR is small (which is of more interest), DL prior still leads to significantly larger sensitivity than the normal prior. Overall, the DL prior outperforms the normal prior, as does the Laplace prior.
Figure 2.2: Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 500$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.

2.6.2 Additional Simulations on Hyperparameter Tuning

To examine the role of $a$ in the DL prior, additional simulations were conducted. Table 2.4 gives the average squared error for the posterior mean based on the 200 same datasets as Section 2.6.1, for the DL priors with $a$ fixed at $1/2$, $1/n$, and $1/p$. The results show that when $p$ is
Figure 2.3: Plot of mean ROC and PRC curves when $\rho = 0.5$ and $\rho = 0.9$, over the 200 datasets for $p = 1000$ predictors, $n = 60$ observations. The left column is the ROC curve, the right column is the PRC curve.

large or there is strong correlation in the dataset, $a = 1/n$ is better than $a = 1/2$. When $p$ is small and there is only moderate correlation for the data, $a = 1/2$ is recommended. Since the performance of different values of $a$ varies relying on the dimension of predictors and the correlation structure of the predictors, fixing $a$ is difficult. Thus either giving a hyperprior for
or using the $R^2$ method proposed in Section 2.5 to tune $a$ is suggested.

Table 2.4: Average squared error for the posterior mean, given Dirichlet-Laplace prior with $a$ fixed at $\frac{1}{2}$, $\frac{1}{n}$ and $\frac{1}{p}$, based on 200 datasets with standard errors in parentheses.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$p = 50$</th>
<th>$p = 500$</th>
<th>$p = 1000$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\frac{1}{2}$ $\frac{1}{n}$ $\frac{1}{p}$</td>
<td>$\frac{1}{2}$ $\frac{1}{n}$ $\frac{1}{p}$</td>
<td>$\frac{1}{2}$ $\frac{1}{n}$ $\frac{1}{p}$</td>
</tr>
<tr>
<td>$\rho = 0.5$</td>
<td>0.772 0.877 0.874 0.0234 0.0325 0.0329</td>
<td>1.292 1.400 1.953 0.0421 0.0519 0.0576</td>
<td>1.470 1.434 2.196 0.0451 0.1070 0.1196</td>
</tr>
<tr>
<td>$\rho = 0.9$</td>
<td>1.989 1.751 1.715 0.0559 0.0737 0.0739</td>
<td>2.193 2.142 2.546 0.0767 0.0981 0.1180</td>
<td>2.299 2.247 2.426 0.1101 0.1178 0.1186</td>
</tr>
</tbody>
</table>

Furthermore, to verify Theorem 5 described in Section 2.5, additional calculations were performed. For each of the above 200 datasets, “Normal_tune” returns a “best” tuned $\gamma$ through conducting the practical procedures as introduced in Section 2.5, and we name it as “Tuned”. Also, by Theorem 5, the theoretic “best” $\gamma$ can be derived based on the eigenvalues of $X^TX/n$ for each dataset, and we name it as “Derived”. In addition, for each of the above 200 datasets, the design matrix $X$ is generated from a multivariate normal distribution with specific and fixed covariance structure. So the eigenvalues of such true covariance matrix, instead of $X^TX/n$, can be used to derive the theoretic “best” $\gamma$, and we name it as “Theoretic” value. Table 2.5 gives the “Theoretic” value, and the mean of “Derived” and “Tuned” value together with the standard error among the 200 datasets, for simulation setups $\rho = 0.5$ and $0.9$. In general, the three values are similar and all of them are close to the value of $p$. So in practice, $\gamma$ can be set as the “Derived” value based on the eigenvalues of $X^TX/n$, or for simplicity, $\gamma = p$ can also be used.
Table 2.5: Theoretic $\gamma$ in the normal prior (2.3) based on Theorem 5, together with mean of the derived and tuned $\gamma$ through methods proposed in Section 2.5, based on 200 datasets with standard errors in parentheses.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Theoretic</th>
<th>Derived ($\rho = 0.5$)</th>
<th>Tuned ($\rho = 0.5$)</th>
<th>Theoretic</th>
<th>Derived ($\rho = 0.9$)</th>
<th>Tuned ($\rho = 0.9$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>47.6</td>
<td>46.6 (0.11)</td>
<td>48.6 (0.24)</td>
<td>40.8</td>
<td>39.9 (0.18)</td>
<td>41.3 (0.25)</td>
</tr>
<tr>
<td>500</td>
<td>490.0</td>
<td>481.8 (0.35)</td>
<td>474.3 (0.83)</td>
<td>482.4</td>
<td>474.1 (0.81)</td>
<td>471.9 (1.23)</td>
</tr>
<tr>
<td>1000</td>
<td>981.7</td>
<td>965.1 (0.51)</td>
<td>947.1 (1.50)</td>
<td>974.0</td>
<td>956.9 (1.18)</td>
<td>944.3 (1.82)</td>
</tr>
</tbody>
</table>

2.7 Real Data Analysis

We now analyze data on mouse gene expression from the experiment conducted by Lan et al. (2006). There were 60 arrays to monitor the expression levels of 22,575 genes consisting of 31 female and 29 male mice. Quantitative real-time PCR were used to measure some physiological phenotypes, including numbers of phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1). The gene expression data and the phenotypic data can be found at GEO (http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330).

First, by ordering the magnitude of marginal correlation between the genes with the three responses from the largest to the smallest, 22,575 genes were screened down to the 999 genes, thus reducing the number of candidate predictors of the three linear regressions. Note that the top 999 genes were not the same for the 3 responses. Then for each of the 3 regressions, the dataset is composed of $n = 60$ observations and $p = 1,000$ predictors (gender along with the 999 genes). After the screening, the Lasso estimator and the penalized credible region method applied on the normal, Laplace and DL priors were used. The hyperparameters in those prior distributions are tuned through the $R^2$ method introduced in Section 2.5, with a target of uniform distribution.
To evaluate the performance of the proposed approach, the first step was to randomly split the sample size 60 into a training set of size 55 and a testing set of size 5. The stopping rule was BIC. To be more specific, the selected model was the one with smallest BIC among all models in which the number of predictors is less than 30. Then the selected model was used to predict the remaining 5 observations, and the prediction error was then obtained. We repeated this for 100 replicates in order to compare the prediction errors. Table 2.6 shows the mean squared prediction error (with its standard error) based on the 100 random splits of the data. The mean selected model size (with its standard error) is also included.

Table 2.6: Mean squared prediction error and model size, with standard errors in parenthesis, based on 100 random splits of the real data.

<table>
<thead>
<tr>
<th></th>
<th>PEPCK</th>
<th></th>
<th>GPAT</th>
<th></th>
<th>SCD1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSPE</td>
<td>Model Size</td>
<td>MSPE</td>
<td>Model Size</td>
<td>MSPE</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.54 (0.026)</td>
<td>25.8 (0.34)</td>
<td>1.43 (0.082)</td>
<td>24.4 (0.56)</td>
<td>0.55 (0.052)</td>
</tr>
<tr>
<td>Normal</td>
<td>0.66 (0.033)</td>
<td>16.8 (0.67)</td>
<td>1.30 (0.099)</td>
<td>16.3 (0.66)</td>
<td>0.71 (0.059)</td>
</tr>
<tr>
<td>Laplace</td>
<td>0.70 (0.037)</td>
<td>17.0 (0.78)</td>
<td>1.19 (0.086)</td>
<td>21.4 (0.56)</td>
<td>0.69 (0.054)</td>
</tr>
<tr>
<td>DL</td>
<td>0.49 (0.032)</td>
<td>18.4 (0.73)</td>
<td>1.37 (0.102)</td>
<td>13.1 (0.68)</td>
<td>0.54 (0.037)</td>
</tr>
</tbody>
</table>

Overall, the results show that the proposed penalized credible region selection method using global-local shrinkage priors such as DL prior performs well. For all 3 responses, the penalized credible region approach with DL prior performs better than the Lasso estimator and has a smaller number of predictors. For PEPCK and SCD1, the DL prior has significant better performance than the normal prior and Laplace prior. For GPAT, there is no significant difference between normal and DL prior. In all, for this dataset, the proposed approach generally improves the performance by replacing the normal prior with the DL prior.
2.8 Discussion

In this chapter, we extend the penalized credible variable selection approach by using global-local shrinkage priors. Simulation studies show that the GL shrinkage priors outperform the original normal prior. Our main result also includes modifying the Dirichlet-Laplace prior to accommodate the linear regression model instead of the simple normal mean problem as in Bhattacharya et al. (2015). In theory, we obtain the selection consistency for the penalized credible region method using the global-local shrinkage priors when $p = o(n)$. Posterior consistency for the normal and DL priors are also shown.

Furthermore, this chapter introduces a new default method to tune the hyperparameters in prior distributions based on the induced prior distribution of $R^2$. The hyperparameter is chosen to minimize a discrepancy between the induced distribution of $R^2$ and a default Beta distribution. For the normal prior, a closed form of the hyperparameters is derived. This method is straightforward and efficient as it only involves the prior distributions. A simulation study illustrates that our proposed tuning method improves upon the usual hyperprior method.

2.9 Proofs

Proof of Theorem 1

Proof. According to Theorem 1 in Armagan et al. (2013b), if under a particular prior, $\beta_n$ satisfies

$$P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) > \exp(-dn)$$

for all $0 < \Delta < \varepsilon d_{min}/(48d_{max})$ and $0 < d < \varepsilon d_{min}/(32\sigma^2) - 3\Delta d_{max}/(2\sigma^2)$ and some $\rho > 0$, then the posterior of $\beta_n$ is consistent. So to get the posterior consistency, the key is to calculate the probability of $\{\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}\}$ under the given prior.
Following the proof of Theorem 2 in Armagan et al. (2013b), we have

\[
P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) = P\left\{ \beta_n : \sum_{j \in A_n} (\beta_{nj} - \beta_{nj}^0)^2 + \sum_{j \notin A_n} \beta_{nj}^2 < \frac{\Delta^2}{n^{\rho}} \right\}
\]

\[
\geq \prod_{j \in A_n} \left\{ P\left( \beta_{nj} : |\beta_{nj} - \beta_{nj}^0| < \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} \right) \right\} \times P\left\{ \beta_{n}^j A : \sum_{j \notin A_n} \beta_{nj}^2 < \frac{(p_n - q_n) \Delta^2}{p_n n^{\rho}} \right\}
\]

\[
\geq \prod_{j \in A_n} \left\{ P\left( \beta_{nj}^0 - \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} < \beta_{nj} < \beta_{nj}^0 + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} \right) \right\} \times \left\{ 1 - \frac{p_n n^\rho E(\sum_{j \notin A_n} \beta_{nj}^2)}{(p_n - q_n) \Delta^2} \right\}
\]

\[
\geq \left\{ 2 \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} f(\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) \right\}^{q_n} \times \left\{ 1 - \frac{p_n n^\rho E(\sum_{j \notin A_n} \beta_{nj}^2)}{(p_n - q_n) \Delta^2} \right\}, \tag{2.10}
\]

where \( f \) is the prior pdf of \( \beta \), symmetric and decreasing when the support is positive. In normal prior (2.3), \( f(\beta_{nj}) = \frac{1}{\sqrt{2\pi\sigma^2/\gamma_n}} \exp\left\{ -\frac{\beta_{nj}^2}{2\sigma^2/\gamma_n} \right\} \), \( E(\beta_{nj}^2) = \sigma^2/\gamma_n \). Following from (2.10), we have

\[
P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) \geq \left\{ 2 \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} \frac{1}{\sqrt{2\pi\sigma^2/\gamma_n}} \exp\left\{ -\frac{(\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}})^2}{2\sigma^2/\gamma_n} \right\} \right\}^{q_n} \times \left\{ 1 - \frac{p_n n^\rho \sigma^2/\gamma_n}{\Delta^2} \right\}
\]

\[
\geq \left\{ 2 \frac{\Delta}{\sqrt{p_n n^{\rho/2}}} \frac{1}{\sqrt{2\pi\sigma^2/\gamma_n}} \exp\left\{ -\frac{(\sup_{j \in A_n} |\beta_{nj}^0|)^2 + \frac{\Delta^2}{\sqrt{p_n n^{\rho/2}}}}{\sigma^2/\gamma_n} \right\} \right\}^{q_n} \times \left\{ 1 - \frac{p_n n^\rho \sigma^2/\gamma_n}{\Delta^2} \right\}.
\]

Taking the negative logarithm of both sides of the above formula, and letting \( \sqrt{\sigma^2/\gamma_n} = \)
\( C/(\sqrt{p}n^{\rho/2} \log n) \), we have

\[
- \log P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) \leq -q_n \log \left\{ \frac{2\Delta}{\sqrt{p}n^{\rho/2}} \sqrt{2\pi \sigma^2} \right\} + q_n \left( \sup_{j \in A_n} |\beta^0_{nj}| \right)^2 + \left( \frac{\Delta^2}{p_n^{\rho}(\log n)^2} \right) - \log \left\{ 1 - \frac{C^2}{\Delta^2(\log n)^2} \right\} + q_n \frac{\Delta^2(\log n)^2}{C^2} + \frac{q_n p_n^{\rho}(\log n)^2 \left( \sup_{j \in A_n} |\beta^0_{nj}| \right)^2}{C^2}.
\]

The last term is the dominating one in the above equation, and \(- \log P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) < \text{dn} \) for all \( d > 0 \) if \( q_n p_n^{\rho}(\log n)^2 = o(n) \), i.e., \( q_n = o\left( \frac{n^{1-\rho}}{p_n(\log n)^2} \right) \). So given the normal prior and assumptions (A1)-(A4), if \( q_n = o\left( \frac{n^{1-\rho}}{p_n(\log n)^2} \right) \) for \( \rho \in (0, 1) \), the prior satisfies \( P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) > \exp(-\text{dn}) \). The posterior consistency is completed by Theorem 1 in Armagan et al. (2013b).

**Proof of Theorem 2**

**Proof.** According to Section 3 in Bhattacharya et al. (2015), the Dirichlet-Laplace prior (2.7) can also be represented as

\[
\beta_j | \xi_j \sim \text{DE}(\xi_j \sigma), \quad \xi_j \sim \text{Ga}(a, 1/2).
\]
And the marginal distribution of $\beta_j$ is

$$f_d(\beta_j) = \int_{\xi_j=0}^{\infty} \left[ \frac{1}{2\xi_j\sigma} \exp\left\{ -\frac{|\beta_j|}{\xi_j\sigma} \right\} \right] \left[ \frac{(1/2)^a}{\Gamma(a)} \xi_j^{a-1} \exp\left\{ -\frac{1}{2}\xi_j \right\} \right] d\xi_j$$

$$= \frac{(1/2)^a}{2\Gamma(a)\sigma} \int_{\xi_j=0}^{\infty} \exp\left\{ -\frac{|\beta_j|}{\xi_j\sigma} \right\} \xi_j^{a-2} \exp\left\{ -\frac{1}{2}\xi_j \right\} d\xi_j.$$

Without loss of generality, we assume $\sigma = 1$. According to the Proposition 3.1 in Bhattacharya et al. (2015), the marginal density function of $\beta_j$ for any $1 \leq j \leq p$ is

$$f_d(\beta_j) = \frac{1}{2^{(1+a)/2}\Gamma(a)} |\beta_j|^{(a-1)/2} K_{1-a}(\sqrt{2|\beta_j|}),$$

where

$$K_{\nu}(x) = \frac{\Gamma(\nu + 1/2)(2x)^\nu}{\sqrt{\pi}} \int_{0}^{\infty} \frac{\cos t}{(t^2 + x^2)^{\nu + 1/2}} dt$$

is the modified Bessel function of the second kind.

Also we have

$$E(\beta_j^2) = \int_{\beta_j=\infty}^{\infty} \int_{\xi_j=0}^{\infty} \left[ \frac{1}{2\xi_j} \beta_j^2 \exp\left\{ -\frac{|\beta_j|}{\xi_j} \right\} \right] \left[ \frac{(1/2)^a}{\Gamma(a)} \xi_j^{a-1} \exp\left\{ -\frac{1}{2}\xi_j \right\} \right] d\xi_j d\beta_j$$

$$= \int_{\xi_j=0}^{\infty} \frac{2\xi_j^2}{\Gamma(a)} \xi_j^{a-1} \exp\left\{ -\frac{1}{2}\xi_j \right\} d\xi_j = 8a(a + 1).$$

Facts: (i) 10.37.1 in DLMF (2015), if $0 \leq \nu < \mu$ and $z$ is a real number, then $|K_\nu(z)| < |K_\mu(z)|$; (ii) 10.39.2 in DLMF (2015), when $\nu = \frac{1}{2}$, $K_{1/2}(z) = \sqrt{\frac{\pi}{2z}} e^{-z}$. For $K_{1-a}(\sqrt{2|\beta_j|})$, as $a \leq \frac{1}{2}$, or $1 - a \geq \frac{1}{2}$; also $\sqrt{2|\beta_j|}$ is a real number for $j = 1, \cdots, p$. So when $\beta_j$ is fixed, $K_{1-a}(\sqrt{2|\beta_j|}) \geq K_{1/2}(\sqrt{2|\beta_j|}) = \sqrt{\frac{\pi}{2\sqrt{2|\beta_j|}}} \exp\left\{ -\sqrt{2|\beta_j|} \right\}$. Combining with the fact that $\Gamma(a) = a^{a-1} - \gamma_0 + O(a) \leq a^{-1} for a close to zero, where $\gamma_0$ is the Euler-Mascheroni
constant, then for \( j = 1, \ldots, p \), we have

\[
f_d(\beta_j) \geq \frac{1}{2^{(1+a)/2} \Gamma(a)} |\beta_j|^{(a-1)/2} \sqrt{\frac{\pi}{2 \times \sqrt{2|\beta_j|}}} \exp\{-\sqrt{2|\beta_j|}\} \geq \frac{\sqrt{\pi} a |\beta_j|^{(2a-3)/4} \exp\{-\sqrt{2|\beta_j|}\}}{2^{(2a+5)/4}}.
\]

Then following from (2.10),

\[
P(\beta_n : ||\beta_n - \beta_0_n|| < \frac{\Delta}{n^{\rho/2}}) \geq \left\{ \begin{array}{c} 2 \Delta \sqrt{\pi a |\sup_{j \in A_n} \beta_0_{nj}| + \frac{\Delta}{\sqrt{n^{\rho/2}}}} \exp\{-2\sup_{j \in A_n} |\beta_0_{nj}| + \frac{\Delta}{\sqrt{n^{\rho/2}}}\} \\
\left\{ 1 - \frac{p_n n^\rho 8a(a + 1)}{\Delta^2} \right\} \end{array} \right\}^q_n.
\]

Taking the negative logarithm of both sides of the above formula, and letting \( a = C/(p_n n^\rho \log n) \),
we have

\[- \log P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) \leq -q_n \log \frac{\sqrt{\pi} \Delta a (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}})^{(2a-3)/4}}{2^{(2a+1)/4} \sqrt{p_n n^{\rho/2}}} + q_n \sqrt{2} (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) \]

\[- \log \left\{ 1 - \frac{p_n n^{\rho} 16a}{\Delta^2} \right\} \]

\[= -q_n \log \frac{\sqrt{\pi} \Delta a}{\sqrt{p_n n^{\rho/2}}} - \frac{q_n a}{2} \log (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) - \log \left\{ 1 - \frac{p_n n^{\rho} 16a}{\Delta^2} \right\} \]

\[+ q_n \sqrt{2} (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) + \frac{3q_n}{4} \log (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) + \frac{q_n (2a + 1) \log 2}{4} \]

\[= -q_n \log (\sqrt{\pi} \Delta C) - \frac{q_n C}{2p_n n^{\rho} \log n} \log (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) - \log \left\{ 1 - \frac{16C}{\Delta^2 \log n} \right\} \]

\[+ q_n \sqrt{2} (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) + \frac{3q_n}{4} \log (\sup_{j \in A_n} |\beta_{nj}^0| + \frac{\Delta}{\sqrt{p_n n^{\rho/2}}}) \]

\[+ \frac{q_n \log 2}{4} \left( \frac{2C}{p_n n^{\rho} \log n} + 1 \right) + q_n \log (p_n^{3/2} n^{3\rho/2} \log n). \]

The last term is the dominating one in the above equation, and \(- \log P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) < dn\) for all \(d > 0\) if \(q_n = o(n/ \log (p_n^{3/2} n^{3\rho/2} \log n))\). Furthermore, \(q_n = o(n/ \log n)\) is a sufficient condition. So given the DL prior and assumptions (A1)-(A4), if \(q_n = o(n/ \log n)\), the prior satisfies \(P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) > \exp(-dn)\). The posterior consistency is completed by Theorem 1 in Armagan et al. (2013b).

\[\Box\]

**Lemma 1.** Given assumptions (A1)-(A4), if \(p_n = o(c_n), c_n \rightarrow \infty\) and \(\gamma_n = o(n)\), then the true parameter \(\beta_n^0\) is contained in the proposed region, i.e., \((\beta_n - \hat{\beta}_n)^T \Sigma_n^{-1} (\beta_n - \hat{\beta}_n) \leq c_n\), with probability increasing to 1.

