

ABSTRACT

NAUGHTON, BRIAN PATRICK. Bayesian Regression Using Priors On The Model Fit. (Under the direction of Howard D. Bondell.)

Bayesian regression models have become a popular tool for many researchers, and offer many advantages over the frequentist approach. For example, Bayesian methods can incorporate prior subjective knowledge, shrink parameter estimates to give more reasonable results, and fit complex models that otherwise would not be possible. However, choosing prior distributions is an active area of research for Bayesian methods, and is the focus of this dissertation. Chapter 1 presents a literature review of typical priors used in regression, and describes how they might be elicited with prior knowledge.

Eliciting informative priors for linear regression models is challenging, especially with many parameters. Chapter 2 proposes a new approach to elicit information through a summary of the model fit, using R^2 . From previous studies, a meta-analysis, or expert knowledge, this quantity is more readily known to researchers using regression models. Putting a prior distribution on R^2 has the added benefit of shrinking the estimated coefficients, and penalizing overfitted models. We demonstrate how to fit a linear regression model with a prior on the R^2 , then present a modification to fit sparse and high dimensional models. Through simulated and real data examples, we show that the proposed model, even with default priors, outperforms other global-local shrinkage priors in estimation and predictive performance.

In Chapter 3, we extend the methods developed for the linear model to binary regression models. We demonstrate how to elicit a prior based on the distribution of a pseudo- R^2 , and sample from the posteriors for logistic and probit regression. A simulation study shows that both informative and default priors for logistic regression improve

performance compared with other default models.

Chapter 4 develops new methods for analyzing rank-ordered data in a Bayesian framework. This type of data is common in stated preference surveys because it allows researchers to extract more information than if the most preferred choice was recorded, such as with multinomial data. We develop an improved sampling algorithm for the rank-ordered probit model. Traditional approaches become intractable in the presence of a large number of choices, so also we introduce a multiple-shrinkage prior for the rank-ordered probit to accommodate high dimensional scenarios. Simulated and real data examples demonstrate the improved sampling performance of the methods.

© Copyright 2018 by Brian Patrick Naughton

All Rights Reserved

Bayesian Regression Using Priors On The Model Fit

by
Brian Patrick Naughton

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

2018

APPROVED BY:

Brian J. Reich

Eric C. Chi

Alyson G. Wilson

Howard D. Bondell
Chair of Advisory Committee

DEDICATION

To my wife, Katie.

BIOGRAPHY

Brian was born in Syracuse, NY, and graduated from Fayetteville-Manlius High School in 2005. He graduated Summa Cum Laude with a Bachelor of Science degree in Mathematics from the University of North Carolina at Charlotte in 2011. While a graduate student in Statistics at NC State, he worked as a Data Science Intern at Red Hat, as a Summer Research Intern at MIT Lincoln Laboratory, and as an instructor for ST311: Intro to Statistics. He received his Masters of Statistics in 2015 and PhD in 2018. Shortly after defending his PhD, Brian started working as a Data Scientist for Google in Mountain View, California.

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Howard Bondell, for your support, guidance and patience over the last few years. Many times I thought that I wouldn't make it, but your advice helped turn a daunting task into something I could achieve, and made this dissertation possible. I also would like to thank Brian Reich, Alyson Wilson and Eric Chi for serving on my committee, and for your valuable wisdom and feedback on my research.

I am grateful for my teaching mentors, Roger Woodard and Herle McGowan, who helped me grow as a teacher and communicator of statistics. The skills I developed teaching opened up many opportunities that I otherwise wouldn't have.

My graduate education would not have been the same without the excellent teachers in the department. Thanks to Len Stefanski, Dennis Boos, Hua Zhou, Emily Griffith, Joe Guinness, Brian Reich and Howard Bondell for teaching interesting classes. A special thanks also goes to Alison McCoy, Lanakila Alexander, and Dana Derosier for everything you've done to make the department such a friendly and pleasant place to work.

I would like to acknowledge my mentors: Keith Watkins and Megan Jones at Red Hat, and Michael Kotson at MIT Lincoln Lab. The business, communication, and research skills I learned working with all of you have greatly contributed to my development as a statistician.

Thanks to all of my friends and family. Your support has helped keep me sane throughout this journey.

Finally, my deepest gratitude goes to my wife, Katie. When I had the crazy idea turn down a job offer and pursue a PhD, you told me to go for it. Your trust in me never wavered, and I could not have done all of this without your love and support.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	viii
Chapter 1 Introduction	1
1.1 Priors for Regression	1
1.2 Prior elicitation	4
1.3 Dissertation Outline	6
Chapter 2 Linear Regression Using a Prior on the Model Fit	7
2.1 Introduction	7
2.2 Model Specification	10
2.2.1 Prior Distribution on R^2	10
2.2.2 Sparse regression and local shrinkage	12
2.3 Posterior Computation	14
2.3.1 Uniform-on-Ellipsoid Model	15
2.3.2 Local Shrinkage Model	16
2.3.3 High Dimensional Data	18
2.3.4 Selecting Hyperparameters	19
2.4 Simulations	21
2.4.1 Subjective Examples	21
2.4.2 Sparse Examples	24
2.5 Real Data Examples	28
2.6 Discussion	31
Chapter 3 Binary Regression Using a Prior on the Model Fit	33
3.1 Introduction	33
3.2 Model Specification	35
3.2.1 Prior Distribution on Pseudo- R^2	35
3.2.2 Posterior Computation	38
3.3 Simulations	38
3.3.1 Using Informative Priors	38
3.3.2 Sparse Regression Example	40
3.4 Discussion	43
Chapter 4 An Efficient Bayesian Rank-Ordered Probit Model	44
4.1 Introduction	44
4.2 Bayesian Rank-Ordered Probit Model	47
4.2.1 Preliminaries	47

4.2.2	Marginal Data Augmentation Algorithm	48
4.2.3	An Efficient MCMC Sampler	50
4.3	Multiple-Shrinkage Rank-Ordered Probit Model	51
4.4	Examples	53
4.4.1	MCMC Sampling Comparison	53
4.4.2	Model Comparison	57
4.5	Discussion	60
References		62
Appendices		71
Appendix A	Chapter 2 Proofs	72
A.1	Proposition 2.1	72
A.2	Proposition 2.2	73
A.3	Proposition 2.3	75
A.4	Choosing Hyperparameters ν and μ	75
Appendix B	MCMC Samplers for Logistic Regression	77
B.1	Gibbs sampler for Uniform-on-ellipsoid Model	77
B.2	MCMC sampler for Local Shrinkage Model	79

LIST OF TABLES

Table 2.1	Average sum of squared errors (and standard errors), from 200 randomly generated data sets. For each data set the true coefficients are generated independently from a $N(0, 1)$ distribution, and the responses from a $N(\mathbf{x}'\boldsymbol{\beta}_0, \sigma^2)$ distribution, where $n = p = 30$	24
Table 2.2	Average sum of squared errors (and standard errors), multiplied by 10 for readability, from 200 randomly generated data sets with $r = 0.5$	26
Table 2.3	Average sum of squared errors (and standard errors), multiplied by 10 for readability, from 200 randomly generated data sets with $r = 0.9$	27
Table 2.4	Average sum of squared error (and standard errors) from 200 randomly generated data sets, split by the SSE due to the zero and non-zero coefficients. The covariates (x_{ij}) are generated from a $N(0, 1)$ distribution with correlation $(x_{ij_1}, x_{ij_2}) = r^{ j_1 - j_2 }$, and q is the number of non-zero coefficients generated from a Uniform(0, 1) distribution.	29
Table 2.5	Average root mean square error (RMSE) (and standard errors) for real data examples.	30
Table 3.1	Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 3$ and correlation $(x_{ij}, x_{ij'}) = r^{ j - j' }$	41
Table 3.2	Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 5$ and correlation $(x_{ij}, x_{ij'}) = r^{ j - j' }$	41
Table 3.3	Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 10$ and correlation $(x_{ij}, x_{ij'}) = r^{ j - j' }$	42
Table 3.4	Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients for the sparse regression example. All errors are multiplied by 10 for readability. The dimension is $p = 20$ and correlation $(x_{ij}, x_{ij'}) = r^{ j - j' }$. Both BEERS put a Uniform distribution on R^2	42
Table 4.1	Median rank correlation coefficients (and IQRs) for the simulation study with $p \in \{5, 10, 20\}$	59
Table 4.2	Median rank correlation coefficients (and IQRs), and the median effective sampling rate (ESS per minute) for the Sushi data sets.	60