**Proof.** Since \(\hat{\beta}_n = (X^T X + \gamma_n I)^{-1} (X^T Y)\), and \(\Sigma_n^{-1} = (\hat{\sigma}^2)^{-1} (X^T X + \gamma_n I)\) with \(\hat{\sigma}^2 \rightarrow \sigma^2\),
then \( \hat{\beta}_n - \beta_0^0 = (X^T X + \gamma_n I)^{-1} (X^T (X\beta_0^0 + \varepsilon)) - \beta_0^0 = \left[ (\frac{X^T X}{n} + \frac{\gamma_n I}{n})^{-1} \frac{X^T X}{n} - I \right] \beta_n^0 + (\frac{X^T X}{n} + \frac{\gamma_n I}{n})^{-1} \frac{X^T X}{n} \varepsilon \). Note \( (\frac{X^T X}{n} + \frac{\gamma_n I}{n})^{-1} \frac{X^T X}{n} - I = -\frac{\gamma_n}{n} (\frac{X^T X}{n} + \frac{\gamma_n I}{n})^{-1} \). So for each fixed \( n \), \( \hat{\beta}_n - \beta_0^0 \sim N(m, V) \), where

\[
m = -\frac{\gamma_n}{n} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} \beta_n^0
\]

and

\[
V = \frac{\sigma^2}{n} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} \frac{X^T X}{n} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1}.
\]

Then \( (\hat{\beta}_n - \beta_0^0 - m)^T V^{-1} (\hat{\beta}_n - \beta_0^0 - m) \sim \chi^2_{p_0} \) for each fixed \( n \). Further,

\[
\lim_{p_0, n \to \infty} \frac{1}{p_0} (\hat{\beta}_n - \beta_0^0 - m)^T V^{-1} (\hat{\beta}_n - \beta_0^0 - m) = \lim_{p_0, n \to \infty} \frac{\chi^2_{p_0}}{p_0} = 1. \tag{2.11}
\]

Furthermore,

\[
(\hat{\beta}_n - \beta_0^0 - m)^T V^{-1} (\hat{\beta}_n - \beta_0^0 - m)
\]

\[=
(\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0) - 2m^T V^{-1} (\hat{\beta}_n - \beta_0^0) + m^T V^{-1} m
\]

\[=
(\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0) - 2m^T V^{-1} \left( m + \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} \frac{X^T X}{n} \varepsilon \right) + m^T V^{-1} m
\]

\[=
(\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0) - m^T V^{-1} m - 2m^T V^{-1} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} \frac{X^T X}{n} \varepsilon. \tag{2.12}
\]
First of all, since

\[ 0 \leq \lim_{p_n, n \to \infty} \frac{1}{p_n} m^T V^{-1} m \]

\[ = \lim_{p_n, n \to \infty} \frac{1}{p_n} n \frac{\gamma_n^2}{\sigma^2} \mathbf{0}^T \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right) \]

\[ = \lim_{p_n, n \to \infty} \frac{1}{p_n} \frac{\gamma_n^2}{\sigma^2} \mathbf{0}^T \left( \frac{X^T X}{n} \right)^{-1} \mathbf{0} = \lim_{p_n, n \to \infty} \frac{1}{p_n} \frac{\gamma_n^2}{n \sigma^2} \mathbf{0}^T \Gamma_n D_n^{-1} \Gamma_n \mathbf{0} \]

\[ \leq \lim_{p_n, n \to \infty} \frac{1}{p_n} \frac{\gamma_n^2}{n \sigma^2} \mathbf{0}^T \Gamma_n \text{diag} \{1/d_{\min}, \cdots, 1/d_{\min}\} \Gamma_n \mathbf{0} \]

\[ = \lim_{p_n, n \to \infty} \frac{1}{p_n} \frac{\gamma_n^2}{n \sigma^2} \frac{1}{d_{\min}} ||\mathbf{0}||^2 = 0, \]

so we have

\[ \lim_{p_n, n \to \infty} \frac{1}{p_n} m^T V^{-1} m = 0. \] (2.13)

Next, let’s get the limit of

\[ -\frac{1}{p_n} m^T V^{-1} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} X^T \varepsilon = \frac{1}{p_n} \frac{\gamma_n}{\sigma^2} \mathbf{0}^T \left( \frac{X^T X}{n} \right)^{-1} X^T \varepsilon. \]

As

\[ \frac{\gamma_n}{\sigma^2} \mathbf{0}^T \left( \frac{X^T X}{n} \right)^{-1} X^T \varepsilon \sim N(0, V^*), \]

where

\[ V^* = \left( \frac{\gamma_n}{\sigma^2} \mathbf{0}^T \left( \frac{X^T X}{n} \right)^{-1} X^T \right) \sigma^2 \left( \frac{\gamma_n}{\sigma^2} \frac{X}{n} \mathbf{0}^T \left( \frac{X^T X}{n} \right)^{-1} \mathbf{0} \right) = \frac{\gamma_n^2}{n \sigma^2} \mathbf{0}^T \Gamma_n D_n^{-1} \Gamma_n \mathbf{0} \]

\[ \leq \frac{\gamma_n^2}{n \sigma^2} \mathbf{0}^T \Gamma_n \text{diag} \{1/d_{\min}, \cdots, 1/d_{\min}\} \Gamma_n \mathbf{0} \leq \frac{\gamma_n^2}{n \sigma^2} \frac{1}{d_{\min}} ||\mathbf{0}||^2 \to 0, \]

so

\[ \lim_{p_n, n \to \infty} -\frac{1}{p_n} m^T V^{-1} \left( \frac{X^T X}{n} + \frac{\gamma_n I}{n} \right)^{-1} X^T \varepsilon = 0. \] (2.14)
According to (2.11), (2.12), (2.13) and (2.14), we have

\[
\lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0) = 1. \tag{2.15}
\]

So

\[
1 = \lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0)
\]

\[
= \lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T \frac{n}{\sigma^2} \Gamma_n \text{diag}\{d_1 + \gamma_n/n, \ldots, (d_p + \gamma_n/n)/d_p\}
\]

\[
\Gamma_n^T (\hat{\beta}_n - \beta_0^0)
\]

\[
\geq \lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T \frac{n}{\sigma^2} \Gamma_n \text{diag}\{d_1 + \gamma_n/n, \ldots, (d_p + \gamma_n/n)\} \Gamma_n^T (\hat{\beta}_n - \beta_0^0)
\]

\[
= \lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T \Sigma_n^{-1} (\hat{\beta}_n - \beta_0^0).
\]

Hence then \((\hat{\beta}_n - \beta_0^0)^T \Sigma_n^{-1} (\hat{\beta}_n - \beta_0^0) = p_n \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T \Sigma_n^{-1} (\hat{\beta}_n - \beta_0^0) \leq c_n\), if \(p_n = o(c_n)\), \(c_n \to \infty\) and \(\gamma_n = o(n)\), together with assumptions (A1)-(A4), the true parameter is contained in the region with probability tending to 1.

\[\square\]

**Lemma 2.** Under assumptions (A1)-(A4), and \(\gamma_n = o(n)\), the posterior mean \(\hat{\beta}_n = (X^T X + \gamma_n I)^{-1} (X^T Y)\) has the property: \(\frac{n}{p_n} ||\hat{\beta}_n - \beta_0^0||^2 = O(1)\), or \(\sqrt{\frac{n}{p_n}} (\hat{\beta}_{nj} - \beta_0^{0j}) = O(1)\) for \(j = 1, \ldots, p\).

**Proof.** Following (2.15) in the proof of Lemma 1, we have

\[
\lim_{p_n,n \to \infty} \frac{1}{p_n} (\hat{\beta}_n - \beta_0^0)^T V^{-1} (\hat{\beta}_n - \beta_0^0) = \lim_{p_n,n \to \infty} \frac{1}{p_n} \frac{n}{\sigma^2} (\hat{\beta}_n - \beta_0^0)^T \Gamma_n \text{diag}\{d_1 + \gamma_n/n, \ldots, (d_p + \gamma_n/n)/d_p\} \Gamma_n^T (\hat{\beta}_n - \beta_0^0),
\]
\( \beta_n^0 = 1 \). So as \( p_n, n \to \infty \), we have

\[
\frac{nd_{\min}}{p_n\sigma^2} (\hat{\beta}_n - \beta_n^0)^T (\hat{\beta}_n - \beta_n^0) = \frac{n}{p_n\sigma^2} (\hat{\beta}_n - \beta_n^0)^T \Gamma_n \text{diag}(d_{\min}, \ldots, d_{\min}) \Gamma_n (\hat{\beta}_n - \beta_n^0) \\
\leq \frac{n}{p_n\sigma^2} (\hat{\beta}_n - \beta_n^0)^T \Gamma_n \text{diag}(d_1, \ldots, d_{p_n}) \Gamma_n (\hat{\beta}_n - \beta_n^0) \\
\leq \frac{n}{p_n\sigma^2} (\hat{\beta}_n - \beta_n^0)^T \Gamma_n \text{diag}\{(d_1 + \gamma_n/n)^2/d_1, \ldots, (d_{p_n} + \gamma_n/n)^2/d_{p_n} \} \Gamma_n (\hat{\beta}_n - \beta_n^0) \to 1.
\]

Then \( \frac{n}{p_n}||\hat{\beta}_n - \beta_n^0||^2 \leq \frac{\sigma^2}{d_{\min}} \), i.e., \( \frac{n}{p_n}||\hat{\beta}_n - \beta_n^0||^2 = O(1) \) or \( \sqrt{\frac{n}{p_n}}(\hat{\beta}_{n_j} - \beta_{n_j}^0) = O(1) \) for \( j = 1, \ldots, p \).

Note: This cannot ensure for every \( j = 1, \ldots, p \), \( \sqrt{n}(\hat{\beta}_{n_j} - \beta_{n_j}^0) = O(1) \). For example, if \( \sqrt{n}(\hat{\beta}_{n1} - \beta_{n1}^0) = \sqrt{p_n} \), and all the other terms are zero, we’d still have \( \frac{n}{p_n}||\hat{\beta}_n - \beta_n^0||^2 = \frac{1}{p_n} \sum_{j=1}^{p_n} (\sqrt{n}(\hat{\beta}_{n_j} - \beta_{n_j}^0))^2 = \frac{1}{p_n}(p_n + 0) = 1 = O(1) \). \( \square \)

**Lemma 3.** If \( c_n \to \infty \) and \( \frac{c_n}{np_n} \to c \) for \( 0 < c \leq \infty \), then \( (\hat{\beta}_{n_j} - \beta_{n_j}^0)^{-1} = O(1) \), for some \( j \), where \( \hat{\beta}_n \) is the solution to the optimization algorithm (2.5) for the particular choice of \( c_n \).

This implies that \( \hat{\beta}_n \) is not a consistent estimator.

**Proof.** Let’s assume \( (\hat{\beta}_{n_j} - \beta_{n_j}^0) \to 0 \) for all \( j \) and obtain a contradiction.

Since the solution occurs on the boundary of the credible set, \( (\hat{\beta}_n - \hat{\beta}_n)^T \Sigma^{-1}(\hat{\beta}_n - \hat{\beta}_n) = (\hat{\beta}_n - \hat{\beta}_n)^T \frac{(X^TX + \gamma_n I)}{\sigma^2}(\hat{\beta}_n - \hat{\beta}_n) = c_n \), or equally, \( (\hat{\beta}_n - \hat{\beta}_n)^T (X^TX + \gamma_n I)(\hat{\beta}_n - \hat{\beta}_n) = c_n\sigma^2 \).

Multiplying both sides by \( \frac{1}{np_n} \), on the right hand side, we have \( \frac{c_n\sigma^2}{np_n} \to c\sigma^2 \). For the left side, we have \( \frac{1}{np_n} (\hat{\beta}_n - \hat{\beta}_n)^T \frac{(X^TX + \gamma_n I)}{n}(\hat{\beta}_n - \hat{\beta}_n) \).

According to Lemma 2, \( \frac{n}{p_n}||\hat{\beta}_n - \beta_n^0||^2 = O(1) \), so \( \frac{1}{np_n}(\hat{\beta}_n - \beta_n^0)^T(\hat{\beta}_n - \beta_n^0) \to 0 \). Furthermore, by the assumption, \( (\hat{\beta}_{n_j} - \beta_{n_j}^0) \to 0 \) for all \( j \), then \( \frac{1}{p_n}(\hat{\beta}_n - \beta_n^0)^T(\hat{\beta}_n - \beta_n^0) \to 0 \). So
\[\frac{1}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T(\tilde{\beta}_n - \hat{\beta}_n) \to 0.\] Then as \(p_n, n \to \infty\), we have

\[
\frac{1}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \left( \frac{X^T X + \gamma_n I}{n} \right)(\tilde{\beta}_n - \hat{\beta}_n)
\]

\[
= \frac{1}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \Gamma_n \text{diag}\{(d_1 + \gamma_n/n), \ldots, (d_{p_n} + \gamma_n/n)\} \Gamma_n^T(\tilde{\beta}_n - \hat{\beta}_n)
\]

\[
\leq \frac{1}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \Gamma_n \text{diag}\{(d_{\max} + \gamma_n/n), \ldots, (d_{\max} + \gamma_n/n)\} \Gamma_n^T(\tilde{\beta}_n - \hat{\beta}_n)
\]

\[
\leq \frac{(d_{\max} + \gamma_n/n)}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T(\tilde{\beta}_n - \hat{\beta}_n) \to 0.
\]

Hence the left side goes to zero, while the right side converges to a non-zero constant. Thus we obtain a contradiction. \(\square\)

**Lemma 4.** Let \(0 < \varepsilon < 1\), and \(\tilde{\beta}_n\) be the solution to the optimization problem for the choice of \(c_n\). If \(\frac{c_n}{n^{\varepsilon}p_n} \to c\), where \(0 < c < \infty\), then \(\frac{n^{1-\varepsilon}}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n) = O(1)\), and \(\left[\frac{n^{1-\varepsilon}}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n)\right]^{-1} = O(1)\).

**Proof.** Suppose \(\frac{c_n}{n^{\varepsilon}p_n} \to c\), since the solution occurs on the boundary of the credible set, we have \(n(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n) = c_n\). Multiplying both sides by \(\frac{1}{n^{\varepsilon}p_n}\), on the right hand side we have \(\frac{c_n}{n^{\varepsilon}p_n} \to c\). So we have \(\frac{n^{1-\varepsilon}}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n) \to c\). As \(0 < c < \infty\), then \(\frac{n^{1-\varepsilon}}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n) = O(1)\), and \(\left[\frac{n^{1-\varepsilon}}{p_n}(\tilde{\beta}_n - \hat{\beta}_n)^T \frac{\Sigma^{-1}}{n}(\tilde{\beta}_n - \hat{\beta}_n)\right]^{-1} = O(1)\). \(\square\)

**Lemma 5.** Let \(0 < \varepsilon < 1\). Then if \(\frac{c_n}{n^{\varepsilon}p_n} \to c\), \(\sqrt{\frac{p_n}{n}}(\tilde{\beta}_{nj} - \hat{\beta}_{nj}) \to \infty\) can be true only for \(j \in A_n^c\).

**Proof.** From Lemma 4, we have \(\sqrt{\frac{p_n}{n}}(\tilde{\beta}_{nj} - \hat{\beta}_{nj}) \to \infty\) for some \(j\). We now prove that it cannot be true for \(j \in A_n^c\).

Without loss of generality, we assume the true parameters are \(\beta_{n1} = \cdots = \beta_{nk} = 0\) and \(\beta_{nk+1} = \cdots = \beta_{np_n} \neq 0\). Assume \(S(\beta_n) = \sum_{j=1}^{p_n} |\tilde{\beta}_{nj}|^2|\beta_{nj}|\), then the solution is the minimizer of \(S(\beta_n)\) among those points within the given credible set.
Suppose $\tilde{\beta}_n$ is the minimizer of $S(\beta_n)$, and suppose that $\sqrt{\frac{n}{p_n}}(\hat{\beta}_{n1} - \tilde{\beta}_{n1}) \to \infty$. Since $\beta_{n1} \in A_{n}^{c}$, it follows that $\sqrt{\frac{n}{p_n}}(\hat{\beta}_{n1} - 0) = O(1)$. Hence, it must be that $\sqrt{\frac{n}{p_n}}\hat{\beta}_{n1} \to \infty$. Also $\sqrt{\frac{n}{p_n}}|\hat{\beta}_{n1}|^2 \to 0$. So, the first term of $S(\tilde{\beta}_n)$, $\sqrt{\frac{n}{p_n}}|\tilde{\beta}_{n1}|^2 \to \infty$. Therefore, $S(\tilde{\beta}_n) \to \infty$.

According to the proof of Lemma 4 in Bondell and Reich (2012), there exists a $\tilde{\beta}_n^0 = \{0, \cdots, 0, \tilde{\beta}_{nk+1}^*, \cdots, \tilde{\beta}_{npn}^*\}$ be the minimizer of $S(\beta_n)$, within the credible set. Also $S(\beta_n^0) < S(\tilde{\beta}_n)$, and hence achieve a contradiction. The proof follows from Lemma 4 in Bondell and Reich (2012), except that $\sqrt{n}$ is replaced with $\sqrt{\frac{n}{p_n}}$. Therefore, $\sqrt{\frac{n}{p_n}}(\hat{\beta}_{nj} - \tilde{\beta}_{nj}) \to \infty$ can be true only for $j \in A_n$.

**Proof of Theorem 3**

*Proof*. Following Bondell and Reich (2012), we show that $P(A_n \cap A_{n}^{c}) \to 0$ and $P(A_{n}^{c} \cap A_n) \to 0$. By Lemma 4, when $c_{n} = o(np_n)$, the credible set is shrinking around $\tilde{\beta}_n$. Hence $\tilde{\beta}_{nj} \to \beta_{nj}$ for all $j$, then $P(A_{n}^{c} \cap A_n) \to 0$.

We now show that $P(A_n \cap A_{n}^{c}) \to 0$. Without loss of generality, assume $\beta_{n1} = 0, \beta_{npn} \neq 0$ and by Lemma 5, let $\sqrt{\frac{n}{p_n}}(\tilde{\beta}_{npn} - \beta_{npn}) \to \infty$. Denote $\sigma_{ij}$ as the $i-j$th element of $\Sigma^{-1}$. Assume $\tilde{\beta}_{n1} \neq 0$, and we’ll get a contradiction. The proof is following the proof of Theorem 1 in Bondell and Reich (2012). As $(\tilde{\beta}_n - \tilde{\beta}_n)^T\Sigma^{-1}(\tilde{\beta}_n - \tilde{\beta}_n) = c_n$, then $\beta_{npn}$ can be represented as a function of the remaining $p_n - 1$ coefficients. Since $\tilde{\beta}_{n1} \neq 0$, the minimizer of $\sum_{j=1}^{p_n} \frac{1}{\hat{\beta}_{nj}}|\beta_{nj}|$ with respect to $\beta_{n1}$ satisfies

$$
\frac{1}{|\tilde{\beta}_{n1}^2|}\text{sign}(\tilde{\beta}_{n1}) + \frac{1}{|\tilde{\beta}_{npn}^2|}\text{sign}(\tilde{\beta}_{npn}) \frac{\partial \beta_{npn}}{\partial \beta_{n1}} \bigg|_{\beta_n} = 0.
\tag{2.16}
$$

Consider the first term on the left hand side. Since $\beta_{n1} = 0$, we have $\hat{\beta}_{n1} \to 0$, then $\frac{1}{|\beta_{n1}^2|} \to \infty$. Hence $\frac{1}{|\beta_{n1}^2|}\text{sign}(\tilde{\beta}_{n1}) \to \infty$ if $\tilde{\beta}_{n1} \neq 0$. For the second term, since $\beta_{npn} \neq 0$, $\frac{1}{|\beta_{npn}|} = O(1)$.
The goal is to show that \( \delta \) any \( \epsilon > 0 \). Proof.

For any global-local shrinkage prior represented as (2.2), the prior precision \( \Theta \) in (2.17) are \( O(1) \). Also, due to the consistency of ordinary least square estimator \( \hat{\beta} \), \( n \) where \( \hat{\beta} \mid \beta \) \( \mid \mid Y \) of \( \hat{\beta} \mid \beta \) \( \mid \mid Y \) yields a contradiction. Hence, the left side of (2.16) diverges, which yields a contradiction.

Proof of Theorem 4

Proof. For any global-local shrinkage prior represented as (2.2), the prior precision \( \delta_j \) is \( 1/(w \xi_j) \). The goal is to show that \( \delta_j = o(n) \) for each \( j = 1, \cdots, p_n \) with posterior probability 1, i.e., for any \( \epsilon > 0, P(1/n \delta_j \geq \epsilon \mid Y) \to 0 \) as \( n \to \infty \).

If the GL prior produces consistent posteriors, i.e., for any \( \epsilon > 0, P(\beta_n : \mid \mid \beta_n - \beta_0^0 \mid \mid \epsilon \mid Y) \to 0 \) as \( p_n, n \to \infty \). Then the posterior mean, \( \hat{\beta}_{GL} \), satisfies \( P(\mid \mid \hat{\beta}_{GL} - \beta_0^0 \mid \mid \epsilon \mid Y) \to 0 \). Also, due to the consistency of ordinary least square estimator \( \hat{\beta}_{n}^{OLS} \), it follows that \( P(\mid \mid \hat{\beta}_n^{GL} - \hat{\beta}_n^{OLS} \mid \mid \epsilon \mid Y) \to 0 \), or \( \hat{\beta}_n^{GL} - \hat{\beta}_n^{OLS} \to 0 \). Let \( \delta = (\delta_1, \cdots, \delta_{p_n}) \). Furthermore, we have

\[
\hat{\beta}_{GL}^n = E_{\delta \mid Y}[E(\beta \mid \delta, Y)] = \int_{\delta} (X^T X + S^{-1})^{-1} X^T Y \pi(\delta \mid Y) d\delta,
\]

where \( S = \text{diag} \{\delta_1, \cdots, \delta_p\} \) and \( \pi(\delta \mid Y) \) is the posterior density function of \( \delta \). Also as \( X_n^T X_n = \Gamma_n D_n \Gamma_n^T \) where \( D_n = \text{diag} \{d_1, \cdots, d_{p_n}\} \), we have

\[
\hat{\beta}_{GL}^n - \hat{\beta}_{OLS}^n = \int_{\delta} \Gamma_n \text{diag} \{\left(1 + \frac{1}{n \delta_1}\right)^{-1} - d_1^{-1}, \cdots, \left(1 + \frac{1}{n \delta_{p_n}}\right)^{-1} - d_{p_n}^{-1}\} \Gamma_n^T X_n^T Y \pi(\delta \mid Y) d\delta.
\]

Differentiating \((\tilde{\beta}_n - \hat{\beta}_n)^T \Sigma_n^{-1}(\tilde{\beta}_n - \hat{\beta}_n) = c_n\) with respect to \( \beta_{n1} \) yields

\[
\frac{\partial \beta_{npn}}{\partial \beta_{n1}} \bigg|_{\beta_n} = \frac{\sum_{j=1}^{p_n} \sigma_j (\hat{\beta}_nj - \hat{\beta}_nj)}{\sum_{j=1}^{p_n} \sigma_j (\hat{\beta}_nj - \hat{\beta}_nj)} = - \frac{\sum_{j=1}^{p_n} \sigma_j \sqrt{n^{1-\epsilon} (\hat{\beta}_nj - \hat{\beta}_nj)}}{\sum_{j=1}^{p_n} \sigma_j \sqrt{n^{1-\epsilon} (\hat{\beta}_nj - \hat{\beta}_nj)}}, \quad (2.17)
\]

where \( 0 < \epsilon < 1 \) is such that \( \frac{c_n}{n^{1-\epsilon}} \to c \). Note that by Lemma 4, both numerator and denominator in (2.17) are \( O(1) \). Also by Lemma 5, the denominator cannot be 0, due to the presence of \((\hat{\beta}_{npn} - \hat{\beta}_{npn})\). Hence \( \frac{\partial \beta_{npn}}{\partial \beta_{n1}} \bigg|_{\beta_n} = O(1) \). Then, \( \frac{1}{|\beta_{npn}|} \text{sign}(\hat{\beta}_{npn}) \frac{\partial \beta_{npn}}{\partial \beta_{n1}} \bigg|_{\beta_n} = O(1) \). Hence, the left side of (2.16) diverges, which yields a contradiction. \( \square \)
Since as $n \to \infty$, $\frac{1}{n}X^TY \sim N(\frac{X^TX}{n}\beta, \sigma^2\frac{X^TX}{n})$ is a random variable with $P(\frac{1}{n}X^TY = 0) = 0$, then it follows that $f_\delta \left((d_j + \frac{1}{n\delta_j})^{-1} - d_j^{-1}\right) \pi(\delta|Y) d\delta \to 0$ for each $j = 1, \cdots, p_n$. So $P(\frac{1}{n\delta_j} \geq \epsilon |Y) = \int \delta I\{\frac{1}{n\delta_j} \geq \epsilon\} \pi(\delta|Y) d\delta \to 0$ as $n \to 0$, for each $j = 1, \cdots, p_n$. The selection consistency of the credible region variable selection follows.

\[\text{Proof of Theorem 5}\]

\textbf{Proof.} According to the assumption that $\frac{X^TX}{n} = \Gamma D \Gamma^T$ where $D = \text{diag}\{d_1, \cdots, d_p\}$ with $d_1, \cdots, d_p$ denoting the eigenvalues. Since $\beta \sim N(0, \sigma^2/\gamma I_p)$, then $\sqrt{\gamma/\sigma^2} \Gamma^T \beta \sim N(0, I_p)$.

Also,

$$R^2 = 1 - \frac{\sigma^2}{\beta^T \frac{X^TX}{n} \beta + \sigma^2} = \frac{\beta^T \frac{X^TX}{n} \beta}{\beta^T \frac{X^TX}{n} \beta + \sigma^2} = \frac{\gamma}{\sigma^2} (\beta^T \Gamma D \Gamma^T \beta) = W + \gamma,$$

where $W = \frac{\gamma}{\sigma^2} (\beta^T \Gamma D \Gamma^T \beta) = d_1 Z_1^2 + \cdots + d_p Z_p^2$, where $Z_1, \cdots, Z_p$ are i.i.d. from $N(0, 1)$, or $Z_1^2, \cdots, Z_p^2$ i.i.d. from $\chi^2_1$. Then $W$ follows a distribution with density denoted as $f_W(\cdot)$, with mean $\sum_{j=1}^p d_j$, and variance $2\sum_{j=1}^p d_j^2$.

On the other hand, if $R^2$ follows a $\text{Beta}(a, b)$ distribution, then the density function of $W = \frac{\gamma R^2}{1-R^2}$ is as follows:

$$f_B(w) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{w^{a-1} \gamma^b}{(\gamma + w)^{a+b}}, \ (w \geq 0).$$

Thus, to get the solution of $\gamma$ to make the distribution of $R^2$ closest to the Beta distribution, or $f_W(\cdot)$ closest to $f_B(\cdot)$, one needs to minimize the Kullback-Liebler directed divergence.
between them, which is given as below:

\[
D(f_W|f_B) = \int_0^\infty f_W(x) \log \frac{f_W(x)}{f_B(x)} \, dx
\]

\[
= \int_0^\infty f_W(x) \log f_W(x) \, dx - \int_0^\infty f_W(x) \log f_B(x) \, dx
\]

\[
= -b \log \gamma - (a - 1) E[\log W] + (a + b) E[\log(W + \gamma)] + C
\]

\[
\approx -b \log \gamma + (a + b) \left( \log(E[W] + \gamma) - \frac{\text{Var}[W]}{2(E[W] + \gamma)^2} \right) + C^*
\]

where \( C \) and \( C^* \) are some constant value with no relation to \( \gamma \), and \( E(\cdot) \) and \( \text{Var}(\cdot) \) denote the expectation and variance of the random variable \( W \) with density \( f_W(\cdot) \). The derivation of the above formula relies on the following facts: (i) \( \log x = \log x_0 + \frac{1}{x_0}(x-x_0) - \frac{(x-x_0)^2}{2x_0^2} + O((x-x_0)^3) \); (ii) When \( p \to \infty \), a third derivative is small, so a second order Taylor expansion around \( x_0 = E[X] \) can be used to approximate \( E[\log X] \):

\[
E[\log X] \approx E[\log x_0] + E\left[ \frac{1}{x_0}(X - x_0) \right] - E\left[ \frac{(X - x_0)^2}{2x_0^2} \right] = \log(E[X]) - \frac{\text{Var}[X]}{2(E[X])^2}.
\]

(iii) Similarly, a second order Taylor expansion around \( x'_0 = x_0 + c = E[X] + c \) is used to approximate \( E[\log(X + c)] \):

\[
E[\log(X + c)] \approx E[\log(x_0 + c)] + E\left[ \frac{X - x_0}{x_0 + c} \right] - E\left[ \frac{(X - x_0)^2}{2(x_0 + c)^2} \right] = \log(E[X] + c) - \frac{\text{Var}[X]}{2(E[X] + c)^2}.
\]

Then taking the derivative of \( D(f_W|f_B) \), i.e., \( D'(f_W|f_B) = -b \gamma + \frac{a+b}{\sum_j d_j + \gamma} + \frac{2(a+b)\sum_j d_j^2}{(\sum_j d_j + \gamma)^2} \),
and letting it equal to 0, we have

\[
\gamma^3 + \frac{2a-b}{a} \left( \sum_{j=1}^{p} d_j \right) \gamma^2 + \left( \frac{2(a+b)}{a} \sum_{j=1}^{p} d_j^2 + \frac{a-2b}{a} \left( \sum_{j=1}^{p} d_j \right)^2 \right) \gamma - \frac{b}{a} \left( \sum_{j=1}^{p} d_j \right)^3 = 0. \tag{2.18}
\]

According to the conclusion in Osler (2002), if \( P = \frac{2a-b}{a} \sum_{j=1}^{p} d_j, \) \( Q = \frac{2(a+b)}{a} \sum_{j=1}^{p} d_j^2 + \frac{a-2b}{a} \left( \sum_{j=1}^{p} d_j \right)^2, \) \( R = -\frac{b}{a} \left( \sum_{j=1}^{p} d_j \right)^3, \) \( C = P^2/9 - Q/3, \) \( A = PQ/6 - P^3/27 - R/2, \) \( B = A^2 - C^3, \) and \( B \geq 0, \) then \( \gamma = (A + \sqrt{B})^{1/3} + (A - \sqrt{B})^{1/3} - P/3 \) is the unique real solution to (2.18). \( \Box \)
Chapter 3

High Dimensional Linear Regression via
the R2-D2 Shrinkage Prior

3.1 Introduction

Consider the linear regression model,

\[ Y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n, \quad (3.1) \]

where \( Y_i \) is the \( i \)th response, \( x_i \) is the \( p \)-dimensional vector of covariates for \( i \)th observation, \( \beta = (\beta_1, \cdots, \beta_p)^T \) is the coefficient vector, and the \( \epsilon_i \)'s are the error terms assumed be normal and independent with \( \text{E}(\epsilon_i) = 0 \) and \( \text{Var}(\epsilon_i) = \sigma^2 \). High-dimensional data with \( p > n \) in this context is common in diverse application areas. It is well known that maximum likelihood estimation performs poorly in this setting, and this motivates a number of approaches in shrinkage estimation and variable selection. In the Bayesian framework, there are two main alternatives to address such problems: two component discrete mixture priors (also referred as spike and slab
priors) and continuous shrinkage priors. The discrete mixture priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ishwaran and Rao, 2005; Narisetty et al., 2014) put a point mass (spike) at $\beta_j = 0$ and a continuous prior (slab) for the terms with $\beta_j \neq 0$. Although these priors have an intuitive and appealing representation, they lead to computational issues due to the spread of posterior probability over the $2^p$ models formed by including subsets of the coefficients to zero.

These issues with discrete mixture priors motivate the second approach, the continuous shrinkage prior. The shrinkage priors are essentially written as global-local (GL) scale mixture Gaussian family as summarized in Polson and Scott (2010), i.e.,

$$
\beta_j | \phi_j, \omega \sim N(0, \omega \phi_j), \quad \phi_j \sim \pi(\phi_j), \quad (\omega, \sigma^2) \sim \pi(\omega, \sigma^2),
$$

where $\omega$ represents the global shrinkage, while $\phi_j$’s are the local variance components. Current existing GL priors exhibit desirable theoretic and empirical properties. The priors are continuous but have high concentration at zero and heavy tails, which reflects the prior that many covariates are irrelevant while a few have large effect, without explicitly having prior probability at $\beta_j = 0$. Some examples include normal-gamma mixtures (Griffin et al., 2010), Horseshoe (Carvalho et al., 2009, 2010), generalized Beta (Armagan et al., 2011), generalized double Pareto (Armagan et al., 2013a), Dirichlet-Laplace (Bhattacharya et al., 2015), and Horseshoe+ (Bhadra et al., 2015). Moreover, several frequentist regularization methods lend themselves to the GL shrinkage priors, such as Bayesian Lasso (Park and Casella, 2008; Hans, 2009), Bayesian elastic net (Li et al., 2010), Bayesian Bridge (Polson et al., 2013), and Bayesian adaptive Lasso (Leng et al., 2014). GL priors have substantial computational advantages over the discrete mixture priors.

We propose a new global-local prior, which we term an $R^2$-induced Dirichlet Decomposi-
The R2-D2 prior. $R^2$, the coefficient of determination, is defined as the square of the correlation coefficient between the original dependent variable and the modeled value. The motivation comes from the fact that it is hard to specify a $p$-dimensional prior on $\beta$ with high dimensional data, however, it is more direct to construct a prior on the 1-dimensional $R^2$. The R2-D2 prior is induced by a prior on $R^2$, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. We show that the R2-D2 prior includes a number of existing priors as special cases, such as the normal-gamma (Griffin et al., 2010) and Horseshoe (Carvalho et al., 2009, 2010) priors. The R2-D2 prior has many appealing properties, such as strongly shrinking small coefficients due to a tight peak at zero, allowing for large coefficients due to the heavy tails, and enjoying computational tractability with a straightforward Gibbs sampler. We also offer a theoretical framework to compare different global-local priors. The R2-D2 compares favorably to the other GL shrinkage priors in terms of both concentration around the origin and tail behavior. We also demonstrate that in the orthogonal design setup, the proposed R2-D2 prior guarantees that the Bayes estimator converges to the truth at a Kullback-Leibler super-efficient rate. In fact, the R2-D2 prior attains a sharper information theoretic bound than the existing GL priors, such as the Horseshoe (Carvalho et al., 2009, 2010) and Horseshoe+ (Bhadra et al., 2015) prior.

In terms of posterior properties, Armagan et al. (2013b) investigates the asymptotic behavior of posterior distributions of regression coefficients as $p$ grows with $n$. They prove the posterior consistency for some shrinkage priors, including the double-exponential prior, Student’s $t$ prior, generalized double Pareto prior, and the Horseshoe-like priors. Under similar conditions, Chapter 2 of this thesis demonstrates posterior consistency for the Dirichlet-Laplace prior. In this chapter, we prove that our proposed R2-D2 prior also leads to consistent posterior distributions.

The rest of this chapter is outlined as follows. Section 3.2 describes the motivation and
details of the R2-D2 prior, along with posterior computation. Section 3.3 provides theoretical properties of the R2-D2 prior including posterior consistency, as well as a comparison with other common shrinkage priors, in terms of tail behavior, concentration at zero and predictive efficiency. Section 3.4 compares different global-local shrinkage priors including our proposed R2-D2 priors through numerical examples. Section 3.5 applies the proposed priors on a PCR mouse gene expression data set. Section 3.6 summarizes the main results and discusses some future directions. All proofs are given in Section 3.7.

3.2 A New Class of Global-Local Shrinkage Priors

3.2.1 Motivation

The primary goal is to estimate the vector $\beta$ and select important covariates. Common Bayesian methods assume a prior on $\beta$ directly. In this chapter, we start by placing a prior on a univariate function of $\beta$ with practical meaning, and then induce a prior on the $p$-dimensional $\beta$.

Suppose that the predictor vectors $x_1, x_2, \cdots, x_n \overset{iid}{\sim} H(\cdot)$, with $E(x_i) = \mu$ and $\text{Cov}(x_i) = \Sigma$. Assume that for each $i = 1, \cdots, n$, $x_i$ is independent of the $n$-vector of errors, $\varepsilon$, and the marginal variance of $Y_i$ is then $\text{Var}(x^T \beta) + \sigma^2$. For simplicity, we assume that the response is centered and covariates are standardized so that there is no intercept term in (3.1), and all diagonal elements of $\Sigma$ are 1. The coefficient of determination, $R^2$, can be calculated as the square of the correlation coefficient between the dependent variable, $Y$, and the modeled value, $x^T \beta$, i.e.,

$$R^2 = \frac{\text{Cov}^2(Y, x^T \beta)}{\text{Var}(Y)\text{Var}(x^T \beta)} = \frac{\text{Cov}^2(x^T \beta + \varepsilon, x^T \beta)}{\text{Var}(x^T \beta + \varepsilon)\text{Var}(x^T \beta)} = \frac{\text{Var}(x^T \beta)}{\text{Var}(x^T \beta) + \sigma^2}. \quad (3.2)$$

Consider a prior for $\beta$ satisfying $E(\beta) = 0$ and $\text{Cov}(\beta) = \sigma^2 \Lambda$, where $\Lambda$ is a diagonal
matrix with diagonal elements $\lambda_1, \cdots, \lambda_p$. Then

$$\text{Var}(x^T \beta) = E_x\{\text{Var}_{x}(x^T \beta | x)\} + \text{Var}_{x}\{E_{x}(x^T \beta | x)\} = E_x\{\sigma^2 x^T \Lambda x\} + \text{Var}_x\{0\}$$

$$= \sigma^2 E_x\{\text{tr}(x^T \Lambda x)\} = \sigma^2 \text{tr}(\Lambda E_x\{xx^T\}) = \sigma^2 \sum_{j=1}^{p} \lambda_j.$$ 

Then plugging into (3.2), $R^2$ is represented as

$$R^2 = \frac{\text{Var}(x^T \beta)}{\text{Var}(x^T \beta) + \sigma^2} = \frac{\sigma^2 \sum_{j=1}^{p} \lambda_j}{\sigma^2 \sum_{j=1}^{p} \lambda_j + \sigma^2} = \frac{\sum_{j=1}^{p} \lambda_j}{\sum_{j=1}^{p} \lambda_j + 1} \equiv \frac{W}{W + 1},$$ (3.3)

where $W \equiv \sum_{j=1}^{p} \lambda_j$ is the sum of the prior variances (scaled by $\sigma^2$).

Suppose $R^2 \sim \text{Beta}(a, b)$, a Beta distribution with shape parameters $a$ and $b$, then the induced prior density for $W = R^2/(1 - R^2)$ is a Beta Prime distribution denoted as $\text{BP}(a, b)$, with probability density function

$$\pi_W(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{x^{a-1}}{(1 + x)^{a+b}}, \ (x > 0).$$

Therefore $W \sim \text{BP}(a, b)$ is equivalent to the prior $R^2 \sim \text{Beta}(a, b)$. The following section will induce a prior on $\beta$ based on the distribution of the sum of prior variances $W$.

### 3.2.2 The R2-D2 Prior

Any prior of the form $E(\beta) = 0$, $\text{Cov}(\beta) = \sigma^2 \Lambda$ and $W = \sum_{j=1}^{p} \lambda_j \sim \text{BP}(a, b)$ induces a Beta$(a, b)$ prior on the $R^2$. To construct a prior with such properties, we follow the global-local prior framework, and express $\lambda_j = \phi_j \omega$ with $\sum_{j=1}^{p} \phi_j = 1$. Then $W = \sum_{j=1}^{p} \phi_j \omega = \omega$ is the total prior variability, and $\phi_j$ is the proportion of total variance allocated to the $j$-th covariate. It
is natural to assume that $\omega \sim \text{BP}(a, b)$ and the variances across covariates have a Dirichlet prior with concentration parameter $(a_\pi, \cdots, a_\pi)$, i.e., $\phi = (\phi_1, \cdots, \phi_p) \sim \text{Dir}(a_\pi, \cdots, a_\pi)$, where small $a_\pi$ would lead to most $\phi$ close to zero, while large $a_\pi$ lead to uniform $\phi$, i.e., $\phi \approx (1/p, \cdots, 1/p)$, thus controlling the sparsity. Assume a prior $K(\cdot)$ on each dimension of $\beta$, with $K(\delta)$ denotes a kernel (density) with mean zero and variance $\delta$. The prior is summarized as

$$
\beta_j | \sigma^2, \phi_j, \omega \sim K(\sigma^2 \phi_j \omega), \phi \sim \text{Dir}(a_\pi, \cdots, a_\pi), \omega \sim \text{BP}(a, b). \quad (3.4)
$$

Such prior is induced by a prior on $R^2$ and the total prior variance of $\beta$ is decomposed through a Dirichlet prior, therefore we refer to the prior as the $R^2$-induced Dirichlet Decomposition (R2-D2) prior.

**Proposition 1.** If $\omega | \xi \sim \text{Ga}(a, \xi)$ and $\xi \sim \text{Ga}(b, 1)$, then $\omega \sim \text{BP}(a, b)$, where $\text{Ga}(\mu, \nu)$ is the Gamma random variable with shape $\mu$ and rate $\nu$.

Hence (3.4) can also be written as

$$
\beta_j | \sigma^2, \phi_j, \omega \sim K(\sigma^2 \phi_j \omega), \phi \sim \text{Dir}(a_\pi, \cdots, a_\pi), \omega | \xi \sim \text{Ga}(a, \xi), \xi \sim \text{Ga}(b, 1).
$$

As shown in the next section, Proposition 1’s representation of a Beta prime variable in terms of two Gamma variables reveals connections among other common shrinkage priors.

**Proposition 2.** If $\omega \sim \text{Ga}(a, \xi)$, $(\phi_1, \cdots, \phi_p) \sim \text{Dir}(a_\pi, \cdots, a_\pi)$, and $a = pa_\pi$, then $\phi_j \omega \sim \text{Ga}(a_\pi, \xi)$ independently for $j = 1, \cdots, p$.

Now, using Proposition 2, reducing to the special case of $a = pa_\pi$ in (3.4) is equivalent to

$$
\beta_j | \sigma^2, \lambda_j \sim K(\sigma^2 \lambda_j), \lambda_j | \xi \sim \text{Ga}(a_\pi, \xi), \xi \sim \text{Ga}(b, 1),
$$
or by applying Proposition 1 again, it can also be represented as

\[ \beta_j | \sigma^2, \lambda_j \sim K(\sigma^2 \lambda_j), \lambda_j \sim \text{BP}(a, b). \]

**Normal kernel**

The class of R2-D2 priors relies on the kernel density \( K \). The R2-D2 prior with normal kernel and \( a = pa_\pi \) is given by

\[ \beta_j | \sigma^2, \lambda_j \sim N(0, \sigma^2 \lambda_j), \lambda_j | \xi \sim \text{Ga}(a_\pi, \xi), \xi \sim \text{Ga}(b, 1). \]

This is a special case of the general normal-gamma priors as proposed in Griffin et al. (2010), by keeping the shape hyperparameter in the Gamma prior for the variance coefficients, \( a_\pi \), fixed, and the rate hyperparameter, \( \xi \), given a particular Gamma hyperprior.

Another equivalent form is

\[ \beta_j | \sigma^2, \lambda_j \sim N(0, \sigma^2 \lambda_j), \lambda_j \sim \text{BP}(a, b), \]

and the density of \( \sqrt{\lambda_j} \) is

\[ \pi(x) = \frac{2 \Gamma(a_\pi + b) x^{2a_\pi - 1}}{\Gamma(a_\pi) \Gamma(b) (1 + x^2)^{a_\pi + b}}. \]

When \( a_\pi = a/p = b = 1/2 \), this is the standard half-Cauchy distribution, i.e., \( \mathcal{C}^+(0,1) \), then the R2-D2 prior is written as

\[ \beta_j | \sigma^2, \tau_j \sim N(0, \sigma^2 \tau_j^2), \tau_j \sim \mathcal{C}^+(0, 1), \]
which is a special case of the Horseshoe prior proposed in (Carvalho et al., 2009) with global shrinkage parameter fixed at 1.

**Double-exponential kernel**

As shown in Section 3.2.2, the choice of normal kernel gives the special case of the normal-gamma family and Horseshoe prior. However, the double-exponential (DE) distribution has more mass around zero and heavier tails than the normal distribution. Thus, to encourage shrinkage, it is reasonable to replace the normal kernel with a DE kernel, i.e., \( \beta_j | \sigma^2, \phi_j, \omega \sim \text{DE}(\sigma \sqrt{\phi_j \omega/2}) \) for \( j = 1, \cdots, p \), with \( \text{DE}(\delta) \) denoting a double-exponential distribution with mean 0 and variance \( 2\delta^2 \). The prior is then summarized as follows:

\[
\beta_j | \sigma^2, \phi_j, \omega \sim \text{DE}(\sigma \sqrt{\phi_j \omega/2}), \ \phi \sim \text{Dir}(a_\pi, \cdots, a_\pi), \ \omega \sim \text{BP}(a, b). \tag{3.5}
\]

In this global-local shrinkage prior, \( \omega \) controls the global shrinkage degree through \( a \) and \( b \), while \( \phi_j \) controls the local shrinkage through \( a_\pi \). In particular, when \( a_\pi \) is small, the prior would lead to large variability between the proportions \( \phi_j \)'s, thus more shrinkage for the regression coefficients; while when \( a_\pi \) is large, less shrinkage is assumed.