LIST OF FIGURES

Figure 2.1	10,000 samples from the prior of $\beta \sigma^2, R^2$ for different choices of ν , with $\mu = 1$	14
Figure 2.2	Histograms of all pairwise correlations of \mathbf{x}_j for the Cereal, Cookie and Multidrug data sets.	30
Figure 4.1	Distributions of effective sample size from 5,000 MCMC iterations for $p = 5$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	54
Figure 4.2	Distributions of effective sample size from 5,000 MCMC iterations for $p = 10$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	55
Figure 4.3	Distributions of effective sample size from 5,000 MCMC iterations for $p = 20$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	55
Figure 4.4	Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 5$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	56
Figure 4.5	Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 10$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	56
Figure 4.6	Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 20$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).	57
Figure 4.7	Mean autocorrelation functions (ACF) across samples of the 12 β_j in the Voters' Preference for Political Parties in Japan data.	58

Chapter 1

Introduction

1.1 Priors for Regression

We begin the dissertation with a brief literature review of priors that are routinely used in Bayesian regression problems. In the absence of subjective knowledge, and desire to fit a regression model in a Bayesian setting, a common approach is to use non-informative priors on the regression coefficients, $\boldsymbol{\beta}$, and the error variance, σ^2 . The Jeffrey's prior puts an improper, uniform prior on $(\boldsymbol{\beta}, \log \sigma)$, that is,

$$p(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \quad (1.1)$$

When there is a large amount of data, this is convenient and acceptable because it yields a proper posterior when \mathbf{X} is full rank, and analytical solutions for the posterior distribution. Namely, $\boldsymbol{\beta} | (\sigma^2, \mathbf{y}) \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}_{LS}, (\mathbf{X}'\mathbf{X})^{-1} \sigma^2)$, where $\hat{\boldsymbol{\beta}}_{LS}$ is the least-squares estimate for $\boldsymbol{\beta}$. Another common approach is to use conjugate priors, namely an inverse Gamma distribution for σ^2 and a multivariate Normal distribution for $\boldsymbol{\beta} | \sigma^2$. The posterior

can be computed analytically or easily simulated via Gibbs sampling, depending on the variance structure of β . In the presence of substantial prior knowledge, the prior mean and variance of β can be specified here.

Another common default prior for regression is the g -prior – which is a multivariate Normal prior for β with covariance proportional to the covariance of $\hat{\beta}_{LS}$ (Zellner, 1986). That is, $\beta | (\sigma^2, g) \sim N(\tilde{\beta}, g \sigma^2 (\mathbf{X}'\mathbf{X})^{-1})$, where $\tilde{\beta}$ is typically $\mathbf{0}$. The parameter g can be fixed based on subjective knowledge, can be a function of n (i.e. $g \propto n$ or $g \propto n/\log n$), or can be given a prior such as the conjugate inverse-Gamma, which can be either informative or non-informative. In addition to the conjugacy of the g -prior for fixed g , it has the nice property that the posterior mean, $\hat{\beta}$ is a linear combination of $\hat{\beta}_{LS}$ and the prior mean $\tilde{\beta}$, i.e., $\hat{\beta} = (g\hat{\beta}_{LS} + \tilde{\beta})/(1+g)$. Small values of g shrink the least squares estimate towards the prior mean, which suggests the unit information prior, $g = n$, as a default choice. Liang et al. (2008) introduced a family of mixing distributions for g -priors for model selection, which were also independently studied by Cui and George (2008). Note that if $\tilde{\beta} = \mathbf{0}$, then $\hat{\beta} = \frac{g}{1+g}\hat{\beta}_{LS}$, where $0 < \frac{g}{1+g} < 1$. These so-called “hyper- g ” priors put a Beta distribution on the shrinkage factor $\frac{g}{1+g}$.

Regularization methods for regression are commonly used for better estimation, prediction, and model selection. Many of the popular Bayesian regularization methods fit into a family of so-called global-local shrinkage priors (Polson and Scott, 2010). The regression coefficients can be represented as a scale mixture of Normals:

$$\begin{aligned}\beta_i &\sim N(0, \sigma^2 \tau^2 \lambda_i^2) \\ \lambda_i^2 | \tau^2 &\sim p(\lambda_i^2) \\ (\tau^2, \sigma^2) &\sim p(\tau^2, \sigma^2),\end{aligned}$$

Note that the Normal distribution can be replaced by an alternative kernel, for example a Laplace. The Laplace itself, however, can be expressed as a scale mixture of Normals, and that is the expression we use here. The global variance component (τ^2) controls shrinkage for all coefficients, and the local variance components (λ_i^2) encourage sparsity by allowing some coefficients to shrink faster than others. One of the most popular methods is the Bayesian Lasso (Park and Casella, 2008) which puts a Laplace prior on β_i , or equivalently independent Exponential(τ^2) priors on λ_i^2 . The global variance component, τ^2 , can be chosen based by empirical Bayes through marginal maximum likelihood, or given a Gamma hyperprior as recommended by Park and Casella (2008) and Kyung et al. (2010). The posterior mode of β is equivalent to the Lasso solution by Tibshirani (1996) with regularization parameter τ . Griffin and Brown (2010) introduced Normal-Gamma priors to more flexibly handle different types of sparsity than the Bayesian Lasso. These put independent Gamma(α, τ^2) distributions on λ_i^2 , and include the Bayesian Lasso as a special case when $\alpha = 1$. Small values of α encourage sparser solutions, that is, only a few of the coefficients account for most of the total variance of β . Bhattacharya et al. (2015) proposed the class of Dirichlet-Laplace (DL) priors to also provide more flexibility for local shrinkage through an additional Dirichlet prior on the variances. Specifically, the DL prior puts a Gamma distribution on τ , and $\lambda_i^2 = \phi_j \psi_j^2$, where ψ is an Exponential(1/2) distribution and ϕ has a Dirichlet distribution with common parameter a . The Horseshoe prior of Carvalho et al. (2010) is useful for sparse signals because it puts infinite mass at the origin, but has Cauchy-like tails that are robust to large coefficients. This is done by putting a half-Cauchy prior on λ_i with scale parameter τ , and a Cauchy prior on τ with scale parameter σ .

The coefficient of determination, R^2 , has also been used to assist with the choice of shrinkage priors. Zhang and Bondell (2016) show how global-local priors can be used

for variable selection through penalized credible regions with specific attention to the Dirichlet-Laplace priors. Selecting the hyperparameters is a routine problem with regularized regression, so R^2 is used to choose the hyperparameters. If a particular Beta(a, b) distribution is desired for R^2 , then this induces a particular choice of the prior. This is accomplished by simulating the prior on a grid of parameters, or by minimizing the Kullback-Leibler (KL) divergence between the induced distribution of R^2 and the desired Beta(a, b) distribution. Zhang et al. (2016) propose the R^2 induced Dirichlet decomposition, or R2-D2 prior, extending the DL prior by putting a Beta prior on the marginal prior distribution of R^2 . This is equivalent to putting a Beta-prime, or inverted-Beta distribution on τ^2 , and the product of a Dirichlet distribution and an Exponential(1/2) distribution on $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)$ in the global-local framework.

1.2 Prior elicitation

Prior elicitation is the process of quantifying subjective beliefs and translating them into a probability distribution for unknown quantities. Despite its difficulty it encourages the expert to think in terms of the parameters being modeled, and brings the analysis closer to the application (Garthwaite et al., 2005). In addition, a subjective Bayesian perspective can be helpful in problems where traditional approaches would not be feasible (Goldstein, 2006). Elicitation of prior distributions has been used in subjective Bayesian analyses in a variety of applications, including for example: Ecology (Choy et al., 2009; Denham and Mengersen, 2007; Kuhnert et al., 2010), Clinical Trials (Johnson et al., 2010), and Political Science (Gill and Walker, 2005).

Kadane and Wolfson (1998) describe two classes of elicitation techniques in the literature – “structural” and “predictive”. Structural methods attempt to elicit the dis-

tributions of parameters directly, and are typically based on observable quantities. This is often not feasible if the parameters are unobserved such as the coefficients in regression. According to the psychological literature on elicitation, people are generally good at quantifying measures of central tendency for symmetric distributions, but not as good at quantifying measures of variability (Garthwaite et al., 2005). Therefore, eliciting the distribution of regression parameters would be quite difficult without substantial expert knowledge.

Predictive elicitation quantifies the parameters by asking a series of questions about the predictive distribution, and conditional on the covariates in regression setup. Kadane et al. (1980) proposed an interactive method to elicit a multivariate Normal distribution for β in the conjugate linear model, by first estimating the predictive distribution. The expert is asked to estimate the median of the predictive distribution, $y_{.5}|\mathbf{x}$, for different design points, \mathbf{x} , which are used to estimate the mean of β . The expert then predicts higher quantiles (e.g. $y_{.75}, y_{.875}, y_{.9375}$) which are used to estimate the hyperparameters and covariance of β . Garthwaite and Dickey (1988, 1992) elicit medians and quantiles of the mean response \bar{y} given the design for variable selection in regression.