Given \( a = p a_\pi \), the R2-D2 prior can also be equivalently written as:

\[
\beta_j | \sigma^2, \lambda_j \sim \text{DE}(\sigma \sqrt{\lambda_j/2}), \ \lambda_j | \xi \sim \text{Ga}(a_\pi, \xi), \ \xi \sim \text{Ga}(b, 1), \tag{3.6}
\]

or

\[
\beta_j | \sigma^2, \lambda_j \sim \text{DE}(\sigma \sqrt{\lambda_j/2}), \ \lambda_j \sim \text{BP}(a_\pi, b). \tag{3.7}
\]

We focus on this double-exponential kernel-based prior for the remainder of the chapter.
3.2.3 Posterior Computation

For posterior computation, the following equivalent representation is useful. The R2-D2 prior (3.5) is equivalent to

\[
\beta_j | \sigma^2, \psi_j, \phi_j, \omega \sim \mathcal{N}(0, \psi_j \phi_j \omega \sigma^2 / 2), \quad \psi_j \sim \text{Exp}(1/2),
\]
\[
\phi \sim \text{Dir}(a_\pi, \ldots, a_\pi), \quad \omega | \xi \sim \text{Ga}(a, \xi), \quad \xi \sim \text{Ga}(b, 1),
\]

(3.8)

where \(\text{Exp}(\delta)\) denotes the exponential distribution with mean \(\delta^{-1}\).

The Gibbs sampling procedure is based on (3.8) with \(a = pa_\pi\). Assume the variance has prior \(\sigma^2 \sim \text{IG}(a_1, b_1)\), an inverse Gamma distribution with shape and scale parameters \(a_1\) and \(b_1\) respectively. Denote \(Z \sim \text{InvGaussian}(\mu, \lambda)\), the inverse Gaussian distribution, if \(\pi(z) = \sqrt{\lambda/(2\pi z^3)} \exp\{-\lambda(z - \mu)^2/(2\mu^2 z)\}\). Denote \(Z \sim \text{giG}(\chi, \rho, \lambda_0)\), the generalized inverse Gaussian distribution, if \(\pi(z) \propto z^{\lambda_0 - 1} \exp\{-\rho z + \chi/z\}/2\). The Gibbs sampling procedure is as follows:

(i) Sample \(\beta|\psi, \phi, \omega, \sigma^2, Y \sim N(\mu, \sigma^2 V)\), where \(\mu = VX^T Y = (X^T X + S^{-1})^{-1}(X^T Y)\),
\[
V = (X^T X + S^{-1})^{-1}, \quad S = \text{diag}\{\psi_1 \phi_1 \omega/2, \ldots, \psi_p \phi_p \omega/2\}, \quad X = (x_1, \ldots, x_n)^T, \quad \text{and} \quad Y = (Y_1, \ldots, Y_n)^T.
\]

(ii) Sample \(\sigma^2|\beta, \psi, \phi, \omega, Y \sim \text{IG}(a_1 + (n + p)/2, b_1 + (\beta^T S^{-1} \beta + (Y - X \beta)^T (Y - X \beta))/2)\).

(iii) Sample \(\psi|\beta, \phi, \omega, \sigma^2\). One can draw \(\psi_j^{-1} \sim \text{InvGaussian}(\mu_j = \sqrt{\sigma^2 \phi_j \omega / 2} / |\beta_j|, \lambda = 1)\), then take the reciprocal to get \(\psi_j\).

(iv) Sample \(\omega|\beta, \psi, \phi, \xi, \sigma^2 \sim \text{giG}(\chi = \sum_{j=1}^p 2\beta_j^2 / (\sigma^2 \phi_j \psi_j), \rho = 2\xi, \lambda_0 = a - p/2)\).

(v) Sample \(\xi|\omega \sim \text{Ga}(a + b, 1 + \omega)\).
(vi) Sample $\phi|\beta, \psi, \xi, \sigma^2$. Motivated by Bhattacharya et al. (2015), if $a = pa_{\pi}$, one can draw $T_1, \cdots, T_p$ independently with $T_j \sim \text{giG}(\chi = 2\beta_j^2/(\sigma^2\psi_j), \rho = 2\xi, \lambda_0 = a_{\pi} - 1/2)$.

Then set $\phi_j = T_j/T$ with $T = \sum_{j=1}^p T_j$.

### 3.3 Theoretical Properties

In this section, a number of theoretical properties of the proposed R2-D2 prior with the DE kernel are established. The properties of the Horseshoe (HS) (Carvalho et al., 2009, 2010), Horseshoe+ (HS+) (Bhadra et al., 2015), and Dirichlet-Laplace (DL) prior (Bhattacharya et al., 2015) are provided as a comparison. Proofs and technical details are given in the Appendix. For simplicity of comparison across approaches, the variance term $\sigma^2$ is fixed at 1. For the R2-D2 prior, the Dirichlet concentration $a_{\pi}$ is set to $a = pa_{\pi}$, so we consider the R2-D2 prior represented as (3.6) or (3.7) in this section.

#### 3.3.1 Marginal Density of the R2-D2 Prior

**Proposition 3.** Given the R2-D2 prior, the marginal density of $\beta_j$ for any $j = 1, \cdots, p$ is

$$
\pi_{R2-D2}(\beta_j) = \frac{1}{\sqrt{2\pi}\Gamma(a_{\pi})\Gamma(b)} G_{13}^{31}\left(\begin{array}{c}
\frac{1}{2} - b \\
\frac{1}{2} - \frac{a_{\pi}}{2}, 0, 0, 0
\end{array} \mid \begin{array}{c}
b \\frac{1}{2} + b \\
2, 2, 2
\end{array} \right)
$$

where $G_{p,q}^{m,n}(z \mid a_1, \cdots, a_p \mid b_1, \cdots, b_q)$ denotes the Meijer G-function.

Figure 3.1 plots the marginal density function of the R2-D2 density along with the Horseshoe, DL, and Cauchy distributions. The hyperparameters for all distributions are selected to ensure the interquartile range is approximately 1 for comparison. The R2-D2 density has the most mass around zero and the heaviest tails; we formally investigate these asymptotic properties in the following sections. Table 3.1 provides the summary results of these properties.
3.3.2 Asymptotic Tail Behaviors

We examine the asymptotic behaviors of tails of the proposed R2-D2 prior in this section. A prior with heavy tails is desirable in high-dimensional regression to allow the posterior to estimate large values for important predictors.

**Theorem 6.** Given $|\beta| \to \infty$, for any $a_\pi > 0$ and $b > 0$, the marginal density of the R2-D2
Table 3.1: Asymptotic properties for Horseshoe, Horseshoe+, R2-D2 and Dirichlet-Laplace priors as discussed in Section 3.3.

<table>
<thead>
<tr>
<th>Prior</th>
<th>Tail Decay</th>
<th>Concentration at zero</th>
<th>Cesàro-average Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horseshoe</td>
<td>$O\left(\frac{1}{</td>
<td>\beta</td>
<td>^b}\right)$</td>
</tr>
<tr>
<td>Horseshoe+</td>
<td>$O\left(\frac{\log</td>
<td>\beta</td>
<td>}{</td>
</tr>
<tr>
<td>R2-D2</td>
<td>$O\left(\frac{1}{</td>
<td>\beta</td>
<td>^{1+2b}}\right)$</td>
</tr>
<tr>
<td>Dirichlet-Laplace</td>
<td>$O\left(\frac{</td>
<td>\beta</td>
<td>^{aD/2-3/4}}{\exp\left(\sqrt{2</td>
</tr>
</tbody>
</table>

prior satisfies $\pi_{R2-D2}(\beta) = O(1/|\beta|^{2b+1})$. Furthermore, when $0 < b < 1/2$,

$$\lim_{|\beta|\to\infty} \frac{\pi_{R2-D2}(\beta)}{|\beta|^{-2}} = \infty,$$

i.e., the R2-D2 prior has heavier tails than the Cauchy distribution.

As a comparison, we also study the tail behavior of the Dirichlet-Laplace prior. The DL prior proposed in Bhattacharya et al. (2015) is

$$\beta_j|\psi_j \sim \text{DE}(\psi_j), \quad \psi_j \sim \text{Ga}(a_D, 1/2).$$

(3.9)

**Theorem 7.** Given $|\beta| \to \infty$, for any $a_D > 0$, the marginal density of the Dirichlet-Laplace prior satisfies $\pi_{DL}(\beta) = O(|\beta|^{a_D/2-3/4}/\exp\{\sqrt{2|\beta|}\})$. Furthermore, the DL prior has lighter tails than the Cauchy distribution, i.e., $\lim_{|\beta|\to\infty} \frac{\pi_{DL}(\beta)}{|\beta|^{-2}} = 0$.

As noted in Carvalho et al. (2010), the Horseshoe prior has exact Cauchy-like tails that decay like $\beta^{-2}$, and the Horseshoe+ prior has a tail of $O(\log |\beta|/\beta^2)$ as illustrated in the proof of Theorem 4.6 in Bhadra et al. (2015). Therefore, the proposed R2-D2 prior leads to the heaviest tail, followed by Horseshoe+, then Horseshoe, and finally the DL prior. With a polynomial
tail heavier than Cauchy distribution, the R2-D2 prior attains a substantial improvement over a large class of global-local shrinkage priors.

### 3.3.3 Concentration Properties

In this section, we study the concentration properties of the R2-D2 prior around the origin. The concentration properties of DL, Horseshoe, and Horseshoe+ priors are also given. We prefer priors with high concentration near zero to reflect the prior that most of the covariates do not have a substantial effect on the response. We now show that the proposed R2-D2 prior has higher concentration at zero to go along with heavier tails than other GL priors.

**Theorem 8.** As $|\beta| \to 0$, if $0 < a_\pi < 1/2$, the marginal density of the R2-D2 prior satisfies

$$
\pi_{R2-D2}(\beta) = O(1/|\beta|^{1-2a_\pi}).
$$

**Theorem 9.** As $|\beta| \to 0$, if $0 < a_D < 1$, the marginal density of the Dirichlet-Laplace prior satisfies

$$
\pi_{DL}(\beta) = O(1/|\beta|^{1-a_D}).
$$

For the Horseshoe prior, as summarized in Carvalho et al. (2010), the marginal density

$$
\pi_{HS}(\beta) = \exp\{\beta^2/2\}E_1(\beta^2/2)/\sqrt{2\pi^3},
$$

where $E_1(z) = \int_1^\infty e^{-zt}/t \, dt$ is the exponential integral function. As $|\beta| \to 0$,

$$
\frac{1}{2\sqrt{2\pi^3}} \log(1 + \frac{4}{\beta^2}) \leq \pi_{HS}(\beta) \leq \frac{1}{\sqrt{2\pi^3}} \log(1 + \frac{2}{\beta^2}).
$$

Therefore around the origin, $\pi_{HS}(\beta) = O(\log(1/|\beta|))$. Also by the proof of Theorem 4.6 in Bhattacharya et al. (2015), as $|\beta| \to 0$, the marginal density of Horseshoe+ prior satisfies $\pi_{HS+}(\beta) = O(\log^2(1/|\beta|))$. It is clear that $2a_\pi$ in the R2-D2 prior plays the same role around
origin as $a_D$ in the DL prior. Accordingly, when given $a_D = 2a_\pi \in (0, 1)$, all four priors (R2-D2, DL, HS, and HS+) possess unbounded density near the origin. However, the R2-D2 and DL prior diverge to infinity with a polynomial order, much faster than the Horseshoe+ (with a squared logarithm order) and the Horseshoe prior (with a logarithm order).

Now we see that the R2-D2 and DL prior put more mass in a small neighborhood of zero compared to the Horseshoe and Horseshoe+ prior. Polson and Scott (2010) established that when the truth is zero, a prior with unbounded density near zero is super-efficient in terms of the Kullback-Leibler risk. The more mass the prior puts around the neighborhood of the origin, the more efficient. Then the four priors are all super-efficient, with R2-D2 and DL more efficient than the HS+, and followed by HS. The following section discusses it in detail.

### 3.3.4 Predictive Efficiency

In this section, we study the predictive efficiency of the shrinkage priors. We focus on the case when the design matrix $X = (x_1^T, \cdots, x_n^T)^T$ is orthogonal, i.e., $X^T X = I_p$. We also assume $\sigma^2$ to be known. Though this may not be a realistic setup in practice, it provides some insight and motivation for measuring the predictive efficiency. In this case, the sufficient statistic for $\beta$ is the ordinary least square estimate, i.e., $\hat{\beta} = X^T Y$, with $\hat{\beta} \sim N(\beta, \sigma^2 I_p)$. For simplicity of notation, without loss of generality, assume $\beta_0$ is the true parameter, we rewrite the sampling model as

$$y_i \sim N(\beta_0, \sigma^2)$$

independently for $i = 1, \cdots, n$. Similar as Carvalho et al. (2010) and Bhadra et al. (2015), we use the Kullback-Leibler divergence between the true model and the Bayes estimator (or the posterior mean estimator) of the density function to measure the predictive efficiency. Denote $\pi(y|\beta)$ as the sampling model (3.10), and $KL(\pi_1, \pi_2) = E_{\pi_1}(\log(\pi_1/\pi_2))$ as the Kullback-
Leibler divergence of $\pi_2$ from $\pi_1$. The results are based on the following Lemma.

**Lemma 6.** (Clarke and Barron, 1990) Define $A_\epsilon = \{\beta : KL(\pi_{\beta_0}, \pi_\beta) \leq \epsilon\}$ as the Kullback-Leibler information set of measure $\epsilon$ centered at $\beta_0$. Let $\mu(d\beta)$ be the prior measure, the observed data $y(n) = (y_1, \ldots, y_n)$, the corresponding posterior distribution is $\mu_n(d\beta|y(n))$, and posterior predictive density $\hat{\pi}_n(y) = \int \pi(y|\beta)\mu_n(d\beta|y(n))$. Assume that $\mu(A_\epsilon) > 0$ for any $\epsilon > 0$, then at $\pi_{\beta_0}$, the prior $\mu(d\beta)$ is information dense. Then the Cesàro-average risk of the Bayes estimator $\hat{\pi}_n$, defined as $R_n = n^{-1} \sum_{j=1}^{n} KL(\pi_{\beta_0}, \hat{\pi}_j)$, satisfies $R_n \leq \epsilon - \log \mu(A_\epsilon)/n$.

Carvalho et al. (2010) proved that when the true parameter $\beta_0$ is zero, the Cesàro-average risk of the maximum likelihood estimator (MLE) is $R_n = O(n^{-1} \log n)$; while the Horseshoe estimator’s risk is $R_n = O\left(n^{-1} \log \left(n/(\log n)^{b_0}\right)\right)$, where $b_0$ is a constant, lower than the MLE. In this sense, the Horseshoe estimator for the sampling density converges to the true model at a super-efficient rate. Bhadra et al. (2015) shows that the Horseshoe+ estimator slightly improves the rate with $R_n = O\left(n^{-1} \log \left(n/(\log n)^{2b_0}\right)\right)$. In this section, we illustrate that our proposed R2-D2 prior achieves a smaller risk. Our result is based on the following theorem.

**Theorem 10.** For $0 < \alpha_\pi < 1/2$, when $\beta_0 = 0$, the Cesàro-average risk of the Bayes estimator under the R2-D2 prior satisfies

$$R_n(R2-D2) = O\left(\frac{1}{n} \log \left(\frac{n}{n^{1/2-\alpha_\pi}}\right)\right).$$

As a complement, we also give the risk bounds for the DL prior.

**Theorem 11.** For $0 < \alpha_D < 1$, when $\beta_0 = 0$, the Cesàro-average risk of the Bayes estimator under the DL prior satisfies

$$R_n(DL) = O\left(\frac{1}{n} \log \left(\frac{n}{n^{1/2-\alpha_D/2}}\right)\right).$$
Therefore the R2-D2 and DL priors have smaller Kullback-Leibler risk than the Horseshoe and Horseshoe+ priors. Hence our proposed R2-D2 prior achieves success in both origin and tails, and demonstrates superior performance in prediction.

### 3.3.5 Posterior Consistency

In this section, we show that R2-D2 prior yields posterior consistency. Assume the true regression parameter is $\beta_n^0$, and the regression parameter $\beta_n$ is given some shrinkage prior. If the posterior of $\beta_n$ converges in probability towards $\beta_n^0$, i.e., for any $\epsilon > 0$, $P(|\beta_n - \beta_n^0| > \epsilon|Y) \rightarrow 0$ as $p_n, n \rightarrow \infty$, we say the prior yields a consistent posterior.

Assume the following regularity conditions:

(A1) $p_n = o(n)$;

(A2) Let $d_{p_n}$ and $d_1$ be the smallest and the largest singular values of $X^TX/n$ respectively. Assume $0 < d_{\min} < \lim \inf_{n \rightarrow \infty} d_{p_n} \leq \lim \sup_{n \rightarrow \infty} d_1 < d_{\max} < \infty$, where $d_{\min}$ and $d_{\max}$ are fixed;

(A3) $\lim \sup_{j=1,\cdots,p_n} |\beta_{nj}^0| < \infty$;

(A4) $q_n = o(n/\log n)$, in which $q_n$ is the number of nonzero components in $\beta_n^0$.

**Theorem 12.** Under assumptions (A1)-(A4), for any $b > 0$, given the linear regression model, and $a_n = C/(p_n^{b/2} n^{\rho b/2} \log n)$ for finite $\rho > 0$ and $C > 0$, the R2-D2 prior yields a consistent posterior.
3.4 Simulation Study

To illustrate the performance of the proposed R2-D2 prior, we conduct a simulation study with various number of predictors and effect size. In each setting, 200 datasets are simulated from the linear model (3.1) with $\sigma^2 = 1$, sample size $n$ fixed at 60 to match the real data example in Section 3.5, and the number of predictors $p$ varying in $p \in \{50, 100, 500\}$. The covariates $x_i, i = 1, \cdots, n$, are generated from multivariate normal distribution with mean zero, and correlation matrix of AR(1) structure with correlation $\rho = 0.5$ or 0.9. For the regression coefficients $\beta$, we consider the following two setups.

**Setup 1:** $\beta = (0_{10}^T, B_1^T, 0_{50}^T, B_2^T, 0_{p=50}^T)^T$ with $0_k$ representing the zero vector of length $k$. $B_1$ and $B_2$ are each of length 5 with elements generated independently from the Student $t$ distribution with 3 degrees of freedom, which is to mimic the heavy tail distribution. Then the theoretical $R^2$ as in equation (3.3) is 0.97.

**Setup 2:** $\beta = (0_{10}^T, s^*B_3^T, 0_{p=15}^T)^T$ with $B_3$ also of length 5 with elements generated independently from the Student $t$ distribution with 3 degrees of freedom, with $s^* = \sqrt{1/15}$ to ensure the total prior variance of $\beta$ is 1 and hence the theoretical $R^2$ as in equation (3.3) is 0.5.

We consider $p = 50, 100, 500$ for setup 1, and only $p = 100$ for setup 2 because the other cases perform similarly. Setup 2 is designed to study the performance of the proposed R2-D2 prior with known $R^2$ information. For each simulated dataset, we use different shrinkage priors for $\beta$. The priors are Horseshoe (HS), Horseshoe+ (HS+), R2-D2$_{(0.5, 0.5)}$ with $a = 0.5, b = 0.5, a_\pi = 1/(2p)$, R2-D2$_{(p/n, 0.5)}$ with $a = p/n, b = 0.5, a_\pi = 1/n$, R2-D2$_{(p/n, 0.1)}$ with $a = p/n, b = 0.1, a_\pi = 1/n$, R2-D2$_{(1, 1)}$ with $a = 1, b = 1, a_\pi = 1/p$, DL$_{1/p}$ with $a_D = 1/p$ and DL$_{2/n}$ with $a_D = 2/n$ and DL$_{1/n}$ with $a_D = 1/n$. For the HS and HS+, MCMC steps are implemented through Stan in R using the code provided by the author of Bhadra et al. (2015). For the R2-D2 and DL, Gibbs samplers are implemented in R. 10,000 samples are collected.
with the first 5,000 samples discarded as burn-in. Table 3.2 gives the CPU time for running 10,000 samples in the same computer for one data set. The computational advantages of the R2-D2 and DL are obvious.

Table 3.2:  CPU time (in hours) for a single dataset to generate 10,000 posterior samples with \( n = 60 \) and varying \( p \) for different shrinkage priors.

<table>
<thead>
<tr>
<th>Prior</th>
<th>( p = 50 )</th>
<th>( p = 100 )</th>
<th>( p = 500 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horseshoe</td>
<td>0.103</td>
<td>0.441</td>
<td>1.875</td>
</tr>
<tr>
<td>Horseshoe+</td>
<td>0.235</td>
<td>0.502</td>
<td>2.294</td>
</tr>
<tr>
<td>R2-D2</td>
<td>0.008</td>
<td>0.015</td>
<td>0.297</td>
</tr>
<tr>
<td>Dirichlet-Laplace</td>
<td>0.008</td>
<td>0.015</td>
<td>0.297</td>
</tr>
</tbody>
</table>

The average value of the sum squared error (SSE) corresponding to the posterior mean across the 200 replicates is provided in Table 3.3 with \( \rho = 0.5 \) and Table 3.4 with \( \rho = 0.9 \) for simulation setup 1. Table 3.5 provides the results for simulation setup 2. Tables 3.3, 3.4 and 3.5 give the total SSE as well as the SSE split into three pieces according to the value of the true \( \beta \) at \( \beta_j = 0, |\beta_j| \in (0, 0.5], \) and \( |\beta_j| > 0.5, j = 1, \cdots, p. \) The averaged area under the Receiver-Operating Characteristic (ROC) curve based on the posterior \( t \)-statistic, i.e., the ratio of the posterior mean and posterior standard deviation, is also given to offer further evaluation of the variable selection performance. We measure the reliability of the ordering of the magnitude of the posterior \( t \)-statistic through the area under the ROC curve, which is labeled as “ROC” in the tables.

Overall, the R2-D2 prior with \( a = p/n \) (or \( a_\pi = 1/n \)) has similar SSE to the HS and HS+ prior, and smaller than the DL prior. HS and HS+ yield good estimators with small SSE. However, HS and HS+ generally have lower ROC area, worse than the R2-D2 and DL priors.
This may be explained by investigating the SSE for zero, small coefficients and large coefficients. Although HS and HS+ estimate the nonzero coefficients quite well, they estimate the zero coefficients poorly, which leads to more false positives and poor ROC performance. This corresponds with the poor concentration properties of the HS and HS+ as discussed in Section 3.3.3. In addition, we also conduct simulations for fixed $p$ with varying $n$, and the performance is similar.

Furthermore, the DL priors exhibit excellent performance in estimating the zero coefficients, but poor estimates of large coefficients, leading to large total SSE. This is due to the good concentration (Section 3.3.3) while relevantly light tail property of DL priors (Section 3.3.2). However, inaccurate estimation at large coefficients does not greatly affect the ROC performance, which is comparable to the R2-D2 priors. For the R2-D2 priors, the value of $b$ slightly affects the estimation. By analogy of the R2-D2 priors with $a = p/n, b = 0.5$ and $a = p/n, b = 0.1$, we gain a key insight that a smaller value of $b$ results in slightly smaller total SSE due to the better estimation at large coefficients. This coincides with the fact that $b$ controls the tail behavior (Section 3.3.2), with smaller $b$ giving heavier tails. $a = 1$ (or $a_\pi = 1/p$) and $a = 0.5$ (or $a_\pi = 1/(2p)$) lead to smaller SSE at zero coefficients than $a = p/n$ (or $a_\pi = 1/n$) when $p > n$, which again, matches the concentration properties of R2-D2 priors as described in Section 3.3.3. There is no significant difference in variable selection performance for the four parametrizations. In all, the R2-D2 prior with $a_\pi = 1/n$ and $b < 0.5$ achieves success in both estimation and variable selection, demonstrating distinguishable performance from the HS, HS+ and DL priors.
Table 3.3: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| \in (0, 0.5]$, and $|\beta_j| > 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the ROC curve, based on 200 datasets generated by Setup 1. Here $n = 60$ and $\rho = 0.5$. Standard errors are in parentheses.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Prior</th>
<th>SSE($\beta = 0$)</th>
<th>SSE((0, 0.5])</th>
<th>SSE(&gt; 0.5)</th>
<th>SSE(Total)</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>HS</td>
<td>0.19 (0.009)</td>
<td>0.12 (0.007)</td>
<td>0.48 (0.029)</td>
<td>0.78 (0.030)</td>
<td>0.89 (0.005)</td>
</tr>
<tr>
<td></td>
<td>HS+</td>
<td>0.15 (0.017)</td>
<td>0.13 (0.007)</td>
<td>0.46 (0.028)</td>
<td>0.74 (0.034)</td>
<td>0.89 (0.005)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(0.5,0.5)</td>
<td>0.05 (0.003)</td>
<td>0.14 (0.008)</td>
<td>0.60 (0.037)</td>
<td>0.79 (0.037)</td>
<td>0.90 (0.005)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0,5)</td>
<td>0.05 (0.003)</td>
<td>0.14 (0.007)</td>
<td>0.55 (0.034)</td>
<td>0.74 (0.034)</td>
<td>0.90 (0.005)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.1)</td>
<td>0.05 (0.003)</td>
<td>0.14 (0.007)</td>
<td>0.58 (0.035)</td>
<td>0.77 (0.035)</td>
<td>0.90 (0.005)</td>
</tr>
<tr>
<td></td>
<td>DL1/p</td>
<td>0.06 (0.003)</td>
<td>0.14 (0.007)</td>
<td>0.54 (0.033)</td>
<td>0.74 (0.033)</td>
<td>0.90 (0.005)</td>
</tr>
<tr>
<td></td>
<td>DL2/n</td>
<td>0.02 (0.004)</td>
<td>0.18 (0.009)</td>
<td>1.52 (0.106)</td>
<td>1.72 (0.107)</td>
<td>0.89 (0.005)</td>
</tr>
<tr>
<td></td>
<td>DL1/n</td>
<td>0.01 (0.003)</td>
<td>0.19 (0.009)</td>
<td>1.78 (0.118)</td>
<td>1.99 (0.118)</td>
<td>0.89 (0.005)</td>
</tr>
<tr>
<td>100</td>
<td>HS</td>
<td>0.32 (0.032)</td>
<td>0.16 (0.009)</td>
<td>0.65 (0.038)</td>
<td>1.13 (0.055)</td>
<td>0.87 (0.006)</td>
</tr>
<tr>
<td></td>
<td>HS+</td>
<td>0.22 (0.020)</td>
<td>0.16 (0.010)</td>
<td>0.63 (0.041)</td>
<td>1.02 (0.052)</td>
<td>0.88 (0.006)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(0.5,0.5)</td>
<td>0.06 (0.004)</td>
<td>0.17 (0.010)</td>
<td>0.90 (0.057)</td>
<td>1.13 (0.057)</td>
<td>0.89 (0.006)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.5)</td>
<td>0.11 (0.006)</td>
<td>0.16 (0.009)</td>
<td>0.70 (0.044)</td>
<td>0.96 (0.045)</td>
<td>0.89 (0.006)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.1)</td>
<td>0.11 (0.006)</td>
<td>0.16 (0.009)</td>
<td>0.69 (0.045)</td>
<td>0.96 (0.046)</td>
<td>0.89 (0.006)</td>
</tr>
<tr>
<td></td>
<td>DL1/p</td>
<td>0.08 (0.005)</td>
<td>0.17 (0.010)</td>
<td>0.77 (0.050)</td>
<td>1.02 (0.051)</td>
<td>0.89 (0.006)</td>
</tr>
<tr>
<td></td>
<td>DL2/n</td>
<td>0.03 (0.002)</td>
<td>0.20 (0.010)</td>
<td>1.75 (0.103)</td>
<td>1.97 (0.103)</td>
<td>0.88 (0.006)</td>
</tr>
<tr>
<td></td>
<td>DL1/n</td>
<td>0.02 (0.001)</td>
<td>0.21 (0.010)</td>
<td>2.22 (0.134)</td>
<td>2.44 (0.134)</td>
<td>0.88 (0.006)</td>
</tr>
<tr>
<td>500</td>
<td>HS</td>
<td>0.79 (0.054)</td>
<td>0.21 (0.014)</td>
<td>1.22 (0.077)</td>
<td>2.22 (0.112)</td>
<td>0.82 (0.006)</td>
</tr>
<tr>
<td></td>
<td>HS+</td>
<td>0.63 (0.047)</td>
<td>0.21 (0.011)</td>
<td>1.14 (0.074)</td>
<td>1.99 (0.096)</td>
<td>0.83 (0.007)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(0.5,0.5)</td>
<td>0.04 (0.002)</td>
<td>0.23 (0.011)</td>
<td>3.76 (0.300)</td>
<td>4.02 (0.299)</td>
<td>0.85 (0.006)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.5)</td>
<td>0.17 (0.008)</td>
<td>0.20 (0.010)</td>
<td>1.56 (0.095)</td>
<td>1.93 (0.100)</td>
<td>0.87 (0.005)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.1)</td>
<td>0.19 (0.008)</td>
<td>0.20 (0.010)</td>
<td>1.46 (0.088)</td>
<td>1.84 (0.093)</td>
<td>0.87 (0.005)</td>
</tr>
<tr>
<td></td>
<td>DL1/p</td>
<td>0.05 (0.003)</td>
<td>0.22 (0.011)</td>
<td>2.81 (0.237)</td>
<td>3.08 (0.237)</td>
<td>0.86 (0.006)</td>
</tr>
<tr>
<td></td>
<td>DL2/n</td>
<td>0.11 (0.007)</td>
<td>0.21 (0.010)</td>
<td>2.35 (0.142)</td>
<td>2.68 (0.145)</td>
<td>0.86 (0.006)</td>
</tr>
<tr>
<td></td>
<td>DL1/n</td>
<td>0.18 (0.095)</td>
<td>0.22 (0.011)</td>
<td>3.16 (0.188)</td>
<td>3.57 (0.217)</td>
<td>0.86 (0.006)</td>
</tr>
</tbody>
</table>

3.5 PCR Data Analysis

We now analyze the mouse gene expression data collected by Lan et al. (2006), consisting of the expression levels of 22,575 genes for $n = 60$ (31 female and 29 male) mice. Real-time PCR
Table 3.4: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| \in (0, 0.5]$, and $|\beta_j| > 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the ROC curve, based on 200 datasets generated by Setup 1. Here $n = 60$ and $\rho = 0.9$. Standard errors are in parentheses.

<table>
<thead>
<tr>
<th>Prior</th>
<th>$p = 50$</th>
<th>$p = 100$</th>
<th>$p = 500$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2-D2(0.5,0.5)</td>
<td>0.08 (0.012)</td>
<td>0.26 (0.021)</td>
<td>3.41 (0.194)</td>
</tr>
<tr>
<td>R2-D2(p/n,0.5)</td>
<td>0.10 (0.016)</td>
<td>0.26 (0.022)</td>
<td>3.22 (0.185)</td>
</tr>
<tr>
<td>R2-D2(p/n,0.1)</td>
<td>0.10 (0.015)</td>
<td>0.26 (0.022)</td>
<td>3.23 (0.185)</td>
</tr>
<tr>
<td>DL1/p</td>
<td>0.03 (0.004)</td>
<td>0.27 (0.018)</td>
<td>4.91 (0.256)</td>
</tr>
<tr>
<td>DL2/n</td>
<td>0.04 (0.005)</td>
<td>0.26 (0.017)</td>
<td>4.67 (0.247)</td>
</tr>
<tr>
<td>DL1/n</td>
<td>0.04 (0.004)</td>
<td>0.27 (0.021)</td>
<td>5.07 (0.264)</td>
</tr>
<tr>
<td>DL1/p</td>
<td>0.04 (0.007)</td>
<td>0.28 (0.025)</td>
<td>5.65 (0.288)</td>
</tr>
<tr>
<td>DL2/n</td>
<td>0.05 (0.006)</td>
<td>0.27 (0.019)</td>
<td>4.78 (0.243)</td>
</tr>
<tr>
<td>DL1/n</td>
<td>0.05 (0.011)</td>
<td>0.29 (0.025)</td>
<td>5.47 (0.279)</td>
</tr>
</tbody>
</table>

was used to measure some physiological phenotypes, including number of phosphoenopyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1). We build regression models for the three phenotypes using gender and...
Table 3.5: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| \in (0, 0.5]$, and $|\beta_j| > 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the ROC curve, based on 200 datasets generated by Setup 2. Here $n = 60$ and $p = 100$. Standard errors are in parentheses.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Prior</th>
<th>SSE($= 0$)</th>
<th>SSE((0, 0.5])</th>
<th>SSE(&gt; 0.5)</th>
<th>SSE(Total)</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>HS</td>
<td>0.04 (0.004)</td>
<td>0.19 (0.008)</td>
<td>0.08 (0.010)</td>
<td>0.27 (0.012)</td>
<td>0.72 (0.010)</td>
</tr>
<tr>
<td></td>
<td>HS+</td>
<td>0.04 (0.004)</td>
<td>0.19 (0.008)</td>
<td>0.08 (0.010)</td>
<td>0.27 (0.012)</td>
<td>0.72 (0.010)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(0.5,0.5)</td>
<td>0.01 (0.001)</td>
<td>0.18 (0.008)</td>
<td>0.13 (0.016)</td>
<td>0.30 (0.017)</td>
<td>0.74 (0.011)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.5)</td>
<td>0.03 (0.002)</td>
<td>0.18 (0.007)</td>
<td>0.09 (0.011)</td>
<td>0.26 (0.012)</td>
<td>0.73 (0.010)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.1)</td>
<td>0.03 (0.002)</td>
<td>0.18 (0.007)</td>
<td>0.09 (0.011)</td>
<td>0.26 (0.012)</td>
<td>0.73 (0.010)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(1,1)</td>
<td>0.02 (0.001)</td>
<td>0.17 (0.007)</td>
<td>0.10 (0.013)</td>
<td>0.28 (0.014)</td>
<td>0.73 (0.010)</td>
</tr>
<tr>
<td></td>
<td>DL1/p</td>
<td>0.00 (0.000)</td>
<td>0.18 (0.008)</td>
<td>0.14 (0.018)</td>
<td>0.32 (0.018)</td>
<td>0.74 (0.010)</td>
</tr>
<tr>
<td></td>
<td>DL2/n</td>
<td>0.01 (0.001)</td>
<td>0.17 (0.007)</td>
<td>0.11 (0.014)</td>
<td>0.28 (0.015)</td>
<td>0.74 (0.011)</td>
</tr>
<tr>
<td></td>
<td>DL1/n</td>
<td>0.01 (0.001)</td>
<td>0.17 (0.007)</td>
<td>0.13 (0.017)</td>
<td>0.31 (0.017)</td>
<td>0.74 (0.010)</td>
</tr>
<tr>
<td>0.9</td>
<td>HS</td>
<td>0.02 (0.004)</td>
<td>0.23 (0.011)</td>
<td>0.23 (0.025)</td>
<td>0.46 (0.027)</td>
<td>0.73 (0.012)</td>
</tr>
<tr>
<td></td>
<td>HS+</td>
<td>0.03 (0.004)</td>
<td>0.24 (0.011)</td>
<td>0.23 (0.025)</td>
<td>0.47 (0.028)</td>
<td>0.73 (0.011)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(0.5,0.5)</td>
<td>0.01 (0.001)</td>
<td>0.21 (0.009)</td>
<td>0.24 (0.026)</td>
<td>0.44 (0.027)</td>
<td>0.78 (0.013)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.5)</td>
<td>0.02 (0.003)</td>
<td>0.22 (0.009)</td>
<td>0.22 (0.024)</td>
<td>0.44 (0.027)</td>
<td>0.76 (0.012)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(p/n,0.1)</td>
<td>0.02 (0.003)</td>
<td>0.22 (0.009)</td>
<td>0.22 (0.024)</td>
<td>0.43 (0.026)</td>
<td>0.75 (0.012)</td>
</tr>
<tr>
<td></td>
<td>R2-D2(1,1)</td>
<td>0.01 (0.002)</td>
<td>0.21 (0.009)</td>
<td>0.23 (0.025)</td>
<td>0.44 (0.027)</td>
<td>0.77 (0.013)</td>
</tr>
<tr>
<td></td>
<td>DL1/p</td>
<td>0.01 (0.001)</td>
<td>0.21 (0.009)</td>
<td>0.24 (0.026)</td>
<td>0.45 (0.029)</td>
<td>0.79 (0.013)</td>
</tr>
<tr>
<td></td>
<td>DL2/n</td>
<td>0.02 (0.002)</td>
<td>0.20 (0.008)</td>
<td>0.22 (0.025)</td>
<td>0.42 (0.026)</td>
<td>0.77 (0.012)</td>
</tr>
<tr>
<td></td>
<td>DL1/n</td>
<td>0.01 (0.001)</td>
<td>0.21 (0.010)</td>
<td>0.24 (0.027)</td>
<td>0.45 (0.029)</td>
<td>0.79 (0.013)</td>
</tr>
</tbody>
</table>

genetic covariates as predictors. The data can be found at http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330.

The 22,575 genes were first screened down to 999 genes based on the ordering of the magnitude of the marginal correlation with the response. Then for each of the 3 regressions, the data set contains $p = 1,000$ predictors with 999 genes and gender and $n = 60$ observations. After screening, linear regression with different global-local shrinkage priors were conducted. To evaluate the performance of different priors, the 60 observations are randomly split into a training set of size 55 and testing set of size 5. Table 3.6 gives the mean squared prediction
error based on 100 random splits of the data. The posterior mean of the regression coefficients from the training set served as the estimate of $\beta$ to make prediction for the testing set. Overall, the results agree with the simulation studies. Our proposed R2-D2 prior with $a = p/n$ performs the best in all cases, and the value of $b$ has little effect on the results.

Table 3.6: Mean squared prediction error, with standard errors in parenthesis, based on 100 random splits of the real data.

<table>
<thead>
<tr>
<th></th>
<th>PEPCK</th>
<th>GPAT</th>
<th>SCD1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>0.148(0.0094)</td>
<td>0.504(0.0364)</td>
<td>0.361(0.0350)</td>
</tr>
<tr>
<td>HS+</td>
<td>0.154(0.0115)</td>
<td>0.646(0.0517)</td>
<td>0.354(0.0277)</td>
</tr>
<tr>
<td>R2-D2(0.5,0.5)</td>
<td>0.162(0.0080)</td>
<td>0.469(0.0354)</td>
<td>0.421(0.0369)</td>
</tr>
<tr>
<td>R2-D2(p/n,0.5)</td>
<td>0.114(0.0065)</td>
<td>0.396(0.0291)</td>
<td>0.281(0.0248)</td>
</tr>
<tr>
<td>R2-D2(p/n,0.1)</td>
<td>0.115(0.0064)</td>
<td>0.403(0.0298)</td>
<td>0.280(0.0253)</td>
</tr>
<tr>
<td>R2-D2(1,1)</td>
<td>0.150(0.0075)</td>
<td>0.450(0.0340)</td>
<td>0.380(0.0329)</td>
</tr>
<tr>
<td>DL_{1/p}</td>
<td>0.316(0.0431)</td>
<td>0.537(0.0459)</td>
<td>0.590(0.0831)</td>
</tr>
<tr>
<td>DL_{1/n}</td>
<td>0.164(0.0167)</td>
<td>0.452(0.0325)</td>
<td>0.354(0.0339)</td>
</tr>
</tbody>
</table>

In addition, we also compared the agreement between methods in terms of variable selection. For each regression, we apply different shrinkage priors on the full data set, then posterior samples of $\beta$ are collected. For each $\beta_j$ ($j = 1, \cdots, p$), the posterior $t$-statistic is calculated by dividing the mean with the standard deviation of those posterior samples. The predictors are ordered by the magnitude of the posterior $t$-statistics from the largest to the smallest. Ideally, the important predictors will be in the beginning of the ordering. Figure 3.2 plots the agreement of the orders between various priors when fit to the full data set for PEPCK. The figures for SCD1 and GPAT perform similarly. Again, it shows that changing the value of $b$ does not result in too much variation of the agreement. In general, the difference of the agreement with different hyperparameter values in the R2-D2 priors is smaller than that of DL prior, and the
Figure 3.2: The agreement of the variable selection between two different priors, based on the induced ordering through the magnitude of posterior $t$-statistics for PEPCK ($n = 60$ and $p = 1000$). The point $(x, y)$ on the line means in the first ordered $x$-length variable list by using the prior on $x$-axis, there are $y$ variables matching with such list by using the prior on $y$-axis.

difference between Horseshoe and Horseshoe+ prior. These results suggest that in terms of variable selection, our proposed R2-D2 prior is less sensitive to hyperparameter selection than the DL or Horseshoe priors.
3.6 Discussion

We propose a class of $R^2$-induced Dirichlet Decomposition (R2-D2) priors in this chapter. The prior starts with placing a Beta prior on $R^2$, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. We demonstrate both theoretically and empirically the advantages of the R2-D2 prior over a number of common shrinkage priors, including the Horseshoe, Horseshoe+, and Dirichlet-Laplace priors. The R2-D2 prior possesses the fastest concentration rate around zero and heaviest tails among the common shrinkage priors, along with the fact that its Bayes estimator is the most efficient. We also show the posterior consistency of the proposed R2-D2 prior.