While these have been the preferred elicitation methods for regression, the expert will still have to elicit responses for many more design points as the dimension grows, requiring prior knowledge about the effects of all predictors. It may also be difficult to make predictions when experts have experience with similar problems, but not necessarily the present experiment. This dissertation proposes solutions to some of these problems.

1.3 Dissertation Outline

This dissertation introduces new methods for a variety of Bayesian regression problems. Chapter 2 proposes a novel method for eliciting a prior distribution for linear regression based on the model fit, specifically through the coefficient-of-determination, R^2 . Chapter 3 extends this approach to binary data, by eliciting priors for logistic and probit regression using a pseudo- R^2 measure. Bayesian regression models for rank-ordered data are considered in Chapter 4. We propose a new sampler for rank-ordered probit models, and introduce a new multiple shrinkage prior to accommodate data with many choices. Each Chapter concludes with a discussion of the contributions to the field, and opportunities for future research. Chapters 2 through 4 are from journal articles that are to be submitted.

Chapter 2

Linear Regression Using a Prior on the Model Fit

2.1 Introduction

Bayesian inference provides a framework to update subjective prior beliefs with observed data. But specifying a probability distribution conveying this subjective knowledge poses a challenge. We consider priors for the p -dimensional coefficient vector, $\boldsymbol{\beta}$, and the error variance σ^2 in the linear regression model, $y = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$, with n observations, and $\varepsilon \sim N(0, \sigma^2)$. Elicitation proves particularly difficult in regression problems, requiring joint distributions for the unobserved regression coefficients. For more than a few variables this becomes intractable, and researchers often choose default priors out of convenience. Additionally, in the high-dimensional case, choosing default priors is also a problem as tuning parameters play a large role.

The subjective Bayesian approach takes advantage of expert knowledge to assist in limited data applications, and it may be the only feasible approach for complex prob-

lems (Goldstein, 2006). It also brings researchers closer to the application by encouraging them to think deeply about the model parameters (Garthwaite et al., 2005). Elicitation of informative prior distributions have contributed to a variety of applications, including for example: Ecology (Choy et al., 2009; Denham and Mengersen, 2007; Kuhnert et al., 2010), Clinical Trials (Johnson et al., 2010), and Political Science (Gill and Walker, 2005). Kadane and Wolfson (1998) describe two classes of elicitation methods, structural and predictive. Structural methods elicit parameters directly, while predictive methods elicit parameters via the predictive distribution, by having the expert make predictions given hypothetical covariates (e.g. Garthwaite and Dickey (1988); Kadane et al. (1980)). The predictive approach is preferred for regression because it elicits priors through observable quantities, but it becomes intractable for more than a few parameters. Garthwaite et al. (2005) recommend eliciting summaries of the marginal distribution for multivariate problems, but leave it as an open question as to which summaries are effective and reliable. We address this problem by directly putting a Beta prior on the well known R^2 , inducing a prior for the regression coefficients. This allows researchers to select priors based on previous studies (e.g. a meta-analysis), or based on domain specific knowledge about the expected variation explained by the predictors. Summaries of previously observed R^2 values can be empirically matched to a Beta distribution, imparting transparency to the subjective approach.

With limited prior knowledge, we often choose regression priors for computational convenience and for their shrinkage properties. The Zellner’s g -prior places a conjugate, multivariate Normal distribution on $\boldsymbol{\beta}$, with prior covariance proportional to the variance of the least-squares estimator, $\widehat{\boldsymbol{\beta}}_{LS}$ (Zellner, 1986). That is, $\boldsymbol{\beta} | (\sigma^2, g) \sim N_p \left(\tilde{\boldsymbol{\beta}}, g \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \right)$, where $\tilde{\boldsymbol{\beta}}$ is typically $\mathbf{0}$. For fixed g , the posterior mean, $\widehat{\boldsymbol{\beta}}$, is a linear combination of $\tilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_{LS}$, with small values of g shrinking $\widehat{\boldsymbol{\beta}}$ toward the prior mean. The hyperparameter,

g , determines the influence of the prior (i.e. $g = n$ is the unit information prior), or has a prior itself, such as a conjugate Inverse-Gamma distribution.

The global-local (GL) shrinkage priors are often used for regression problems (e.g. Polson and Scott, 2010). These scale mixture-of-Normals priors have variance components controlling for global shrinkage of all coefficients, and local shrinkage of some components faster than others. Examples include the Bayesian Lasso (Park and Casella, 2008), the Normal-Gamma priors (Griffin and Brown, 2010), the Dirichlet-Laplace priors (Bhattacharya et al., 2015), and the Horseshoe (Carvalho et al., 2010). By itself, R^2 is a poor measure of model fitness because it increases as the model over-fits the data. A Beta prior on R^2 can shrink large values of R^2 , thus regularizing β in a global sense for a better model fit. We will show that a particular class of g -priors is a special case of our proposed class of priors, but we also introduce additional variance components encouraging local shrinkage.

The value of R^2 has been used to specify priors, and select hyperparameters for regularization problems. Scott and Varian (2014) elicit an informative distribution for the error variance, σ^2 , based on the expected R^2 , and the response, \mathbf{y} . Zhang and Bondell (2016) choose hyperparameters for GL shrinkage priors by minimizing the Kullback-Leibler divergence between the expected distribution of R^2 and a Beta distribution. Expanding on this idea, the R^2 -induced Dirichlet Decomposition, or R2-D2 prior, assumes the marginal distribution of R^2 (after integrating out β and \mathbf{X}) has a Beta distribution (Zhang et al., 2016). We propose placing a Beta prior on the conditional distribution of R^2 given the model design, \mathbf{X} . This induces a prior directly on the coefficients, consistent with the usual design-based interpretation of R^2 . Additionally, a prior distribution on R^2 also acts as a shrinkage prior, to prevent overfitting and encourage regularization in high-dimensional problems. This approach gives better estimates with subjective prior

knowledge, and more flexibly handles sparse data compared with previous approaches. We also recommend some default priors that perform well even in high-dimensional cases.

The rest of this chapter is presented as follows. Section 2.2 describes how we induce a prior distribution on the regression coefficients from a Beta distribution on R^2 . We incorporate ideas from global-local shrinkage priors to improve its performance for sparse and high-dimensional regression problems. Section 2.3 explains how to efficiently sample from the posterior distributions using Markov chain Monte Carlo (MCMC) methods. In Section 2.4 we present a simulation study demonstrating the performance of the proposed model with prior knowledge, and its effectiveness as a shrinkage prior in high-dimensional settings. Section 2.5 explores the predictive performance on several high-dimensional data sets. Section 2.6 summarizes the impact of the proposed model, and discusses future directions for this research. The proofs for all propositions and results are given in Appendix A.

2.2 Model Specification

2.2.1 Prior Distribution on R^2

We consider the linear model with independent Normally distributed errors. That is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

where \mathbf{y} is the n -dimensional vector of responses, \mathbf{X} is an $n \times p$ -dimensional matrix of covariates, $\boldsymbol{\beta}$ is the p -dimensional vector of regression coefficients, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_p)$, and σ^2 is the error variance. The typical Bayesian approach specifies a joint distribution on the model parameters, namely for the regression coefficients and error variance. Instead,

we specify a distribution for R^2 and σ^2 , and then require a conditional distribution for $\boldsymbol{\beta}$. To formulate the model in terms of R^2 , we first define R^2 as the squared correlation between $E y = \mathbf{x}'\boldsymbol{\beta}$, and the observed response $y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. That is,

$$R^2 = \frac{\text{Cov}^2(y, \mathbf{x}'\boldsymbol{\beta})}{\text{Var}(y)\text{Var}(\mathbf{x}'\boldsymbol{\beta})} = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta})}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}) + \sigma^2}. \quad (2.2)$$

We consider R^2 as a function of $\boldsymbol{\beta}$ and σ^2 , via conditioning on the parameters in the variance calculation to obtain

$$R^2(\boldsymbol{\beta}, \sigma^2) = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta} | \boldsymbol{\beta})}{\text{Var}(\mathbf{x}'\boldsymbol{\beta} | \boldsymbol{\beta}) + \sigma^2} = \frac{\boldsymbol{\beta}'\text{Var}(\mathbf{x})\boldsymbol{\beta}}{\boldsymbol{\beta}'\text{Var}(\mathbf{x})\boldsymbol{\beta} + \sigma^2} = \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta} + \sigma^2}, \quad (2.3)$$

where $\boldsymbol{\Sigma}_X = \text{Var}(\mathbf{x})$. We specifically write $R^2(\boldsymbol{\beta}, \sigma^2)$ to reflect the fact that R^2 depends on the unknown vector $\boldsymbol{\beta}$ as well as σ^2 . We center the columns of \mathbf{X} to have mean 0, and plug in our sample estimate, $\mathbf{X}'\mathbf{X}/n$, for $\boldsymbol{\Sigma}_X$, giving the following expression for R^2 :

$$R^2(\boldsymbol{\beta}, \sigma^2) = \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + n\sigma^2} \quad (2.4)$$

Notice that (2.4) is the familiar statistic, \widehat{R}^2 , if the maximum likelihood estimates were substituted for $\boldsymbol{\beta}$ and σ^2 . Conditional on σ^2 and \mathbf{X} , a distribution on $R^2(\boldsymbol{\beta}, \sigma^2)$ induces a distribution on the quadratic form, $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$.

As a natural choice, we choose a Beta(a, b) prior for R^2 , where the choices of a and b will be discussed in more detail in Section 2.3.4. An Inverse-Gamma(α_0, β_0) prior is used for σ^2 , but note that any other distribution will also work. A prior for $\boldsymbol{\beta} | (R^2, \sigma^2)$ is thus defined on the surface of the ellipsoid: $\{\boldsymbol{\beta} : \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = nR^2\sigma^2/(1 - R^2)\}$. We call the proposed prior the Beta-Ellipsoid R-Squared (BEERS) prior. When $\mathbf{X}'\mathbf{X}$ is full rank we may choose $\boldsymbol{\beta}$ to be uniformly distributed on this ellipsoid; that is, the distribution of $\boldsymbol{\beta}$ is

constant given the quadratic form. The following proposition shows that this “uniform-on-ellipsoid” prior has a closed-form elliptical distribution for $\boldsymbol{\beta}$ after integrating out R^2 .

Proposition 2.1. *If $R^2|\sigma^2$ has a $Beta(a,b)$ distribution and $\boldsymbol{\beta}|(R^2, \sigma)$ is uniform on the ellipsoid, then $\boldsymbol{\beta}|\sigma^2$ has the probability density function:*

$$p(\boldsymbol{\beta}|\sigma^2) = \frac{\Gamma(p/2) |\mathbf{S}_X|^{1/2}}{B(a, b) \pi^{p/2}} (\sigma^2)^{-a} (\boldsymbol{\beta}'\mathbf{S}_X\boldsymbol{\beta})^{a-p/2} (1 + \boldsymbol{\beta}'\mathbf{S}_X\boldsymbol{\beta}/\sigma^2)^{-(a+b)}, \quad (2.5)$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{S}_X = \mathbf{X}'\mathbf{X}/n$, and B denotes the Beta function.

As a special case, if $a = p/2$ and $b = 1/2$, then $\boldsymbol{\beta}$ has a multivariate Cauchy distribution with spread parameter \mathbf{S}_X/σ^2 . Zellner and Siow (1980) recommended these Cauchy priors for model selection problems. The next proposition shows that for $a \leq p/2$, the distribution in (2.5) is equivalent to a mixture of Normals g -prior, with a hyperprior on g that is the product of Beta and Inverse-Gamma distributions.

Proposition 2.2. *If $\boldsymbol{\beta}|(\sigma^2, z, w) \sim N_p(\mathbf{0}, zw\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, $z \sim \text{Inverse-Gamma}(b, n/2)$, $w \sim \text{Beta}(a, p/2 - a)$, and $a \leq p/2$, then $\boldsymbol{\beta}|\sigma^2$ has the distribution given by the density in (2.5).*

This convenient representation eases the posterior computations yielding a straightforward Gibbs sampler discussed in Section 2.3.

2.2.2 Sparse regression and local shrinkage

The prior on R^2 regulates $\boldsymbol{\beta}$ through the quadratic form $\boldsymbol{\beta}'\mathbf{S}_X\boldsymbol{\beta}$, which can shrink the regression coefficients globally, but lacks the flexibility to handle different forms of sparsity. In addition, the posterior is not a proper distribution when \mathbf{X} is not full rank (e.g.

when $n < p$). Rather than letting $\boldsymbol{\beta} | (R^2, \sigma)$ be uniformly distributed on the ellipsoid, we put a Normal-Gamma prior on $\boldsymbol{\beta}$ (Griffin and Brown, 2010), but restrict its support to lie on the surface of the ellipsoid. Specifically, we let

$$\boldsymbol{\beta} | R^2, \sigma^2, \boldsymbol{\Lambda} \sim N_p \left(\mathbf{0}, \frac{\sigma^2 R^2}{1 - R^2} \boldsymbol{\Lambda} \right) \mathbb{1} \left\{ \boldsymbol{\beta}' \mathbf{S}_X \boldsymbol{\beta} = \frac{\sigma^2 R^2}{1 - R^2} \right\} \quad (2.6)$$

$$\lambda_j \sim \text{Gamma}(\nu, \mu), \text{ for } j = 1, \dots, p, \quad (2.7)$$

where $\boldsymbol{\Lambda} = \text{diag} \{ \lambda_1, \dots, \lambda_p \}$ and $\mathbb{1} \{ \cdot \}$ is the indicator function. Note that this prior no longer requires \mathbf{S}_X to be full rank for it to be proper. Proposition 3 shows that the induced model described in the previous section is a special case of the hierarchical model proposed here with a non-diagonal, fixed $\boldsymbol{\Lambda}$ instead.

Proposition 2.3. *If $\boldsymbol{\beta} | (R^2, \sigma^2, \boldsymbol{\Lambda})$ has the distribution in (2.6), and $\boldsymbol{\Lambda} = \mathbf{X}'\mathbf{X}^{-1}$, then $\boldsymbol{\beta} | \sigma^2$ has the distribution in (2.5).*

That is, if the contours of the Normal distribution align with the ellipsoid, then we recover the uniform-on-ellipsoid prior.

The conditional distribution of $\boldsymbol{\beta}$ is similar to a Bingham distribution, which is a multivariate Normal distribution conditioned to lie on the unit sphere. This Bingham distribution has density $f(\mathbf{w} | \mathbf{A}) = C_A^{-1} \exp\{-\mathbf{w}'\mathbf{A}\mathbf{w}\}$ with respect to the uniform measure on the $p - 1$ dimensional unit sphere, where C_A is the Normalizing constant (Bingham, 1974). Here, $\boldsymbol{\beta} | (R^2, \sigma^2, \boldsymbol{\Lambda})$ is a Bingham distributed random vector that has been rotated and scaled to lie on the ellipsoid defined in Section 2.2.1, rather than the unit sphere. The matrix \mathbf{X} determines the rotation of the ellipsoid, R^2 and σ^2 determine the size of the ellipsoid, and the conditional prior on $\boldsymbol{\beta}$ determines the direction to the surface. If the local variance components (λ_j) are small, then regions of the ellipsoid near the axes

will be favored, encouraging sparser estimates. Like the Normal-Gamma priors, this is primarily controlled by the shape parameter, ν . Figure 2.1 illustrates the local shrinkage properties of the prior, showing 10,000 samples of $\beta | (\sigma^2, R^2)$. In the next section, we discuss how to select the hyperparameters ν and μ .

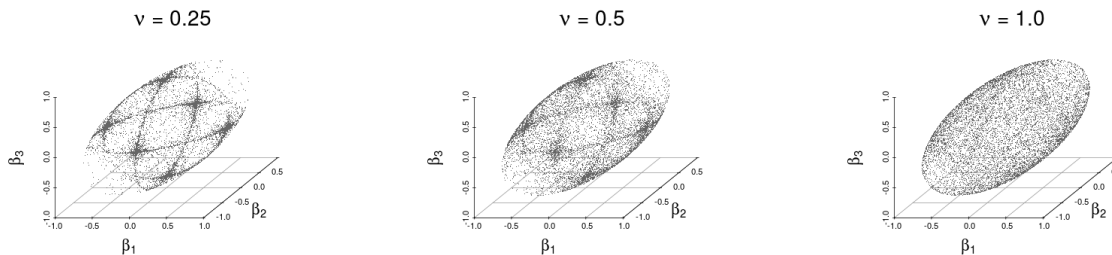


Figure 2.1: 10,000 samples from the prior of $\beta | \sigma^2, R^2$ for different choices of ν , with $\mu = 1$.