In this chapter we did not discuss converting the posterior to a sparse solution, but there are several methods to deal with this. Unlike the discrete mixture priors which directly give sparse estimates, GL priors require additional steps to convert the posterior distribution to a sparse solution. Besides the traditional method relying on a choice of threshold to decide which predictors to be included in the model, there are two major approaches: penalized variable selection based on posterior credible regions, and decoupling shrinkage and selection. Bondell and Reich (2012) proposed the penalized credible region variable selection method, which fits the full model under a normal prior, and then selects the sparsest solution within the posterior credible region. Chapter 2 of this thesis extends this approach by replacing the normal prior with any global-local prior. It is easy to adapt our R2-D2 prior to find the sparsest solution using the credible regions. Hahn and Carvalho (2015) proposed the decoupling shrinkage and selection method, which uses a loss function combining a posterior summarizer with an explicit parsimony penalty to induce sparse estimator.
3.7 Proofs

Proof of Proposition 1

Proof.

\[ \pi(\omega) = \int_0^\infty \pi(\omega|\xi)\pi(\xi)d\xi = \int_0^\infty \frac{\xi^a}{\Gamma(a)}\omega^{a-1}e^{-\xi \omega} \frac{1}{\Gamma(b)}\xi^{b-1}e^{-\xi}d\xi \]

\[ = \frac{1}{\Gamma(a)\Gamma(b)}\omega^{a-1}\int_0^\infty \xi^{a+b-1}e^{-(1+\omega)\xi}d\xi \]

\[ = \frac{1}{\Gamma(a)\Gamma(b)}\omega^{a+b} (\omega > 0). \]

\[ \square \]

Proof of Proposition 2

Proof. The proposition follows from Lemma IV.3 of Zhou and Carin (2015): Suppose \( y \) and \( (y_1, \cdots, y_K) \) are independent with \( y \sim \text{Ga}(\phi, \xi) \), and \( (y_1, \cdots, y_K) \sim \text{Dir}(\phi p_1, \cdots, \phi p_K) \), where \( \sum_{k=1}^K p_k = 1 \). Let \( x_k = yy_k \), then \( x_k \sim \text{Ga}(\phi p_k, \xi) \) independently for \( k = 1, \cdots K \).

\[ \square \]

Proof of Proposition 3

Proof. The marginal density of \( \beta \) for the R2-D2 prior is

\[ \pi_{\text{R2-D2}}(\beta) = \frac{\Gamma(a_\pi + b)}{\Gamma(a_\pi)\Gamma(b)} \int_0^\infty \frac{1}{2\sqrt{\lambda/2}} \exp\{-\frac{|\beta|}{\sqrt{\lambda/2}}\} \frac{\lambda^{a_\pi-1}}{(1 + \lambda)^{a_\pi+b}} d\lambda \]

\[ x = \frac{1}{\sqrt{\lambda/2}} \quad \frac{2a_\pi \Gamma(a_\pi + b)}{\Gamma(a_\pi)\Gamma(b)} \int_0^\infty \exp\{-|\beta|x\} \frac{x^{2b}}{(x^2 + 2)^{a_\pi+b}} dx. \quad (3.11) \]
Let \( \mu = |\beta|, \nu = b + 1/2, u^2 = 2, \) and \( \rho = 1 - a_\pi - b, (|\arg u| < \frac{\pi}{2}, \Re \mu > 0, \) and \( \Re \nu > 0) \), then according to 3.389.2 in Zwillinger (2014),

\[
\pi_{R2-D2}(\beta) = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \int_0^{\infty} \exp\{-\mu x\} x^{2\nu-1} \left(x^2 + u^2\right)^{\rho-1} dx
\]

\[
\pi_{R2-D2}(\beta) = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \frac{u^{2\nu+2\rho-2}}{2\sqrt{\pi} \Gamma(1-\rho)} G_{13}^{31}\left(\frac{\mu^2 u^2}{4} \left|\frac{1-\nu}{1-\rho-\nu,\frac{1}{2}}\right\right)
\]

\[
\pi_{R2-D2}(\beta) = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \frac{2^{1/2-a_\pi}}{2\sqrt{\pi} \Gamma(a_\pi + b)} G_{13}^{31}\left(\frac{\beta^2}{2} \left|\frac{1-b}{a_\pi - \frac{1}{2},0,\frac{1}{2}}\right\right)
\]

\[
\pi_{R2-D2}(\beta) = \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)} G_{13}^{31}\left(\frac{\beta^2}{2} \left|\frac{1-b}{a_\pi - \frac{1}{2},0,\frac{1}{2}}\right\right)
\]

where \( G(\cdot) \) denotes the Meijer G-Function. Also, according to 16.19.1 in DLMF (2015),

\[
G_{13}^{31}\left(\frac{\beta^2}{2} \left|\frac{1-b}{a_\pi - \frac{1}{2},0,\frac{1}{2}}\right\right) = G_{31}^{13}\left(2\frac{\beta^2}{\beta^2} \left|\frac{\frac{3}{2}-a_\pi-1,\frac{1}{2}}{\frac{1}{2}+b}\right\right)
\]

so Proposition 3 follows. \( \square \)

**Proof of Theorem 6**

**Proof.** For the proof of Theorem 6, we will use the following lemma found in Miller (2006).

**Lemma 7.** (Watson’s Lemma) Suppose \( F(s) = \int_0^\infty e^{-st} f(t) dt, f(t) = t^\alpha g(t) \) where \( g(t) \) has an infinite number of derivatives in the neighborhood of \( t = 0, \) with \( g(0) \neq 0, \) and \( \alpha > -1. \)

Suppose \( |f(t)| < Ke^{ct} \) for any \( t \in (0, \infty), \) where \( K \) and \( c \) are independent of \( t. \) Then, for \( s > 0 \) and \( s \to \infty, \)

\[
F(s) = \sum_{k=0}^{n} \frac{g^{(k)}(0)}{k!} \frac{\Gamma(\alpha + k + 1)}{s^{\alpha+k+1}} + O\left(\frac{1}{s^{\alpha+n+2}}\right).
\]
According to equation (3.11) in the proof of Proposition 3, we have
\[
\pi_{R^2-D^2}(\beta) = \frac{2^{a+1}\Gamma(a_\pi + b)}{\Gamma(a_\pi)\Gamma(b)} \int_0^\infty \exp\left\{-|\beta|x\right\} \frac{x^{2b}}{(x^2 + 2)^{a_\pi + b}} \, dx = \int_0^\infty e^{-|\beta|x} f(x) \, dx \equiv F(|\beta|),
\]
where \(f(t) = C^* \frac{t^{2b}}{(t^2 + 2)^{a_\pi + \pi}} \equiv t^{2b} g(t)\), \(C^* = \frac{2^{a+1}\Gamma(a_\pi + b)}{\Gamma(a_\pi)\Gamma(b)}\), and \(g(t) = C^* (t^2 + 2)^{a_\pi - b}\) with \(g(t)\) has an infinite number of derivatives in the neighborhood of \(t = 0\), with \(g(0) \neq 0\). So the marginal density of \(R^2-D^2\) prior is the Laplace transforms of \(f(\cdot)\). By Watson’s Lemma, since \(|f(t)| < Ke^{ct}\) for any \(t \in (0, \infty)\), where \(K\) and \(c\) are independent of \(t\), then as \(|\beta| \to \infty\),
\[
F(|\beta|) = \sum_{k=0}^n \frac{g^{(k)}(0)}{k!} \frac{\Gamma(2b + k + 1)}{|\beta|^{2b+k+1}} + O\left(\frac{1}{|\beta|^{2b+n+2}}\right),
\]
and setting \(n = 2\) gives
\[
F(|\beta|) = C^* \left(2^{-a_\pi - b} \frac{\Gamma(2b + 1)}{|\beta|^{2b+1}} + 0 \frac{\Gamma(2b + 2)}{|\beta|^{2b+2}} + (-a_\pi - b)2^{-a_\pi - b} \frac{\Gamma(2b + 3)}{|\beta|^{2b+3}}\right) + O\left(\frac{1}{|\beta|^{2b+4}}\right)
\]
\[
= C^* 2^{-a_\pi - b} \left( \frac{\Gamma(2b + 1)}{|\beta|^{2b+1}} - (a_\pi + b) \frac{\Gamma(2b + 3)}{|\beta|^{2b+3}}\right) + O\left(\frac{1}{|\beta|^{2b+4}}\right) = O\left(\frac{1}{|\beta|^{2b+1}}\right).
\]
Hence, when \(b < 1/2\), as \(|\beta| \to \infty\), we have
\[
\pi_{R^2-D^2}(\beta) = C^* 2^{-a_\pi - b} \left( \frac{\Gamma(2b + 1)}{|\beta|^{2b+1}} - (a_\pi + b) \frac{\Gamma(2b + 3)}{|\beta|^{2b+3}}\right) + O\left(\frac{1}{|\beta|^{2b+2}}\right) \to \infty.
\]

Proof of Theorem 7

Proof. According to 10.25.3 in DLMF (2015), when both \(\nu\) and \(z\) are real, if \(z \to \infty\), then \(K_{\nu}(z) \approx \sqrt{\pi/(2z)}e^{-z}\). Then as \(|\beta| \to \infty\), the marginal density of the DL prior given in
Bhattacharya et al. (2015) satisfies
\[
\pi_{DL}(\beta) = \frac{1}{2(1+a_D)^{2/2}} |\beta|^{(a_D-1)/2} K_{1-a_D}(\sqrt{2|\beta|})
\]
\[
\approx \frac{1}{2(1+a_D)^{2/2}} |\beta|^{(a_D-1)/2} \sqrt{\frac{\pi}{2}} |\beta| e^{-\sqrt{2|\beta|}}
\]
\[
= C_0 |\beta|^{a_D/2-3/4} e^{-\sqrt{2|\beta|}} = O\left(\frac{|\beta|^{a_D/2-3/4}}{e^{\sqrt{2|\beta|}}}\right),
\]
where \(C_0 = \sqrt{\frac{\pi}{2}} \frac{\sqrt{\pi}}{(2\sqrt{2})^{1/2}} \frac{1}{\Gamma(a_D)}\) is a constant value. Furthermore, as \(|\beta| \to \infty\),
\[
\frac{\pi_{DL}(\beta)}{1/\beta^2} = C_0 |\beta|^{a_D/2+5/4} e^{-\sqrt{2|\beta|}} \to 0.
\]

**Proof of Theorem 8**

Proof. For the proof of Theorem 8, we use the following lemma from Fields (1972). Some useful notations used in the below proof: Denote \(a_P = (a_1, \cdots, a_p)\), as a vector, similarly, \(b_Q = (b_1, \cdots, b_q)\), \(c_M = (c_1, \cdots, c_m)\), and so on. Let \(\Gamma_n(a_P - t) = \prod_{k=n+1}^{p} \Gamma(c_k - t)\), with \(\Gamma_n(a_P - t) = 1\) when \(n = p\), \(\Gamma(c_M - t) = \Gamma_0(c_M - t) = \prod_{k=1}^{m} \Gamma(c_k - t)\), \(\Gamma^*(a_i - a_N) = \prod_{k=1, k \neq i}^{n} \Gamma(a_i - a_k)\), and
\[
_{p}F_{q}\left(\frac{a_P}{b_Q}, w\right) = \sum_{k=0}^{\infty} \frac{\Gamma(a_P + k) \Gamma(b_Q)}{\Gamma(b_Q + k) \Gamma(a_P)} w^k = \sum_{k=0}^{\infty} \prod_{j=1}^{p} \Gamma(a_j + k) \prod_{j=1}^{q} \Gamma(b_j) \prod_{j=1}^{p} \Gamma(a_j) w^k.
\]

**Lemma 8.** (Theorem 1 in Fields (1972)) Given (i) \(0 \leq m \leq q, 0 \leq n \leq p\); (ii) \(a_i - b_k\) is not a positive integer for \(j = 1, \cdots, p\) and \(k = 1, \cdots, q\); (iii) \(a_i - a_k\) is not an integer for
\(i, k = 1, \ldots, p, \text{ and } i \neq k; \text{ and (iv) } q < p \text{ or } q = p \text{ and } |z| > 1, \text{ we have}

\[
G_{m,n}^{p,q}(z \mid a_1, \ldots, a_p, b_1, \ldots, b_q) = \sum_{i=1}^{n} \Gamma^*(a_i - a_N)\Gamma(1 + b_M - a_i) \Gamma_n(1 + a_P - a_i) \Gamma_m(a_i - b_Q) z^{-1+a_i} F_p \left( \begin{array}{c} 1, 1+b_Q-a_i \end{array} \left| 1+a_P-a_i \right| \frac{(-1)^{q-m-n}}{z} \right).
\]

Now to prove Theorem 8, we have from Proposition 3 that, the marginal density of the R2-D2 prior has

\[
\pi_{R2-D2}^{\beta_j} = \frac{1}{\sqrt{2\pi(\alpha_x)^r(b^r)}} G_{m,n}^{p,q}(z \mid a_1, \ldots, a_p, b_1, \ldots, b_q)
\]

with \(m = 1, n = 3, p = 3, q = 1, a_1 = \frac{3}{2} - a_x, a_2 = 1, a_3 = \frac{1}{2}, b_1 = \frac{1}{2} + b, \text{ and } z = \frac{2}{\beta_2}.\) Conditions (i)-(iv) in Lemma 8 are
satisfied for $|\beta|$ near 0, since $0 < a_\pi < \frac{1}{2}$. Hence

$$\pi_{R,2}(\beta) = \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)} \sum_{i=1}^{3} \frac{\Gamma^*(a_i - a_M) \Gamma(1 + b - a_i)}{\Gamma(1 + a_p - a_i) \Gamma(1 + a_p - a_i)} \left( \frac{2}{\beta^2} \right)^{a_i - 1} F_2 \left( \begin{array}{c} 1, 1 + b - a_i \\ 1 + a_p - a_i \end{array} \right) - \frac{\beta^2}{2}$$

$$= \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)} \sum_{i=1}^{3} \frac{\prod_{k=1;k \neq i}^{3} \Gamma(a_i - a_k) \Gamma(1 + b - a_i)}{\Gamma(1 + a_p - a_i) \prod_{k=1}^{3} \Gamma(1 + a_k - a_i) \Gamma(a_i - b_k)} \left( \frac{2}{\beta^2} \right)^{1 - a_i} F_2 \left( \begin{array}{c} 1, 1 + b - a_i \\ 1 + a_p - a_i \end{array} \right) - \frac{\beta^2}{2}$$

$$= \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)} \sum_{i=1}^{3} \left\{ \frac{\prod_{k=1;k \neq i}^{3} \Gamma(a_i - a_k) \Gamma(1 + b - a_i)}{1} \left( \frac{2}{\beta^2} \right)^{1 - a_i} \right\}$$

$$\sum_{k=0}^{\infty} \frac{\Gamma(1 + k) \Gamma(1 + b - a_i + k) \prod_{j=1}^{3} \Gamma(1 + a_j - a_i)}{\Gamma(1 + a_\pi + b) \Gamma(1 + a_\pi + b) \frac{(-\beta^2)^k}{k!}}$$

$$= \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)} \left\{ \Gamma(1 - a_\pi) \Gamma(1 - a_\pi) \frac{(2 \beta^2)^{1/2}}{\Gamma(2 a_\pi + 1)} \sum_{k=0}^{\infty} \frac{\Gamma(1 + k) \Gamma(1 + b + k) \Gamma(3 / 2 - a_\pi) \Gamma(1) \Gamma(1 / 2) \Gamma(1 / 2 + b)}{\Gamma(1 + k) \Gamma(1 + k) \Gamma(1 + k) \Gamma(1 + k) \Gamma(1 + b)} \frac{(-\beta^2)^k}{k!} \right\}$$

$$+ \Gamma(1 + a_\pi - 1) \Gamma(1 - a_\pi) \frac{(2 \beta^2)^{-1/2}}{\Gamma(2 a_\pi + 1)} \sum_{k=0}^{\infty} \frac{\Gamma(1 + k) \Gamma(1 + b + k) \Gamma(2 - a_\pi) \Gamma(3 / 2) \Gamma(1) \Gamma(1) \Gamma(1)}{\Gamma(2 - a_\pi + k) \Gamma(3 / 2 + k) \Gamma(1 + k) \Gamma(1 + b)} \frac{(-\beta^2)^k}{k!} \right\}$$

$$= \left\{ \frac{\Gamma(1 - a_\pi) \Gamma(1 - a_\pi) \Gamma(1 + b + k) \Gamma(1 + b + k) \Gamma(3 / 2 - a_\pi) \Gamma(4 / 2 - a_\pi)}{\Gamma(2 - a_\pi + k) \Gamma(3 / 2 + k) \Gamma(1 + k) \Gamma(1 + k)} \frac{(-\beta^2)^k}{k!} \right\}$$

$$= \frac{1}{\sqrt{2\pi} \Gamma(a_\pi) \Gamma(b)}.$$
Let

\[ U_1(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(a_\pi + b + k)}{\Gamma(\frac{1}{2} + a_\pi + k)\Gamma(a_\pi + k)} \frac{(-1)^k (\frac{\beta^2}{2})^{k+a_\pi - 1/2}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_1(k, \beta^2) \]

\[ U_2(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(\frac{3}{2} + b + k)}{\Gamma(\frac{3}{2} - a_\pi + k)\Gamma(\frac{1}{2} + k)} \frac{(-1)^k (\frac{\beta^2}{2})^{k}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_2(k, \beta^2) \]

\[ U_3(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(1 + b + k)}{\Gamma(2 - a_\pi + k)\Gamma(\frac{3}{2} + k)} \frac{(-1)^k (\frac{\beta^2}{2})^{k+1/2}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_3(k, \beta^2), \]

then

\[ \pi_{R^2-D^2}(\beta) = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_\pi)\Gamma(b)}{\Gamma(1/2)} \left\{ \Gamma(1/2 - a_\pi)\Gamma(1 - a_\pi)\Gamma(a_\pi)\Gamma(1/2 + a_\pi)U_1(\beta^2) + \Gamma(a_\pi - 1/2)\Gamma(1/2)\Gamma(3/2 - a_\pi)U_2(\beta^2) + \Gamma(a_\pi - 1)\Gamma(-1/2)\Gamma(3/2)\Gamma(2 - a_\pi)U_3(\beta^2) \right\} \equiv C_1^* U_1(\beta^2) + C_2^* U_2(\beta^2) + C_3^* U_3(\beta^2), \]

where

\[ C_1^* = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_\pi)\Gamma(b)}{\Gamma(1/2)} \Gamma(1/2 - a_\pi)\Gamma(1 - a_\pi)\Gamma(a_\pi)\Gamma(1/2 + a_\pi) > 0 \]

\[ C_2^* = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_\pi - 1/2)\Gamma(1/2)\Gamma(3/2 - a_\pi)}{\Gamma(1/2)} \Gamma(a_\pi - 1/2)\Gamma(1/2)\Gamma(3/2 - a_\pi) < 0 \]

\[ C_3^* = \frac{1}{\sqrt{2\pi}} \frac{\Gamma(a_\pi - 1)\Gamma(-1/2)\Gamma(3/2)\Gamma(2 - a_\pi)}{\Gamma(1/2)} \Gamma(a_\pi - 1)\Gamma(-1/2)\Gamma(3/2)\Gamma(2 - a_\pi) > 0. \]

For fixed \( \beta \) near the neighborhood of zero, \( u_1(k, \beta^2), u_2(k, \beta^2), \) and \( u_3(k, \beta^2) \) are all monotone decreasing, and converge to zero as \( k \to \infty \). Thus, by alternating series test, \( U_1(\beta^2), U_2(\beta^2), \)
and $U_3(\beta^2)$ all converge. Also, we have

$$C_0|\beta|^{2a_\pi-1} - C_1|\beta|^{2a_\pi+1} = u_1(0, \beta^2) - u_1(1, \beta^2) \leq U_1(\beta^2) \leq u_1(0, \beta^2) = C_0|\beta|^{2a_\pi-1},$$

$$C_2 - C_3|\beta|^2 = u_2(0, \beta^2) - u_2(1, \beta^2) \leq U_2(\beta^2) \leq u_2(0, \beta^2) = C_2,$$

$$C_4|\beta| - C_5|\beta|^3 = u_3(0, \beta^2) - u_3(1, \beta^2) \leq U_3(\beta^2) \leq u_3(0, \beta^2) = C_4|\beta|,$$

where $C_0$, $C_1$, $C_2$, $C_3$, and $C_4$ are all positive constants. So given that $|\beta|$ in the neighborhood of zero and $a_\pi \in (0, \frac{1}{2})$,

$$C_1^*(C_0|\beta|^{2a_\pi-1} - C_1|\beta|^{2a_\pi+1}) + C_2^*C_2 + C_3^*(C_4|\beta| - C_5|\beta|^3)$$

$$\leq \pi_{R2-D2}(\beta) \leq C_1^*C_0|\beta|^{2a_\pi-1} + C_2^*(C_2 - C_3|\beta|^2) + C_3^*C_4|\beta|,$$

then $\pi_{R2-D2}(\beta) = O(|\beta|^{2a_\pi-1}).$ \hfill \qed

**Proof of Theorem 9**

**Proof.** According to 10.30.2 in DLMF (2015), when $\nu > 0$, $z \to 0$ and $z$ is real, $K_{\nu}(z) \approx \frac{1}{2\Gamma(\nu)}(\frac{z}{2})^\nu$. So given $0 < a_D < 1$ and $|\beta| \to 0$,

$$\pi_{DL}(\beta) = \frac{|\beta|^{(a_D-1)/2}K_{1-a_D}(\sqrt{2|\beta|})}{2^{1+a_D}/2\Gamma(a_D)} \approx \frac{|\beta|^{(a_D-1)/2}1/2\Gamma(1-a_D)(\sqrt{2|\beta|})^{a_D-1}}{2^{1+a_D}/2\Gamma(a_D)} = C|\beta|^{a_D-1},$$

where $C = \frac{\Gamma(1-a_D)}{2^{1+a_D}\Gamma(a_D)}$ is a constant value. Theorem 9 follows then. \hfill \qed

**Proof of Theorem 10**

**Proof.** As shown in Theorem 8, when $|\beta|$ is close to zero and $0 < a_\pi < \frac{1}{2}$, $\pi_{R2-D2}(\beta) \approx
\( C_1 + \frac{C_2}{\beta^{1-2a\pi}} \), where \( C_1 \) and \( C_2 \) are some constants, so

\[
\int_0^{1/\sqrt{n}} \pi_{\text{R2-D2}}(\beta) \, d\beta \approx \int_0^{1/\sqrt{n}} (C_1 + \frac{C_2}{\beta^{1-2a\pi}}) \, d\beta = \frac{1}{\sqrt{n}} \frac{C_2}{2a\pi} \left( n^{1-a\pi} + C_1 \frac{2a\pi}{C_2} \right).
\]

By applying Lemma 6, we have

\[
R_n(\text{R2-D2}) \leq \frac{1}{n} - \frac{1}{n} \log \left( \frac{1}{\sqrt{n}} \frac{C_2}{2a\pi} \left( n^{1-a\pi} + C_1 \frac{2a\pi}{C_2} \right) \right) \leq \frac{1}{n} \left( 1 + \frac{1}{2} \log n - \log(n^{1-a\pi}) + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{n^{1/2-a\pi}} \right) \right),
\]

much smaller than the risk of the Horseshoe and Horseshoe+ prior, i.e., \( O \left( \frac{1}{n} \log \left( \frac{n}{(\log n)^{b_0}} \right) \right) \), where \( b_0 \) is some constant value (note: \( b_0 \) is different for Horseshoe and Horseshoe+ prior).