2.3 Posterior Computation

In this section we describe how to sample from the posterior distributions using Markov chain Monte Carlo (MCMC) samplers, starting with the uniform-on-ellipsoid model, and then the local shrinkage model. Section 2.3.3 illustrates how to modify the samplers to handle high-dimensional data (i.e. $p > n$), or when \mathbf{X} is not full rank. We conclude the section with some guidelines for selecting the hyperparameters with and without prior knowledge.

2.3.1 Uniform-on-Ellipsoid Model

In Section 2.2.1, we showed that $\boldsymbol{\beta}|\sigma^2$ has a mixture of Normals representation for $a \leq p/2$. If σ^2 is Inverse-Gamma(α_0, β_0) distributed, then $\boldsymbol{\beta}, \sigma^2, z$ and w are drawn from their full conditional distributions in the Gibbs sampler described as follows:

- (a) Set initial values for $\boldsymbol{\beta}, \sigma^2, z$ and w .
- (b) Sample $w | (\boldsymbol{\beta}, \sigma^2, z)$ by first sampling $u \sim \text{Gamma}(p/2 - a, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} / (2z\sigma^2))$, and setting $w = 1/(1 + u)$.
- (c) Sample $z | (\boldsymbol{\beta}, \sigma^2, w) \sim \text{Inverse-Gamma}(p/2 + b, \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} / (2w\sigma^2) + n/2)$.
- (d) Sample $\sigma^2 | (\boldsymbol{\beta}, z, w) \sim \text{Inverse-Gamma}((n + p)/2 + \alpha_0, \text{SSE}/2 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} / (2wz) + \beta_0)$, where $\text{SSE} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.
- (e) Sample $\boldsymbol{\beta} | (\sigma^2, z, w) \sim \text{Normal}(c\widehat{\boldsymbol{\beta}}_{LS}, c\sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, where $c = zw/(zw + 1)$ is the shrinkage factor, and $\widehat{\boldsymbol{\beta}}_{LS}$ is the least-squares estimate of $\boldsymbol{\beta}$.
- (f) Repeat steps b-e until convergence.

If $a > p/2$, then we can use any standard MCMC algorithm to sample from the posterior of $p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta} | \sigma^2) p(\sigma) p(\mathbf{y} | \boldsymbol{\beta}, \sigma^2)$. In particular we use a Metropolis random walk to sample $\boldsymbol{\beta}_{(t)} \sim \text{N}(\boldsymbol{\beta}_{(t-1)}, V_\beta)$, and $\log \sigma^2_{(t)} \sim \text{N}(\log \sigma^2_{(t-1)}, V_{\sigma^2})$. The jumping variances, V_β and V_{σ^2} , are tuned during the burn-in phase to achieve optimal acceptance rates.

2.3.2 Local Shrinkage Model

We now develop an MCMC algorithm for the local shrinkage model, sampling from the full conditionals using a Metropolis-Hastings sampler. First, we take the eigen-decomposition of $\mathbf{S}_X = \mathbf{X}'\mathbf{X}/n = \mathbf{V}\mathbf{D}\mathbf{V}'$, where $\mathbf{V}_{p \times p}$ is an orthogonal matrix of eigenvectors, and $\mathbf{D}_{p \times p}$ is a diagonal matrix of eigenvalues. R^2 is transformed such that $\theta = R^2/(1 - R^2)$ has a Beta-Prime (or Inverted-Beta distribution), with density $p(\theta) = \theta^{a-1} (1 + \theta)^{-a-b} / B(a, b)$. We also transform $\boldsymbol{\beta}$ to lie on the unit sphere conditional on the other variables; that is, $\boldsymbol{\gamma} = \mathbf{D}^{1/2}\mathbf{V}'\boldsymbol{\beta}/\sqrt{\theta\sigma^2}$. Then $\boldsymbol{\gamma}|\boldsymbol{\Lambda}$ has a Bingham distribution, and we can write the full model as follows.

$$\begin{aligned} \mathbf{y} | (\boldsymbol{\gamma}, \sigma^2, \theta) &\sim N_n \left(\sqrt{\theta\sigma^2} \mathbf{X}\mathbf{V}\mathbf{D}^{-1/2}\boldsymbol{\gamma}, \sigma^2 \mathbf{I} \right) \\ \boldsymbol{\gamma} | \boldsymbol{\Lambda} &\sim \text{Bingham} \left(\mathbf{D}^{-1/2}\mathbf{V}'\boldsymbol{\Lambda}^{-1}\mathbf{V}\mathbf{D}^{-1/2}/2 \right) \\ \sigma^2 &\sim \text{Inverse-Gamma}(\alpha_0, \beta_0) \\ \theta &\sim \text{Beta-Prime}(a, b) \\ \lambda_j &\sim \text{Gamma}(\nu, \mu), \text{ for } j = 1, \dots, p \end{aligned}$$

This parametrization models the direction of $\boldsymbol{\beta}$ independently of σ^2 and R^2 . However, the Bingham distribution of the direction, $\boldsymbol{\gamma}|\boldsymbol{\Lambda}$, contains an intractable normalizing constant, $C_{X,\boldsymbol{\Lambda}}$, depending on \mathbf{X} and $\boldsymbol{\Lambda}$. Specifically, $C_{X,\boldsymbol{\Lambda}}$ is a confluent hypergeometric function with matrix argument $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{V}'\boldsymbol{\Lambda}^{-1}\mathbf{V}\mathbf{D}^{-1/2}/2$.

The full conditional posterior distribution of $\boldsymbol{\gamma}$ is a Fisher-Bingham($\boldsymbol{\mu}, \mathbf{A}$) distribution (Kent, 1982), where $\boldsymbol{\mu} = \mathbf{y}'\mathbf{X}\mathbf{V}\mathbf{D}^{-1/2}\sqrt{\theta/\sigma^2}$. This is equivalent to a $N_p(\mathbf{A}^{-1}\boldsymbol{\mu}, \mathbf{A}^{-1})$ distribution conditioned to lie on the $p - 1$ dimensional unit sphere. We sample from this using the rejection sampler proposed by Kent et al. (2013) with an Angular Central

Gaussian (ACG) envelope distribution. Sampling efficiently from the ACG distribution is possible because it is just the marginal unit direction of a multivariate Normal distribution with mean $\mathbf{0}$, and thus only requires draws from a Normal distribution. Any standard MCMC algorithm can sample θ and σ^2 , but the Adaptive Metropolis algorithm (Haario et al., 2001) automatically accounts for the strong negative correlation between the parameters without the need for manual tuning. A bivariate Normal proposal distribution is used for (σ^2, θ) with covariance proportional to the running covariance of the samples during the burn-in phase.

The density function of the full conditional posteriors for variance parameters, λ_j , almost have Generalized Inverse Gaussian distributions (GIG) if it were not for the intractable term $C_{X,\Lambda}$. A Metropolis-Hastings algorithm would require computing this quantity. Our solution is to propose each candidate λ_j^* from a GIG distribution, and introduce auxiliary variables, $\boldsymbol{\gamma}^*|\boldsymbol{\Lambda}^*$, from a Bingham distribution in which the constant, C_{X,Λ^*} , appears in the density. We calculate the Metropolis-Hastings acceptance probability for $(\boldsymbol{\Lambda}^*, \boldsymbol{\gamma}^*)$, and since C_{X,Λ^*} appears in the posterior and the proposal distribution, we avoid computing it. This is the so-called ‘‘Exchange Algorithm’’ proposed by Murray et al. (2006) for doubly-intractable distributions, and also used by Fallaize and Kypraios (2016) for Bayesian inference of the Bingham distribution.

The entire sampler for the local shrinkage model is described as follows:

- (a) Set initial values for $\boldsymbol{\gamma}, \sigma^2, \theta$, and $\boldsymbol{\Lambda}$.
- (b) Sample $\boldsymbol{\gamma}|\boldsymbol{\Lambda}$ from a Fisher-Bingham($\boldsymbol{\mu}, \mathbf{A}$) distribution, where $\boldsymbol{\mu} = \mathbf{y}'\mathbf{X}\mathbf{V}\mathbf{D}^{-1/2}\sqrt{\theta/\sigma^2}$ and $\mathbf{A} = \mathbf{D}^{-1/2}\mathbf{V}'\boldsymbol{\Lambda}^{-1}\mathbf{V}\mathbf{D}^{-1/2}/2$.
- (c) Sample (σ^2, θ) jointly using an Adaptive Metropolis algorithm from a bivariate Normal distribution.

(d) Exchange algorithm to sample Λ :

- (i) Sample $\lambda_j^* | u_j \sim \text{GIG}(\nu, 2\mu, u_j^2)$, for $j = 1, \dots, p$, where $\mathbf{u} = \mathbf{V}\mathbf{D}^{-1/2}\boldsymbol{\gamma}$.
- (ii) Sample \mathbf{u}^* from a Bingham(Λ^*) distribution, where $\Lambda^* = \mathbf{D}^{-1/2}\mathbf{V}'\Lambda^{*-1}\mathbf{V}\mathbf{D}^{-1/2}/2$
- (iii) Accept $(\Lambda^*, \mathbf{u}^*)$ with probability:

$$\frac{p(\Lambda^*|\mathbf{u})}{p(\Lambda|\mathbf{u})} \times \frac{q(\Lambda|\mathbf{u})}{q(\Lambda^*|\mathbf{u})} \times \frac{q(\mathbf{u}^*|\Lambda)}{q(\mathbf{u}^*|\Lambda^*)} = \exp \left\{ \sum_{j=1}^p u_j^{*2} (1/\lambda_j^{*2} - 1/\lambda_j^2) \right\}$$

(e) Repeat steps b-d until convergence, and calculate $\boldsymbol{\beta} = \sqrt{\theta\sigma^2}\mathbf{V}\mathbf{D}^{-1/2}\boldsymbol{\gamma}$ for each sample.

In step (d), $p(\Lambda|\mathbf{u})$ is the conditional posterior distribution of Λ ; $q(\Lambda|\mathbf{u})$ is the GIG proposal distribution with density: $p(x; d, a, b) \propto x^{d-1} \exp\{-(ax + b/x)/2\}$, for $-\infty < d < \infty$, and $a, b > 0$; and $q(\mathbf{u}|\Lambda)$ is the Bingham proposal distribution which has the same constant as in $p(\Lambda|\mathbf{u})$. We only need to keep Λ^* at each step, and can discard \mathbf{u}^* . We can efficiently sample from the Bingham distribution because it is a special case of the Fisher-Bingham with $\boldsymbol{\mu} = \mathbf{0}$ (Kent et al., 2013). All modeling is done in terms of $\boldsymbol{\gamma}, \sigma^2, \theta$ and Λ . Since we don't sample the regression parameters directly, at each step we calculate $\boldsymbol{\beta}$ as a function of the sampled parameters by setting it equal to $\sqrt{\theta\sigma^2}\mathbf{V}\mathbf{D}^{-1/2}\boldsymbol{\gamma}$.

2.3.3 High Dimensional Data

Next we address how to fit these models with high-dimensional data, where $p > n$ and $\mathbf{X}'\mathbf{X}$ is not full rank. The restriction on $\boldsymbol{\beta} : \boldsymbol{\beta}'\mathbf{S}_X\boldsymbol{\beta} = \sigma^2\theta$, is no longer an ellipsoid, but an unbounded subspace in p -dimensions (e.g. parallel lines for $p = 2$, and an infinite cylinder for $p = 3$). We assume that the $\text{rank}(\mathbf{X}) = n$, and partition $\mathbf{D} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ and $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2)$. Note that \mathbf{D}_1 is the $n \times n$ diagonal matrix of positive eigenvalues, \mathbf{V}_1 is the

matrix of corresponding eigenvectors, and \mathbf{V}_2 is the matrix of the $p - n$ eigenvectors spanning the null space of \mathbf{X} . We define $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)' = (\mathbf{D}_1^{1/2}\mathbf{V}_1'\boldsymbol{\beta}, \mathbf{V}_2'\boldsymbol{\beta})'/\sqrt{\theta\sigma^2}$, so that $\boldsymbol{\gamma}$ is multivariate Normal with the constraint that $\boldsymbol{\gamma}_1'\boldsymbol{\gamma}_1 = 1$. Marginally, $\boldsymbol{\gamma}_1 = \mathbf{D}_1^{1/2}\mathbf{V}_1'\boldsymbol{\beta}/\sqrt{\theta\sigma^2}$ is defined on the $n - 1$ dimensional unit sphere, and has a Fisher-Bingham distribution just like the full rank case. However, the reverse transformation $\boldsymbol{\beta} = \sqrt{\theta\sigma^2}\mathbf{V}_1\mathbf{D}_1^{-1/2}\boldsymbol{\gamma}_1$ is defined on the lower $n - 1$ dimensional ellipsoid within the entire constrained space $\{\boldsymbol{\beta} : \boldsymbol{\beta}'\mathbf{S}_X\boldsymbol{\beta} = \sigma^2\theta\}$. For example, if $p = 3$ and $n = 2$, this is the slice of the 3-dimensional ellipsoid with the minimum L_2 -norm. The problem is that this lower dimensional ellipsoid will not favor the sparsity or local shrinkage encouraged by $\boldsymbol{\Lambda}$. That would be equivalent to principal components regression using the top n principal components. To allow for shrinkage of the original coefficients and not the principal components, we then sample $\boldsymbol{\gamma}_2|\boldsymbol{\gamma}_1$, which is multivariate Normal, and make the reverse transformation $\boldsymbol{\beta} = \sqrt{\theta\sigma^2}(\mathbf{V}_1\mathbf{D}_1^{-1/2}\boldsymbol{\gamma}_1 + \mathbf{V}_2\boldsymbol{\gamma}_2)$. Since \mathbf{V}_2 spans the null space of \mathbf{X} , $\boldsymbol{\beta}$ is still in the constrained region, but covers the full p -dimensional space.

2.3.4 Selecting Hyperparameters

Next we address how to elicit the parameters, a and b , and ν and μ for the local shrinkage model. In practice, the Beta prior on R^2 can be chosen empirically, for example if summaries of R^2 are available from similar experiments or collected in a meta analysis, and converted to the shape parameters a and b . Alternatively, the distribution of R^2 can be elicited directly from expert opinion by asking about the proportion of variation in the response that is expected to be explained by the predictors. As a reference, if the null model is correct and none of the predictors are useful, then the least-squares statistic, \widehat{R}^2 , has a $\text{Beta}(\frac{p}{2}, \frac{n-p-1}{2})$ distribution under $H_0 : R^2 = 0$ (Muirhead, 1982).

Without prior knowledge, a and b can be chosen based on the desired shrinkage properties of the prior. We use the distribution in (2.5) to understand how $p(\boldsymbol{\beta}|\sigma^2)$ is affected by these choices. Unlike global-local shrinkage priors, the prior distribution for $\boldsymbol{\beta}$ depends on \mathbf{X} , so the quadratic form $\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \equiv \|\boldsymbol{\beta}\|_{\mathbf{X}}^2$ is penalized instead of the individual β_j . The shape parameter a controls the prior around the origin (as $\|\boldsymbol{\beta}\|_{\mathbf{X}}^2$ goes to 0) while b controls the tail behavior (as $\|\boldsymbol{\beta}\|_{\mathbf{X}}^2$ goes to infinity). The prior $p(\boldsymbol{\beta}|\sigma^2)$ has an infinite spike around the origin for $a < p/2$, is a bounded unimodal density for $a = p/2$, and has no mass at the origin for $a > p/2$. When $b = 1/2$, the prior has multivariate Cauchy-like tails, and heavier tails for $b < 1/2$. For regularization we may choose $a < p/2$ and $b \leq 1/2$ to have sufficient mass near zero so as to shrink small coefficients but to also have heavier tails that are robust to larger coefficients, not unlike other sparse regression methods (e.g. Bhattacharya et al., 2015; Carvalho et al., 2010). Zhang et al. (2016) make similar recommendations and consider $a = p/n$ for the R2-D2 model. Because a and b only control global shrinkage of the vector $\boldsymbol{\beta}$ through \mathbf{X} , this motivates the need for the prior on the local variance components.

Similar to the implementation of Normal-Gamma priors (Griffin and Brown, 2010) we choose the shape parameter, ν , to favor sparser solutions. We elicit ν based on the expected number of non-zero coefficients, or more generally, the number of coefficients that account for most of the variation. Specifically, if we expect that there are $q < p$ true signals, ν is chosen such that median of $\left(\sum_{k=1}^q \lambda_{(k)} / \sum_{j=1}^p \lambda_j\right) = 1 - \varepsilon$, where $\lambda_{(1)} > \dots > \lambda_{(p)}$ are the order statistics, and ε is small (e.g. 0.05). That is, we choose ν such that the q largest variances account for most of the variation in $\boldsymbol{\beta}$. We can find such a ν by sampling repeatedly from a Gamma distribution on a grid of possible values for ν until this equality is met. Alternatively, we could take a fully Bayesian approach and put a prior on ν and μ . Like the Normal-Gamma priors, a Gamma distribution for

μ will be conjugate, but ν is not. As a default choice without prior knowledge, we may choose $q = 0.5p$ similar to variable selection priors that prefer $p/2$ predictors. For high-dimensional problems, we might also choose $q = n$, preferring models that estimate n non-zero coefficients.