**Proof of Theorem 11**

*Proof.* As shown in the proof of theorem 9, when \( |\beta| \) is close to zero and \( 0 < a_D < 1 \),

\( \pi_{\text{DL}}(\beta) \approx C|\beta|^{a_D-1} \), where \( C = \frac{\Gamma(1-aD)}{2^{1+aD} \Gamma(aD)} \), so

\[
\int_0^{1/\sqrt{n}} \pi_{\text{DL}}(\beta) \, d\beta \approx C \int_0^{1/\sqrt{n}} |\beta|^{a_D-1} \, d\beta = \frac{C}{a_D} n^{-aD/2} = \frac{C}{a_D} \frac{1}{\sqrt{n}} n^{1-aD/2}.
\]

By applying Lemma 6, we have

\[
R_n(\text{DL}) \leq \frac{1}{n} - \frac{1}{n} \log \left( \frac{C}{a_D} \frac{1}{\sqrt{n}} n^{1-aD/2} \right) \leq \frac{1}{n} \left( 1 + \frac{1}{2} \log n - \log(n^{1-aD/2}) + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{n^{1/2-aD/2}} \right) \right),
\]

much smaller than the risk of the Horseshoe and Horseshoe+ prior, i.e., \( O \left( \frac{1}{n} \log \left( \frac{n}{(\log n)^{b_0}} \right) \right) \), where \( b_0 \) is some constant value (note: \( b_0 \) is different for Horseshoe and Horseshoe+ prior).
Proof of Theorem 12

Proof. Denote the estimated set of non-zero coefficients is $A_n = \{ j : \beta_{nj} \neq 0, j = 1, \ldots , p_n \}$. Also $\sigma^2$ is fixed at 1. Given the R2-D2 prior, the probability assigned to the region $(\beta_n : ||\beta_n - \beta_n^0|| < t_n)$ is

$$ P(\beta_n : ||\beta_n - \beta_n^0|| < t_n) = P\left( \beta_n : \sum_{j \in A_n} (\beta_{nj} - \beta_{nj}^0)^2 + \sum_{j \notin A_n} \beta_{nj}^2 < \frac{t_n^2}{p_n} \right) $$

$$ \geq P\left( \beta_{nj}^0 : \sum_{j \in A_n} (\beta_{nj} - \beta_{nj}^0)^2 < \frac{q_n t_n^2}{p_n} \right) \times P\left( \beta_{nj}^0 : \sum_{j \notin A_n} \beta_{nj}^2 < \frac{(p_n - q_n) t_n^2}{p_n} \right) $$

$$ \geq \prod_{j \in A_n} \left\{ P\left( \beta_{nj} : |\beta_{nj} - \beta_{nj}^0| < \frac{t_n}{\sqrt{p_n}} \right) \times P\left( \beta_{nj} : \beta_{nj}^2 < \frac{t_n^2}{p_n} \right) \right\} \times \left( 1 - \left\{ P\left( \beta_{nj}^0 : \beta_{nj}^2 \geq \frac{t_n^2}{p_n} \right) \right\}^{p_n - q_n} \right) $$

$$ \geq \prod_{j \in A_n} \left\{ P\left( - \sup_{j \in A_n} |\beta_{nj}^0| - \frac{t_n}{\sqrt{p_n}} < \beta_{nj} < \sup_{j \in A_n} |\beta_{nj}^0| + \frac{t_n}{\sqrt{p_n}} \right) \right\} \times \left( 1 - \left\{ P\left( \beta_{nj}^0 : |\beta_{nj}^0| \geq \frac{t_n}{b/2} \right) \right\}^{p_n - q_n} \right) $$

$$ \geq \prod_{j \in A_n} \left\{ P\left( - \sup_{j \in A_n} |\beta_{nj}^0| - \frac{t_n}{\sqrt{p_n}} < \beta_{nj} < \sup_{j \in A_n} |\beta_{nj}^0| + \frac{t_n}{\sqrt{p_n}} \right) \right\} \times \left\{ 1 - \left( \frac{b/2 E(|\beta_{nj}^0|)}{t_n^{b/2}} \right)^{p_n - q_n} \right\} $$

$$ \geq \left\{ 2 \frac{t_n}{\sqrt{p_n}} \pi \left( \sup_{j \in A_n} |\beta_{nj}^0| + \frac{t_n}{\sqrt{p_n}} \right) \right\}^{q_n} \times \left\{ 1 - \left( \frac{b/2 E(|\beta_{nj}^0|)}{t_n^{b/2}} \right)^{p_n - q_n} \right\} $$

where $\pi$ is the marginal density function of $\beta_j$, symmetric and decreasing when the support is positive, and the last but one “$\geq$” is directly got from Markov’s inequality.
Also, based on prior (3.6), for any \( b > 0 \), conditional expectations give

\[
E(\| \beta_j^b \|) = E\{E(E(\| \beta_j^b \| | \lambda_j) | \xi)\} = E_x[\| \lambda_j \| | \xi] = \frac{b \Gamma\left(\frac{b}{2}\right) \Gamma\left(a_\pi + \frac{b}{2}\right)}{2b^2 \Gamma(a_\pi)}.
\]

Now assume \( t_n = \frac{\Delta}{n^{\rho/2}} \) and assumptions (A1) - (A4) are satisfied. Then since \( \limsup_{j=1, \ldots, p_n} |\beta_{nj}^0| < \infty \), there exists a sequence \( k_n = o(n) \) such that \( \sup_{j=1, \ldots, p_n} |\beta_{nj}^0| < k_n \) and \( k_n \to \infty \) as \( n \to \infty \).

For the R2-D2 prior, based on equation (3.11), the marginal density is a decreasing function on the positive supports. Then together with the tail approximation of the marginal density as in the proof of Theorem 6, i.e., equation (3.12), we have

\[
\pi(\sup_{j \in \mathcal{A}_n} |\beta_{nj}^0| + \frac{t_n}{\sqrt{p_n}}) \geq \pi(k_n + \frac{t_n}{\sqrt{p_n}}) \geq \frac{\Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} 2^{-b} \frac{\Gamma(2b + 1)}{(k_n + \frac{\Delta}{n^{\rho/2} \sqrt{p_n}})^{2b+1}}.
\]

Considering the fact that \( \Gamma(a) = a^{-1} - \gamma_0 + O(a) \) for \( a \) close to zero with \( \gamma_0 \) the Euler-Mascheroni constant, now we have

\[
P(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) \geq \left\{ \frac{2 \Delta}{n^{\rho/2} \sqrt{p_n}} \frac{\Gamma(a_\pi + b) \Gamma(b)}{\Gamma(a_\pi) \Gamma(b)} 2^{-b} \frac{\Gamma(2b + 1)}{(k_n + \frac{\Delta}{n^{\rho/2} \sqrt{p_n}})^{2b+1}} \right\} \times \left\{ 1 - \left( \frac{p_n^{b/2} n^{\rho b/2} \Gamma\left(\frac{b}{2}\right) \Gamma(a_\pi + \frac{b}{2})}{\Delta b^{2b/2}} \right)^{p_n-q_n} \right\}
\]

\[
\geq \left\{ \frac{2 \Delta}{n^{\rho/2} \sqrt{p_n}} \frac{\Gamma(a_\pi + b) \alpha_\pi 2^{-b} \Gamma(2b + 1)}{(k_n + \frac{\Delta}{n^{\rho/2} \sqrt{p_n}})^{2b+1}} \right\} \times \left\{ 1 - \left( \frac{p_n^{b/2} n^{\rho b/2} \Gamma\left(\frac{b}{2}\right) \Gamma(a_\pi + \frac{b}{2}) a_\pi}{\Delta b^{2b/2}} \right)^{p_n-q_n} \right\}.
\]

85
Taking the negative logarithm of both sides, and letting \( a = C/(p_n^{b/2} n^{b/2} \log n) \), we have

\[
- \log P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) \\
\leq -q_n \log \left\{ \frac{2\Delta CT(a + b)2^{-b}\Gamma(2b + 1)}{\sqrt{p_n}n^{b/2}n^{b/2}\log n \Gamma(b)} \right\} + q_n (2b + 1) \log(k_n + \frac{\Delta}{n^{\rho/2} \sqrt{p_n}}) \\
- q_n \log(1 - \left( \frac{p_n^{b/2} n^{b/2} b \Gamma(\frac{b}{2}) \Gamma(a + \frac{b}{2}) C}{\Delta^b 2b/2 n^b/2 \log n} \right)^{p_n - q_n}) \\
= -q_n \log \left\{ \frac{2\Delta CT(a + b)2^{-b}\Gamma(2b + 1)}{\Gamma(b)} \right\} - q_n \log(1 - \left( \frac{b \Gamma(\frac{b}{2}) \Gamma(a + \frac{b}{2}) C}{\Delta^b 2b/2 \log n} \right)^{p_n - q_n}) \\
+ q_n (2b + 1) \log(k_n + \frac{\Delta}{n^{\rho/2} \sqrt{p_n}}) + q_n \log \log n + \frac{b + 1}{2} q_n \log p_n + \frac{b + 1}{2} q_n \rho \log n
\]

The dominating term is the last one, and if \( q_n = o(n/\log n) \), \(- \log P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) < dn \) for all \( d > 0 \), so \( P(\beta_n : ||\beta_n - \beta^0_n|| < \frac{\Delta}{n^{\rho/2}}) > \exp(-dn) \). The posterior consistency is completed by applying Theorem 1 in Armagan et al. (2013b). □
Chapter 4

A Regression-based Method for Matrix Completion with Informative Missingness:
Application to Movie Rating Data

4.1 Introduction

Statistical methods to develop recommender systems have become a rapidly-advancing research area. We consider the MovieLens (http://grouplens.org/datasets/movielens/) rating system as a motivating example. This data set consists of a matrix of rating entries: each row represents a user, each column a movie, and each entry a rating associated with the particular user and movie. Because most users rate only a small subset of the movies, most entries are missing. The goal is to predict these missing entries in the partially revealed matrix based on a history of user’s preferences and movie’s features so that personalized recommendations can be made.

In the literature, there are three main approaches in recommendation systems: content-
based filtering, collaborative filtering, and hybrid recommender systems. Content-based filtering uses the profile of the user’s past preferences and item features to make predictions. For more details, one can refer to Pazzani and Billsus (2007), Adomavicius and Tuzhilin (2005), Cheng et al. (2010), Naveed et al. (2011), etc.

Instead of using the similarity of items, collaborative filtering makes recommendations based on the similarity of user with other collected users. The famous Netflix Prize (Bennett and Lanning, 2007) is a popular example. According to Breese et al. (1998), there are two categories of collaborative filtering algorithms. The first is memory-based methods, which makes prediction for the queried user and item using the entire rest of the rating matrix. Related literature includes Nakamura and Abe (1998), Sarwar et al. (2001), Debnath et al. (2008), etc. The other category is the model-based methods, which fit a parameterized model to the entire rating matrix, and then make predictions based on the learned model. Related methods can be found in Pavlov and Pennock (2002), Xue et al. (2005), Miyahara and Pazzani (2000), Vucetic and Obradovic (2005), Agarwal and Chen (2009), etc. Recently developed methods involve matrix factorization by learning a low-rank model, such as singular value decomposition (Billsus and Pazzani, 1998), non-negative matrix factorization (Lee and Seung, 1999), convex optimization (Candès and Recht, 2009), Bayesian Probabilistic Matrix Factorization (Salakhutdinov and Mnih, 2008), nonparametric matrix factorization (Yu et al., 2009), etc. For a more detailed review of classic and state-of-art algorithms, one may refer to Lee et al. (2012).

The two methods above perform unsatisfactorily without sufficient participation of users (Adomavicius and Tuzhilin, 2005). The hybrid approach avoids this shortcoming by combining the two methods, using both the rating matrix and user and item features for prediction. Related methods include Fab’s system (Balabanović and Shoham, 1997), Pennock et al. (2000), Popescul et al. (2001), Melville et al. (2002), Agarwal and Chen (2009), etc.

Most of these methods ignore the pattern of missing data. However, we suspect that missing
data patterns are often not ignorable. For example, we analyze MovieLens-100K data, which consists of 100,000 ratings from $I = 943$ users and $J = 1682$ movies. Movies are rated from 1 to 5 stars. All users in this dataset rated at least 20 movies and users rated an average of 6% of the movies. Figure 4.1 plots the number of ratings and mean rating by user (left) and movie (right). This plot clearly demonstrates that movies with higher average ratings are rated more often; ignoring this pattern leads to questionable inferences. We hypothesize that a missing rating may in fact be informative because users are more likely to rate movies that match their preferences.

In this paper, we propose a Bayesian hybrid recommender system that accounts for informative missingness. We used a shared-variable approach for matrix completion that jointly models the rating and the probability of each movie being rated. A regression-based method is used to model the association between the expected rating with user, item and user-item
specific features. Our model is fitted through MCMC sampling. The computation is stable, accurate and easy to implement. The application of MovieLens data illustrates that by adding informative missingness into the model, the prediction is significantly improved.

The rest of this chapter is outlined as follows. Section 4.2 introduces our proposed regression-based informative missingness model. Section 4.3 discusses the details of fitting the model. Section 4.4 applies the proposed model on the MovieLens-100K data together with a comparison with other models. Finally, Section 4.5 makes a summary, and points out some future work and research directions. This chapter is based on joint work with Brian Reich, Howard Bondell, Curtis Storlie, and Hua Zhou.

4.2 Model Details

4.2.1 Notations

Let $i = 1, 2, \cdots, I$ denote the users, and $j = 1, 2, \cdots, J$ denote the items (or movies), where $I$ and $J$ are the total number of users and items respectively. Let $u_i$ denote the $p \times 1$ feature vector for user $i$, $v_j$ denote the $q \times 1$ feature vector for item $j$, and $x_{ij}$ denote the $r \times 1$ user-item specific feature vector for dyad $(i, j)$. In the $I \times J$ rating matrix, if user $i$ rated item $j$, then the entry is $Y_{ij}^* \in \{1, 2, 3, 4, 5\}$, if there is no information about the rating, then it is missing. We denote $N(\mu, \sigma^2)$ as the univariate normal distribution, $MVN(\mu, \Sigma)$ as the multivariate normal distribution, and $IG(\alpha, \beta)$ an Inverse Gamma distribution with shape $\alpha$ and scale $\beta$.

4.2.2 Hybrid Model

The goal is to study the relationship between $Y_{ij}^*$ and $(x_{ij}, u_i, v_j)$ based on the observed data, so that predictions can be made for the unobserved entries. First, we use a latent variable approach
to account for the discrete nature of the ratings. Define $Y_{ij}$ as a continuous latent variable with

$$Y_{ij}^* = \underset{k \in \{1, 2, 3, 4, 5\}}{\text{argmin}} |Y_{ij} - k|.$$  \hfill (4.1)

Our model is $Y_{ij} \overset{iid}{\sim} \mathcal{N}(\mu_{ij}, \sigma^2)$ where $\mu_{ij}$ is the mean of the continuous latent rating. Motivated by the regression-based latent factor model (RLFM) in Agarwal and Chen (2009), which incorporates user/item features and their interaction into the prediction process through a latent factor model. We model the mean as

$$\mu_{ij} = \mathbf{x}_i' \mathbf{f}_0 + \mathbf{u}_i' \mathbf{g}_0 + \mathbf{v}_j' \mathbf{h}_0 + \alpha_i + \beta_j + \gamma_{ij},$$  \hfill (4.2)

where $\mathbf{f}_0, \mathbf{g}_0, \mathbf{h}_0$ are $r, p, q$-vector of regression coefficient respectively, $\alpha_i, \beta_j$ and $\gamma_{ij}$ are random effects for the user, item, and their interaction, respectively. Furthermore, we assume the interaction effects are of low rank, $L$, i.e.,

$$\gamma_{ij} = \mathbf{b}_i' \mathbf{d}_j = \sum_{l=1}^{L} b_{il} d_{lj},$$

where $\mathbf{b}_i = (b_{i1}, \cdots, b_{iL})'$, and $\mathbf{d}_j = (d_{1j}, \cdots, d_{Lj})'$. Assume the random effects $\alpha_i \overset{iid}{\sim} \mathcal{N}(\mu_\alpha, \sigma^2_\alpha), \beta_j \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\beta)$, $\mathbf{b}_i \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \mathbf{A}_b)$, and $\mathbf{d}_j \overset{iid}{\sim} \text{MVN}(\mathbf{0}, \mathbf{A}_d)$. For identification purpose, we fix $\mathbf{A}_d = \mathbf{I}_L$. Furthermore, we assume $\mathbf{A}_b$ is of diagonal structure. If this is not true, by rotation of $\mathbf{b}_i$, we can always make the covariance diagonal. For example, if $\mathbf{A}_b = \mathbf{\Gamma \Lambda \Gamma}'$ with $\Lambda$ diagonal with elements of eigenvalues of $\mathbf{A}_b$, and each column of $\Gamma$ is the relevant eigenvectors with $\mathbf{\Gamma \Gamma}' = \mathbf{\Gamma}' \mathbf{\Gamma} = \mathbf{I}$. Then by making $\mathbf{\Gamma}' \mathbf{b}_i \sim \text{MVN}(\mathbf{0}, \Lambda)$. Model (4.2) is a hybrid recommender system as it considers both the user and item features as well as trying to learn a parameterized model based on the rating matrix.
4.2.3 Hybrid+Info Model

Most of the rating \( Y_{ij}^* \)'s are missing for the matrix completion problem. Define \( \delta_{ij} = 1 \) if user \( i \) rated movie \( j \) and \( \delta_{ij} = 0 \) if \( Y_{ij}^* \) is missing. We propose a shared-variable approach to accommodate informative missingness. In this approach, the probability of each movie being rated depends on the expected rating. Denote \( \text{Prob}(\delta_{ij} = 1) = \Phi(\tilde{\mu}_{ij}) \), where \( \Phi \) is the standard normal distribution function, and

\[
\tilde{\mu}_{ij} = x'_{ij}\tilde{f}_0 + u'_i\tilde{g}_0 + v'_j\tilde{h}_0 + \tilde{\alpha}_i + \tilde{\beta}_j + \tilde{b}_i'd_j + c_0b_i'd_j.
\]  

(4.3)

Here \( \tilde{\alpha}_i \) and \( \tilde{\beta}_j \) are the random effects for user and movie, and \( \tilde{b}_i'd_j \) is the low-rank interaction effect similar to (4.2). As in the model (4.2), the random effects are modeled as \( \tilde{\alpha}_i \sim iid N(\tilde{\mu}_0 + c_1\alpha_i, \tilde{\sigma}^2_\alpha) \), \( \tilde{\beta}_j \sim iid N(c_2\beta_j, \tilde{\sigma}^2_\beta) \), \( \tilde{b}_i \sim iid \text{MVN}(0, \text{diag}\{\tilde{\sigma}^2_{b_i}\}) \), and \( \tilde{d}_j \sim iid \text{MVN}(0, I_L) \), so that \( \tilde{\mu}_0 \) controls the overall probability of rating movies. The coefficients \( c_0, c_1, \) and \( c_2 \), determine the degree and nature of informative missingness. If these coefficients are all zero, then the probability of rating a movie is independent of user preference, and missing values can be ignored. Positive \( c_1 \) implies that users that generally give favorable ratings are more likely to watch and rate movies; positive \( c_2 \) implies that well-liked movies are rated more often than others; and positive \( c_0 \) implies that user-specific movie preferences affect the movies they choose to rate.

To fit the informative missingness model, (4.2) and (4.3) are jointly estimated through Markov chain Monte Carlo (MCMC) (Section 4.3). In the remainder of the paper, we label model (4.2) based only on the observed data “Hybrid”, and the joint model (4.2) and (4.3) with informative missingness is labeled “Hybrid+Info”.

92
### 4.2.4 Special Cases

Several popular models are special cases of our model. Note that although Hybrid model (4.2) is motivated by the RLFM model of Agarwal and Chen (2009), there are several differences here. RLFM consists of a two-stage hierarchical latent factor model, with the first stage modeling the association between the response and both observed features and latent variables, while the second stage specifies the model of the relationship between latent variables with user and item features together with their random effects. Unlike RLFM, we combine the two stages into one single model. Another major difference is that we have an overall mean parameter in our model, i.e., $\mu_0$, while the RLFM assumes the overall mean is zero, which shrinks the profiles close to zero and results in biased estimator. Except for these differences, model (4.2) can be seemed as a special case of RLFM by assuming the low-rank decomposition components $b_i$ and $d_j$ following a zero mean normal priors but not associated with the user and item features.

Without considering the informative missingness, assuming $f_0, g_0, \text{and } h_0$ all equal to zero in Hybrid (4.2) leads to the random-effect-only (RE) model. This is similar as the “ZeroMean” model referred in Agarwal and Chen (2009), except that ZeroMean is based on a zero-mean normal prior on the factors, while the RE model adds an overall mean parameter $\mu_0$ into the model. Similarly, adding an intercept term into the ZeroMean model leads to less biased estimator for the RE model. The pure feature-based model, or “Feature” model can be obtained by assuming all the variance terms ($\sigma^2_\alpha$, $\sigma^2_\beta$, $A_b$, and $A_d$) equal to zero for the random effects in Hybrid model (4.2). When $c_0 = c_1 = c_2 = 0$, it means that the informative missingness model (4.3) will not be considered. By adding informative missingness to the RE model, we get the random-effect-only model but with informative missingness considered. These models are summarized in Table 4.1.
Table 4.1: Special cases of informative missingness joint model in (4.2) and (4.3).

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters set to zero</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature</td>
<td>$\sigma^2, \sigma^2_\alpha, \sigma^2_\beta, A^b, A^d, c_0, c_1, c_2$</td>
</tr>
<tr>
<td>RE</td>
<td>$f_0, g_0, h_0, c_0, c_1, c_2$</td>
</tr>
<tr>
<td>Hybrid</td>
<td>$c_0, c_1, c_2$</td>
</tr>
<tr>
<td>RE+Info</td>
<td>$f_0, g_0, h_0, \tilde{f}_0, \tilde{g}_0, \tilde{h}_0$</td>
</tr>
<tr>
<td>Hybrid+Info</td>
<td>-</td>
</tr>
</tbody>
</table>

4.3 Model Fitting

4.3.1 Prior Specification

The unknown parameters to estimate include $f_0, g_0, h_0, \tilde{f}_0, \tilde{g}_0, \tilde{h}_0, \mu_0, \tilde{\mu}_0, c_1, c_2, c_0, \sigma^2, \sigma^2_\alpha, \sigma^2_\beta, A^b, A^d, c_0, c_1, c_2$ and $\sigma^2_{\beta l}, \tilde{\sigma}^2_{\beta l}, \tilde{\sigma}^2_{\beta l}$, and $\tilde{\sigma}^2_{\beta l}$. The latent factors are $\alpha_i, \beta_j, b_i, d_j, \tilde{b}_i$ and $\tilde{d}_j$. In this paper, a full Bayesian approach is used to fit this model. We specify the following prior distributions. $\mu_0, \tilde{\mu}_0, c_1, c_2, c_0$ are given $N(0, s^2)$ prior; $\sigma^2_\alpha, \sigma^2_\beta, \tilde{\sigma}^2_\alpha, \tilde{\sigma}^2_\beta, \tilde{\sigma}^2_{\beta l}$, and $\sigma^2_{\beta l}$ are given $IG(e, e)$ prior. To specify prior distributions on the $L$ variance parameters $\sigma^2_{bl}$ ($l = 1, \cdots, L$), one has to take the label switching problem into consideration. For example, switching the value of $\sigma^2_{b_1}$ and $\sigma^2_{b_2}$ while fixing $\sigma^2_{bl}$ ($l = 3, \cdots, L$), leads to a different estimator. For identification purposes, we artificially make $\sigma^2_{bl}$ monotone by assuming $\sigma^2_{bl} = \sigma^2_{b l} \phi^{l-1}$, and similarly $\tilde{\sigma}^2_{bl} = \tilde{\sigma}^2_{b l} \tilde{\phi}^{l-1}$, where $\sigma^2_{bl}$ and $\tilde{\sigma}^2_{bl}$ are also given $IG(e, e)$ prior, and the hyperparameters $\phi$ and $\tilde{\phi}$ have Uniform $(0, 1)$ priors and are updated using Metropolis-Hastings algorithms. The regression coefficients are given the normal prior, $f_0 \sim MVN(0, \sigma^2_f I_r), g_0 \sim MVN(0, \sigma^2_g I_p), h_0 \sim MVN(0, \sigma^2_h I_q), \tilde{f}_0 \sim MVN(0, \tilde{\sigma}^2_f I_r), \tilde{g}_0 \sim MVN(0, \tilde{\sigma}^2_g I_p), \tilde{h}_0 \sim MVN(0, \tilde{\sigma}^2_h I_q)$, with hyperparameters $\sigma^2_f, \sigma^2_g$, $\sigma^2_h, \tilde{\sigma}^2_f, \tilde{\sigma}^2_g$, and $\tilde{\sigma}^2_h$ all given $IG(e, e)$ prior.
4.3.2 Matrix Representation

To estimate some parameters jointly, we rewrite our model in matrix notation. The expected value of the $I \times J$ matrix $Y = \{Y_{ij}\}$ is $\mu = \{\mu_{ij}\}$, where

$$\mu = \{X_{f0}\} + Ug_0 + Vh_0 + \alpha \otimes 1_J^T + 1_I \otimes \beta^T + BD,$$

with $X = (x_{11}, \ldots, x_{1J}, x_{21}, \ldots, x_{IJ})'$ is $IJ \times r$ user-item specific feature matrix, $U = (u_1, \ldots, u_I)'$ is $I \times p$ user feature matrix, $V = (v_1, \ldots, v_J)'$ ia $J \times q$ item feature matrix, $\{X_{f0}\}$ is $I \times J$ matrix by using the inverse of vectorization operation, $\alpha = (\alpha_1, \ldots, \alpha_I)^T$, $\beta = (\beta_1, \ldots, \beta_J)^T$, $B = \{b_{ij}\} = (b'_1, \ldots, b'_J)^T$ is $I \times L$ matrix, $D = \{d_{jl}\} = (d_1, \ldots, d_J)$ is $L \times J$ matrix, $1_n = (1, \ldots, 1)^T$ is vector of $n$ ones, and $\otimes$ denotes the Kronecker product operation on two matrices. The matrix representation of (4.3) is

$$\tilde{\mu} = \{X_{\tilde{f}0}\} + U\tilde{g}_0 + V\tilde{h}_0 + \tilde{\alpha} \otimes 1_J^T + 1_I \otimes \tilde{\beta}^T + \tilde{B}D + c_0BD.$$ 

4.3.3 MCMC Details

Discretization of the response and missing data are handled using a latent variable approach. The latent discretization variable $Y_{ij}$ is defined in (4.1). Similarly, for missing observations, we introduce latent variables $\tilde{Y}_{ij} \sim N(\tilde{\mu}_{ij}, 1)$ and relate the latent variables to the missing data indicators as $\delta_{ij} = I(\tilde{Y}_{ij} > 0)$. Both latent variables $Y_{ij}$ and $\tilde{Y}_{ij}$ are updated from their truncated normal full conditional distributions. The latent variable $Y_{ij}$ is restricted to $A_{ij} = (-\infty, \infty)$ if $\delta_{ij} = 0$; if $\delta_{ij} = 1$ then $Y_{ij}$ is restricted to $A_{ij} = (-\infty, 1.5)$ if $Y_{ij}^* = 1$, $A_{ij} = (Y_{ij}^* - 0.5, Y_{ij}^* + 0.5)$ if $Y_{ij}^* \in \{2, 3, 4\}$, and $A_{ij} = (4.5, \infty)$ if $Y_{ij}^* = 5$. The latent variable $\tilde{Y}_{ij}$ is restricted to
\[ \tilde{A}_{ij} = (-\infty, 0) \text{ if } \delta_{ij} = 0 \text{ and } \tilde{A}_{ij} = (0, \infty) \text{ if } \delta_{ij} = 1. \] The full conditionals are then

\[ Y_{ij}|\cdot \sim \text{TN}_{A_{ij}}(\mu_{ij}, \sigma^2) \quad \text{and} \quad \tilde{Y}_{ij}|\cdot \sim \text{TN}_{\tilde{A}_{ij}}(\tilde{\mu}_{ij}, 1), \quad (4.4) \]

where \( \text{TN}_A(m, s^2) \) is the truncated normal distribution with support \( A \).

MCMC begins by specifying initial values for all parameters and then updating each parameter in sequence. Hyperparameter \( s \) is fixed at 100, and \( e \) is fixed at 0.01. Metropolis-Hastings algorithms are used to update hyperparameter \( \phi \), with a Beta\( (c\tilde{\phi}^*, c(1 - \tilde{\phi}^*)) \) candidate distribution centered on the current value \( \tilde{\phi}^* \) and with \( c \) tuned to give acceptance probability near 0.3. \( \tilde{\phi} \) is updated in the same way. All other parameters are updated through Gibbs samplers, with details given in Section 4.6. We generate 50,000 samples and discard the first 20,000 as burn-in. Convergence is monitored using autocorrelation plots and visual inspection for several representative parameters.

### 4.3.4 Prediction

After fitting the joint model using both observed data and missingness data indicators, we obtain the posterior distribution of the parameters. Then to predict the response for a particular user \( i \) and item \( j \), we draw samples for \( Y_{ij}^* \) from the posterior predictive distribution. At MCMC iteration \( s \), a sample of the parameters, i.e., \( \hat{x}_{ij}, \hat{g}_0, \hat{h}_0, \hat{\alpha}_i, \hat{\beta}_j, \hat{b}_i, \text{ and } \hat{d}_j \) are collected, then we simultaneously output the prediction point estimator,

\[ \hat{Y}_{ij}^{(s)} = \hat{x}_{ij} f_0 + u_i g_0 + v_j h_0 + \hat{\alpha}_i + \hat{\beta}_j + \hat{b}_i d_j, \]

and then \( \hat{Y}_{ij}^*(s) \) is transformed as \( \hat{Y}_{ij}^* \) by (4.1). The posterior probability that \( Y_{ij}^* = y \) for \( y \in \{1, 2, 3, 4, 5\} \) is thus calculated based on the proportion of \( \hat{Y}_{ij}^* = y \) in the total \( S \) MCMC
iterations. The final prediction is made by computing the posterior mean for the data, i.e.,

\[
\text{Prob}(Y_{ij}^* = y) = \frac{1}{S} \sum_{s=1}^{S} I(\hat{Y}_{ij}^{*(s)} = y), \quad \text{and then} \quad \hat{Y}_{ij} = \sum_{y=1}^{5} y \text{Prob}(Y_{ij}^* = y).
\]

(4.5)

4.4 Analysis of MovieLens Data

The MovieLens data (http://grouplens.org/datasets/movielens/), collected as part of the GroupLens Research Project at the University of Minnesota (Herlocker et al., 1999), is analyzed in this section. We conduct experiments on the MovieLens-100K data. The user features are age, gender, occupation, and ZIP code (similar to Agarwal and Chen (2009), we only use the first digit). The occupation variable contains 21 categories, administrator, artist, doctor, educator, engineer, entertainment, executive, healthcare, homemaker, lawyer, librarian, marketing, programmer, retired, salesman, scientist, student, technician, writer, other, and none. We use the “student” category, which has the highest number of ratings, as the baseline, and include dummy variables for all the remaining occupations. There are 11 categories for ZIP code (only the first digit is considered), including 0 to 9 in total 10 integer numbers, as well as an “other” type consisting of letters (letters suggest Canadian addresses). By using ZIP code “9” as the baseline, we create 10 dummy variable for the ZIP code. Finally there are \(p = 1 + 1 + 20 + 10 = 32\) user features. The single movie feature is genre, which contains 19 categories: unknown, Action, Adventure, Animation, Children’s, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, and Western. Movies can be in several genres at once. So all 19 categories are transformed as dummy variables, with 1 indicating the movie is of that genre, and 0 otherwise, resulting in a \(q = 19\) item feature vector. In this section, we estimate the reduction in prediction error attributable to our handling of informative missingness by five training-testing MovieLens data
sets.

Five pre-specified splits of training-testing data are available for MovieLens-100K. We fit the model on each training set, and then get the prediction error on the testing data. For a model with no informative missingness considered, we use (4.5) as the prediction. For the informative missingness model, we do not know whether the rating is missing or not in the testing data. Using the posterior mean (4.5) results in biased prediction, as conditional on the movie being rated leads to a higher rating. In this case, the prediction is based on the mean estimator of the ratings conditionally on the posterior samples being rated (i.e., for those samples with \( \delta_{ij} = 1 \)), i.e.,

\[
\hat{Y}_{ij} = \sum_{y=1}^{5} y \text{Prob}(Y^*_ij = y | \tilde{Y}_{ij}^{(s)} > 0).
\]

This will give a higher predictive rating than the no informative missingness model.

We use the root mean squared error (RMSE) to measure predictive performance. Table 4.2 reports the average RMSEs over the five training-testing splits for each model. Note that even as small an improvement as 0.01 RMSE leads to a significant difference in the ranking system. By adding informative missingness into the observed data only model, the prediction is significantly improved. However, incorporating the user and item features does not improve the prediction of the random effect model.

<table>
<thead>
<tr>
<th>RMSE</th>
<th>Feature RE</th>
<th>Hybrid</th>
<th>RE+Info</th>
<th>Hybrid+Info</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.235</td>
<td>0.818</td>
<td>0.817</td>
<td>0.792</td>
</tr>
</tbody>
</table>

Agarwal and Chen (2009) noted that since there are no new users and movies in the test set,
the improvement of their RLFM compared to the random effect only model is not significant. By re-splitting the training/testing data based on time, Agarwal and Chen (2009) obtain significant boost with the inclusion of features. We did not try the time-based splits as it is not of our main focus here. We note that many of the features have statistically significant effect, both for predicting the mean rating and probability of rating the movie according to the posterior estimation of $g_0$, $h_0$, $\tilde{g}_0$ and $\tilde{h}_0$.

4.5 Discussion and Future Work

In this paper, we propose a new hybrid recommender system for matrix completion for movie ratings that accounts for informative missingness. We use a shared-variable regression-based approach which jointly models the response and probability of missingness as a function of the mean rating. We applied the proposed method to the MovieLens data and found prediction errors that are highly competitive with other approaches (Table 4.2). Therefore, it seems informative missingness is an issue that should be considered in future analyses of movie ratings.

A limitation of the proposed approach is that it is entirely model-based and makes several parametric assumptions such as linearity and normality. Our analysis suggests that these assumptions lead to good results for the MovieLens data, but a more flexible model may be required in other applications. For spatial data, Reich et al. (2013) provide a non-parametric Bayesian model for informatively missing data that could be adapted for matrix completion.

There are many exciting possibilities for future work. First, computation and MCMC steps proposed in this chapter requires modification. Our method takes approximately 130 minutes for 1,000 MCMC iterations for the MovieLens data using R and an ordinary PC. Many of the steps given in Appendix are embarrassingly parallel. Also, a sensitivity analysis to the hyperparameters in the model such as $L$ needs to be checked.
4.6 Appendix: Computational Details

Denote $A_i$ denote the $i$th row for the matrix $A$, and $A_{(j)}$ denote $j$th column of $A$. The MCMC details of the Hybrid+Info model (4.2) and (4.3) are as follows.

1. Sample random effects.

   - Sample $\alpha_i$ and $\beta_j$. Let $R_1 = \{R_{1ij}\} = Y - \mu + \alpha \otimes 1^T_j$, $R_2 = \{R_{2ij}\} = Y - \mu + 1_I \otimes \beta^T$.

     \[
     \alpha_i \mid \cdot \sim N\left( \frac{\sigma^{-2} \sum_j R_{1ij} + \mu_0 \sigma^{-2}_\alpha + c_1 \sigma^2_\alpha (\tilde{\alpha}_i - \tilde{\mu}_0)}{J \sigma^{-2} + \sigma^{-2}_\alpha + c_1^2 \sigma^2_\alpha} , \frac{1}{J \sigma^{-2} + \sigma^{-2}_\alpha + c_1^2 \sigma^2_\alpha} \right)
     \]

     \[
     \beta_j \mid \cdot \sim N\left( \frac{\sigma^{-2} \sum_i R_{2ij} + c_2 \sigma^2_\beta (\tilde{\beta}_j)}{I \sigma^{-2} + \sigma^{-2}_\beta + c_2^2 \sigma^2_\beta} , \frac{1}{I \sigma^{-2} + \sigma^{-2}_\beta + c_2^2 \sigma^2_\beta} \right)
     \]

   - Sample $\tilde{\alpha}_i$ and $\tilde{\beta}_j$. Let $\tilde{R}_1 = \{\tilde{R}_{1ij}\} = \tilde{Y} - \tilde{\mu} + \tilde{\alpha} \otimes 1^T_j$, and $\tilde{R}_2 = \{\tilde{R}_{2ij}\} = \tilde{Y} - \tilde{\mu} + 1_I \otimes \tilde{\beta}^T$.

     \[
     \tilde{\alpha}_i \mid \cdot \sim N\left( \frac{\sum_j \tilde{R}_{1ij} + \tilde{\sigma}^{-2}_\alpha (\tilde{\mu}_0 + c_1 \alpha_i)}{J + \tilde{\sigma}^{-2}_\alpha} , \frac{1}{J + \tilde{\sigma}^{-2}_\alpha} \right)
     \]

     \[
     \tilde{\beta}_j \mid \cdot \sim N\left( \frac{\sum_i \tilde{R}_{2ij} + \tilde{\sigma}^{-2}_\beta (c_2 \beta_j)}{I + \tilde{\sigma}^{-2}_\beta} , \frac{1}{I + \tilde{\sigma}^{-2}_\beta} \right)
     \]

   - Sample $b_i, d_j$. Let $M = Y - \mu + BD$, $\tilde{M} = \tilde{Y} - \tilde{\mu} + c_0 BD$, and $\Sigma_{bl} = \text{diag}\{\sigma^2_{bl}\}$.

     \[
     b_i \mid \cdot \sim \text{MVN} \left( \Sigma_b \left[ \sigma^{-2} DM^T + c_0 D\tilde{M}^T \right]_{(i)} , \Sigma_b \right)
     \]
where $\Sigma_b^{-1} = \sigma^{-2}DD^T + \Sigma_{bl}^{-1} + c_0^2DD^T$.

$$d_j \mid \sim MVN \left( \Sigma_d \left[ \sigma^{-2}B^Tm + c_0B^T\tilde{M}_{(j)} \right], \Sigma_d \right)$$

where $\Sigma_d^{-1} = \sigma^{-2}B^TB + I_L + c_0^2B^TB$.

- Sample $\tilde{b}_i, \tilde{d}_j$. Let $\tilde{N} = \{\tilde{N}_{ij}\} = \tilde{Y} - \tilde{\mu} + \tilde{BD}, \tilde{\Sigma}_{bl} = \text{diag}\{\tilde{\sigma}_b^2\}$.

  $$\tilde{b}_i \mid \sim MVN \left( \left( \tilde{D}D^T + \tilde{\Sigma}_{bl}^{-1} \right)^{-1} \left( \tilde{D}\tilde{N}^T \right)_{(i)}, \left( \tilde{D}D^T + \tilde{\Sigma}_{bl}^{-1} \right)^{-1} \right)$$

  $$\tilde{d}_j \mid \sim MVN \left( \left( \tilde{B}^T\tilde{B} + I_L \right)^{-1} \left( \tilde{B}^T\tilde{N} \right)_{(j)}, \left( \tilde{B}^T\tilde{B} + I_L \right)^{-1} \right)$$

2. Sample regression coefficients.

- Sample $f_0, g_0, h_0$. Let $S_f = Y - \mu + \{Xf_0\}, S_g = Y - \mu + Ug_0$, and $S_h = Y - \mu + Vh_0$.

  $$f_0 \mid \sim MVN \left( \left( \sigma^{-2}X^TX + \sigma_f^{-2}I_r \right)^{-1} \left( \sigma^{-2}X^T \text{Vec}(S_f) \right), \left( \sigma^{-2}X^TX + \sigma_f^{-2}I_r \right)^{-1} \right)$$

  $$g_0 \mid \sim MVN \left( \left( \sigma^{-2}JU^TU + \sigma_g^{-2}I_p \right)^{-1} \left( \sigma^{-2}U^T \sum_j S_g \right), \left( \sigma^{-2}JU^TU + \sigma_g^{-2}I_p \right)^{-1} \right)$$

  $$h_0 \mid \sim MVN \left( \left( \sigma^{-2}JV^TV + \sigma_h^{-2}I_q \right)^{-1} \left( \sigma^{-2}V^T \sum_i S_h \right), \left( \sigma^{-2}JV^TV + \sigma_h^{-2}I_q \right)^{-1} \right)$$

- Sample $\tilde{f}_0, \tilde{g}_0, \tilde{h}_0$. Let $\tilde{S}_F = \tilde{Y} - \tilde{\mu} + \{X\tilde{f}_0\}, \tilde{S}_g = \tilde{Y} - \tilde{\mu} + \tilde{U}g_0$, and $\tilde{S}_h = \tilde{Y} - \tilde{\mu} + \tilde{V}h_0$.

  $$\tilde{f}_0 \mid \sim MVN \left( \left( \tilde{\sigma}^{-2}X^TX + \tilde{\sigma}_f^{-2}I_r \right)^{-1} \left( \tilde{\sigma}^{-2}X^T \text{Vec}(\tilde{S}_f) \right), \left( \tilde{\sigma}^{-2}X^TX + \tilde{\sigma}_f^{-2}I_r \right)^{-1} \right)$$
\[ \tilde{g}_0 | \sim \text{MVN}((\tilde{\sigma}^{-2} J U^T U + \tilde{\sigma}_g^{-2} I_p)^{-1} \left( \tilde{\sigma}^{-2} U^T \sum_j \tilde{S}_g \right), (\tilde{\sigma}^{-2} J U^T U + \tilde{\sigma}_g^{-2} I_p)^{-1}) \]

\[ \tilde{h}_0 | \sim \text{MVN}((\tilde{\sigma}^{-2} V^T V + \tilde{\sigma}_h^{-2} I_q)^{-1} \left( \tilde{\sigma}^{-2} V^T \sum_i \tilde{S}_h \right), (\tilde{\sigma}^{-2} V^T V + \tilde{\sigma}_h^{-2} I_q)^{-1}) \]

3. Sample variance components.

\[ \sigma^2 | \sim \text{IG}(IJ/2 + e, \sum_{i,j} (Y_{ij} - \mu_{ij})^2/2 + e) \]

\[ \sigma_{\alpha}^2 | \sim \text{IG}(I/2 + e, \sum_i (\alpha_i - \mu_0)^2/2 + e) \]

\[ \sigma_{\beta}^2 | \sim \text{IG}(J/2 + e, \sum_j \beta_j^2/2 + e) \]

\[ \tilde{\sigma}_{\alpha}^2 | \sim \text{IG}(I/2 + e, \sum_i (\tilde{\alpha}_i - \tilde{\mu}_0 - c_1 \alpha_i)^2/2 + e) \]

\[ \tilde{\sigma}_{\beta}^2 | \sim \text{IG}(J/2 + e, \sum_j (\tilde{\beta}_j - c_2 \beta_j)^2/2 + e) \]

\[ \sigma_f^2 | \sim \text{IG}(r/2 + e, f'_0 f_0/2 + e) \]

\[ \sigma_g^2 | \sim \text{IG}(p/2 + e, g'_0 g_0/2 + e) \]

\[ \sigma_h^2 | \sim \text{IG}(q/2 + e, h'_0 h_0/2 + e) \]

\[ \sigma_b^2 | \sim \text{IG}(IL/2 + e, \sum_i b'_i \text{diag}\{\phi^{-(l-1)}\} b_i/2 + e) \]

\[ \tilde{\sigma}_b^2 | \sim \text{IG}(IL/2 + e, \sum_i \tilde{b}'_i \text{diag}\{\tilde{\phi}^{-(l-1)}\} \tilde{b}_i/2 + e) \]
4. Sample mean variables $\mu_0, \tilde{\mu}_0$.

$$
\mu_0 \sim N((\sigma_\alpha^{-2} I + s^{-2})^{-1} \sigma_\alpha^{-2} \sum_i \alpha_i, (\sigma_\alpha^{-2} I + s^{-2})^{-1})
$$

$$
\tilde{\mu}_0 \sim N((\tilde{\sigma}_\alpha^{-2} I + s^{-2})^{-1} \tilde{\sigma}_\alpha^{-2} \sum_i (\tilde{\alpha}_i - \tilde{c}_1 \alpha_i), (\tilde{\sigma}_\alpha^{-2} I + s^{-2})^{-1})
$$

5. Sample $c_1, c_2, c_0$.

$$
c_1 \sim N((\sigma_\alpha^{-2} \sum_i \alpha_i^2 + s^{-2})^{-1} \sigma_\alpha^{-2} \sum_i \alpha_i (\tilde{\alpha}_i - \tilde{\mu}_0), (\sigma_\alpha^{-2} \sum_i \alpha_i^2 + s^{-2})^{-1})
$$

$$
c_2 \sim N((\tilde{\sigma}_\beta^{-2} \sum_j \beta_j^2 + s^{-2})^{-1} \tilde{\sigma}_\beta^{-2} \sum_j \beta_j (\tilde{\beta}_j), (\tilde{\sigma}_\beta^{-2} \sum_j \beta_j^2 + s^{-2})^{-1})
$$

$$
c_0 \sim N((\sum_{i,j} (b'_i d_j)^2 + s^{-2})^{-1} \sum_{i,j} b'_i d_j (\bar{Y}_{ij} - \tilde{\mu}_{ij} + c_0 b'_i d_j), (\sum_{i,j} (b'_i d_j)^2 + s^{-2})^{-1})
$$
REFERENCES


104


Efron, B. and Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM.


106