Although the shape parameter primarily controls the relative sizes of the λ_j , the rate parameter, μ , affects model performance via the absolute sizes of the variances. The same q that was used to choose ν can also be used to choose μ . If we expect there are q important predictors, and therefore $p - q$ coefficients close to 0, then the distribution on the ellipsoid should be close to 0 at the axes for $p - q$ directions. We can choose μ such that the Normal distribution for β without the ellipsoid restriction is mostly contained inside the ellipsoid for $p - q$ directions. Similar to choosing ν , we can find this value quickly through simulations, and the details are given in the Appendix.

2.4 Simulations

We conduct a simulation study using two examples to compare the proposed approach with other Bayesian regression models. The first example considers scenarios with prior subjective knowledge, and the second considers sparse regression scenarios. Since estimation is the primary objective in many analyses, we compare models using the sum of squared errors (SSE) between the true coefficients and the estimates. This study also considers the impact of dimensionality and covariate correlation on model performance.

2.4.1 Subjective Examples

We first study the subjective approach, where some prior knowledge about the data generating process is known through the distribution of R^2 . Consider 200 data sets generated

from the linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I),$$

where the sample size is $n = 30$ with $p = 30$ predictors. The covariates, x_{ij} , are Normally distributed with mean 0, variance 1, and with correlation $\text{correlation}(x_{ij_1}, x_{ij_2}) = 0.5^{|j_1 - j_2|}$. The true coefficients, $\boldsymbol{\beta}_0$, are generated componentwise from a standard Normal distribution, and the error variance (σ) is 1, 3, 5, or 7. By varying σ , and therefore the true R^2 , this setup allows us to understand how subjective information about the true model fit affects performance. We obtained similar results from other distributions for $\boldsymbol{\beta}_0$, such as a t -distribution with 3 degrees of freedom and a Uniform distribution.

To simulate having prior knowledge, we choose a $\text{Beta}(a^*, b^*)$ distribution for R^2 under the data generation process described above. First, we sample from the true distribution of R^2 by simulating 10,000 data sets, and estimate a^* and b^* of the approximate Beta distribution by their method-of-moments estimators. We then select the hyperparameters ν and μ assuming that 50% of the predictors account for most of the coefficient variability, that is, $q = 15$. The approximate Beta distribution of R^2 represents the prior knowledge, and varies depending on σ . For example, R^2 is between 0.92 and 0.99 with 95% probability when $\sigma = 1$, and between 0.19 and 0.54 when $\sigma = 7$. Table 2.1 shows the 95% probability intervals for each $\text{Beta}(a^*, b^*)$ distribution (based on the the .025 and .975 quantiles). Note that this prior knowledge about R^2 is derived before fitting any models, and describes only the data generating process, not the true $\boldsymbol{\beta}_0$ for a given data set.

We also fit the R2-D2 prior (Zhang et al., 2016) and Bayesian Lasso prior (Park and Casella, 2008) using the same subjective knowledge. The R2-D2 prior uses the same choices for a^* and b^* , but the Dirichlet parameter for local shrinkage is automatically chosen to be $a_\pi = a^*/p$. The Bayesian Lasso is parameterized by the penalty parameter,

λ^2 , and must be chosen for each data set depending on \mathbf{X} . We choose λ^2 by simulating β on a grid of λ^2 values from the prior until the induced distribution of R^2 given \mathbf{X} best approximates a $\text{Beta}(a^*, b^*)$ distribution, according to the Kolmogorov-Smirnov test statistics. For comparison, we fit some possible “default” priors without subjective knowledge: the BEERS and R2-D2 prior with a Uniform prior on R^2 ($a = 1$ and $b = 1$), and the Horseshoe prior (Carvalho et al., 2010). Each model ran for 20,000 MCMC iterations, using the first 10,000 as a burn-in period, and we estimated β using the posterior means of the remaining 10,000. The models were fit using R code provided by the authors of Zhang et al. (2016) for the R2-D2, the `monomvn` package in R for the Bayesian Lasso, and the `bayesreg` package in R for the Horseshoe. Performance is assessed using the sum of squared errors (SSE):

$$SSE(\hat{\beta}) = (\hat{\beta} - \beta_0)' (\hat{\beta} - \beta_0) = \sum_{j=1}^p (\hat{\beta}_j - \beta_{0j})^2 \quad (2.8)$$

where β_0 are the true coefficients and $\hat{\beta}$ are the estimated coefficients. Table 2.1 shows the average sum of squared errors (ASSE) across the 200 data sets for each example, and standard errors are given in parentheses.

For all examples, the BEERS prior with subjective knowledge is preferred to the default options from the same model. Additionally, it is superior to the R2-D2 using the same subjective information. While the R2-D2 is very effective in sparse and high-dimensional problems, it performs poorly in these examples where the coefficients should not be shrunk to zero. The advantage of the subjective approach is more pronounced when the prior knowledge suggests that R^2 should be small, such as when σ is large. The subjective Bayesian Lasso is preferred to the proposed approach in these cases, but falls behind when R^2 is expected to be large.

Table 2.1: Average sum of squared errors (and standard errors), from 200 randomly generated data sets. For each data set the true coefficients are generated independently from a $N(0, 1)$ distribution, and the responses from a $N(\mathbf{x}'\boldsymbol{\beta}_0, \sigma^2)$ distribution, where $n = p = 30$.

σ	1	3	5	7
a^*	135.3	22.6	12.7	9.6
b^*	5.5	7.9	11.9	17.3
Beta(a^*, b^*) 95% Interval	(0.92, 0.99)	(0.57, 0.88)	(0.32, 0.71)	(0.19, 0.54)
BEERS(a^*, b^*)	8.7 (0.3)	17.4 (0.4)	22.4 (0.4)	25.0 (0.5)
BEERS(1, 1)	8.3 (0.3)	19.5 (0.6)	28.0 (0.9)	35.7 (1.3)
R2-D2(a^*, b^*)	13.9 (0.4)	24.9 (0.5)	28.6 (0.5)	29.8 (0.6)
R2-D2(1, 1)	20.5 (0.5)	24.2 (0.5)	26.4 (0.5)	27.8 (0.6)
BLasso(a^*, b^*)	11.5 (0.3)	17.0 (0.4)	20.4 (0.4)	22.9 (0.5)
Horseshoe	11.6 (0.4)	19.1 (0.4)	23.6 (0.5)	26.8 (0.6)

2.4.2 Sparse Examples

The following simulation study is patterned after those of Zhang et al. (2016) and Bondell and Reich (2012) to study the performance of regression models when the true coefficient vector is sparse. This reflects applications where only a few important predictors affect the response, with many weak or unimportant predictors. We consider 200 data sets generated from the linear model, each with 60 observations (n), and the number of predictors (p) is either 50, 100, or 500. The covariates are generated in the same manner as the previous simulation, but with varying correlation; x_{ij} , are Normally distributed with mean 0, variance 1, and with correlation(x_{ij_1}, x_{ij_2}) = $r^{|j_1-j_2|}$, where r is either 0.5 or 0.9. The true coefficient vector is $\boldsymbol{\beta}_0 = (\mathbf{0}_{10}, \mathbf{B1}, \mathbf{0}_{20}, \mathbf{B2}, \mathbf{0}_{p-40})$, where $\mathbf{B1}$ and $\mathbf{B2}$ are each a random vector of length 5, generated componentwise from a Uniform(0,1) distribution. This choice simulates scenarios with a few small signals, and are difficult to estimate with classical methods, particularly under high-correlation. To better understand the

effect of sparsity levels, we also consider an additional example for the $p = 500$ case with $\beta_0 = (\mathbf{0}_{10}, \mathbf{B1}, \mathbf{0}_{20}, \mathbf{B2}, \mathbf{0}_{20}, \mathbf{B3}, \mathbf{0}_{20}, \mathbf{B4}, \mathbf{0}_{p-90})$, where $\mathbf{B3}$ and $\mathbf{B4}$ are generated in the same manner as $\mathbf{B1}$ and $\mathbf{B2}$.

In addition to the subjective and default models used in Section 2.4.1, we consider the effect of local shrinkage parameters, ν and μ . We choose them as described in Section 2.3.4 based on expected sparsity levels of: 2%, 10%, and 20%. These choices are denoted as $\nu_{.02}, \nu_{.10}$ and $\nu_{.20}$ in the results. For the subjective BEERS prior they are chosen based on the true sparsity percentage. Without any subjective knowledge, we might assume approximately n of the p predictors account for most of the variation in high-dimensional models, and choose the parameters based on $(100*n/p)\%$ sparsity. Following the recommendations of Zhang et al. (2016) and the discussion in Section 3.4, we also consider $a = p/n$ and $b = 0.5$ as a possible default choice, shrinking estimates globally toward the origin, but still allowing for Cauchy-like tails.

Tables 2.2 and 2.3 show the average sum of squared errors (ASSE) for these sparse examples. With prior knowledge, the subjective R2-D2 and Bayesian Lasso priors are generally preferred. Having strong prior information about R^2 controls the size of the ellipsoid in the BEERS prior, and this restricts the ability to shrink coefficients to zero, especially under high-correlation. However, the proposed prior with default parameters gives better estimates than the subjective models and all other default models, including the Horseshoe and R2-D2. Choosing ν and μ based on the true sparsity level gives better predictions in the lower dimensional examples, but shrinks parameters too much when p is 500. These results suggest choosing ν and μ based on 10% to 20% of the coefficients being important for sparse problems, or based on n/p for high-dimensional examples. Although the Bayesian Lasso was superior for many of the subjective simulations, it cannot compete in these sparse examples which require the flexibility of a local shrinkage

prior.

Table 2.2: Average sum of squared errors (and standard errors), multiplied by 10 for readability, from 200 randomly generated data sets with $r = 0.5$.

	$p = 50$ $q = 10$	$p = 100$ $q = 10$	$p = 500$ $q = 10$	$p = 500$ $q = 20$
BEERS(a^*, b^*)	6.7 (0.2)	9.2 (0.3)	11.1 (0.4)	32.8 (1.0)
R2-D2(a^*, b^*)	5.9 (0.2)	8.3 (0.3)	19.0 (0.6)	56.2 (0.9)
BLasso(a^*, b^*)	6.0 (0.1)	11.2 (0.2)	26.7 (0.5)	53.3 (0.8)
BLasso	7.1 (0.2)	18.3 (0.4)	29.6 (0.6)	58.8 (0.8)
Horseshoe	6.4 (0.2)	8.0 (0.3)	13.5 (0.6)	39.3 (1.3)
BEERS(1, 1, $\nu_{0.02}$)	6.0 (0.2)	8.4 (0.3)	11.6 (0.5)	32.5 (1.0)
BEERS(1, 1, $\nu_{0.1}$)	5.7 (0.2)	6.2 (0.2)	9.8 (0.3)	28.2 (0.9)
BEERS(1, 1, $\nu_{0.2}$)	5.6 (0.2)	6.0 (0.2)	9.2 (0.3)	27.3 (0.8)
BEERS(1, 1, $\nu_{n/p}$)	–	6.5 (0.2)	9.6 (0.3)	27.7 (0.8)
R2-D2(1, 1)	6.3 (0.2)	7.4 (0.3)	13.5 (0.5)	48.9 (1.3)
BEERS($p/n, 0.5, \nu_{0.02}$)	6.0 (0.2)	9.2 (0.4)	11.8 (0.5)	32.8 (1.1)
BEERS($p/n, 0.5, \nu_{0.1}$)	5.7 (0.2)	6.2 (0.2)	9.9 (0.3)	28.0 (0.9)
BEERS($p/n, 0.5, \nu_{0.2}$)	5.5 (0.2)	6.0 (0.2)	9.3 (0.3)	27.0 (0.7)
BEERS($p/n, 0.5, \nu_{n/p}$)	–	6.5 (0.2)	9.7 (0.3)	27.6 (0.9)
R2-D2($p/n, 0.5$)	6.4 (0.2)	7.2 (0.3)	11.3 (0.5)	34.9 (1.1)

To better understand the relative performance of the models, Table 2.4 shows the average SSE partitioned into the errors from the non-zero coefficients and from those equal to zero. Aggressively shrinking individual coefficients, and correctly estimating the ones that are truly zero, is the strength of the R2-D2 prior. The ellipsoid restriction of the BEERS prior prevents the local variance components from shrinking the coefficients as fast as the independent priors. This is more pronounced when the covariates are highly correlated. But the R^2 prior can better estimate the true signals for a smaller SSE overall, and because of this distinction, we prefer the R2-D2 prior for variable selection, and the

Table 2.3: Average sum of squared errors (and standard errors), multiplied by 10 for readability, from 200 randomly generated data sets with $r = 0.9$.

	$p = 50$ $q = 10$	$p = 100$ $q = 10$	$p = 500$ $q = 10$	$p = 500$ $q = 20$
BEERS(a^*, b^*)	19.3 (0.7)	23.8 (0.7)	24.4 (0.9)	43.6 (1.4)
R2-D2(a^*, b^*)	11.3 (0.4)	13.1 (0.5)	22.1 (0.9)	35.0 (0.7)
BLasso(a^*, b^*)	10.1 (0.3)	12.9 (0.3)	22.9 (0.4)	46.2 (0.6)
BLasso	12.5 (0.4)	16.3 (0.3)	25.9 (0.5)	52.9 (0.7)
Horseshoe	18.6 (0.8)	21.7 (0.8)	34.3 (1.3)	74.3 (2.5)
BEERS(1, 1, $\nu_{0.02}$)	18.2 (0.7)	19.7 (0.7)	28.0 (0.9)	47.6 (1.4)
BEERS(1, 1, $\nu_{0.1}$)	15.6 (0.6)	15.0 (0.5)	22.6 (0.7)	40.1 (1.1)
BEERS(1, 1, $\nu_{0.2}$)	14.2 (0.6)	13.7 (0.5)	18.1 (0.5)	32.5 (0.8)
BEERS(1, 1, $\nu_{n/p}$)	–	10.5 (0.3)	21.8 (0.7)	38.1 (1.1)
R2-D2(1, 1)	18.6 (0.8)	20.4 (0.8)	25.7 (1.0)	56.2 (2.1)
BEERS($p/n, 0.5, \nu_{0.02}$)	18.6 (0.7)	20.2 (0.7)	30.2 (1.0)	54.3 (1.7)
BEERS($p/n, 0.5, \nu_{0.1}$)	15.7 (0.6)	15.1 (0.5)	24.3 (0.7)	43.6 (1.2)
BEERS($p/n, 0.5, \nu_{0.2}$)	14.3 (0.6)	13.7 (0.5)	19.4 (0.6)	34.9 (0.9)
BEERS($p/n, 0.5, \nu_{n/p}$)	–	10.6 (0.3)	23.2 (0.7)	41.6 (1.1)
R2-D2($p/n, 0.5$)	18.9 (0.8)	20.1 (0.8)	24.4 (0.9)	52.7 (1.8)

BEERS prior for estimation.

2.5 Real Data Examples

We study the predictive performance of the BEERS prior through a simulation study of real examples. These three data sets have many more parameters than observations, and have very different correlation structures. The Cereal data consists of starch content measurements from 15 observations with 145 infrared spectra measurements as predictors. The data is provided with the `chemometrics` R package. The Cookie data arises from an experiment testing the near-infrared (NIR) spectroscopy of biscuit dough in which the fat content is measured on 72 samples, with 700 NIR spectra measurements as predictors. The data was generated in the experiment by Osborne et al. (1984), and is available in the `ppls` R package. The Multidrug data are from a pharmacogenomic study investigating the relationship between the drug concentration (at which 50% growth is inhibited for a human cell line) and expression of the adenosine triphosphate binding cassette transporter (Szakács et al., 2004). The data consists of 853 drugs as predictors, 60 samples of human cell lines using the ABCA3 transporter as the response, and is available in the `mixOmics` R package. In the statistics literature, the Cereal and Multidrug data were both studied by Polson and Scott (2012) and Griffin and Brown (2013); and the Cookie data was studied by Brown et al. (2001) and Ghosh and Ghattas (2015). The Multidrug covariates are weakly correlated, the Cookie covariates are positively correlated, and the Cereal covariates are either positively or negatively correlated. Figure 2.2 shows histograms of all pairwise correlations for the data sets.

We randomly split each data set into a training and testing sets to evaluate the out-of-sample predictive performance. For each example, 75% of the observations were used

Table 2.4: Average sum of squared error (and standard errors) from 200 randomly generated data sets, split by the SSE due to the zero and non-zero coefficients. The covariates (x_{ij}) are generated from a $N(0, 1)$ distribution with correlation $\text{correlation}(x_{ij_1}, x_{ij_2}) = r^{|j_1 - j_2|}$, and q is the number of non-zero coefficients generated from a Uniform(0, 1) distribution.

r	p	Model	SSE ($\beta > 0$)	SSE ($\beta = 0$)	SSE Total
0.5	50	BEERS(1, 1, $\nu_{0.1}$)	5.4 (0.2)	0.4 (0.1)	5.7 (0.2)
		R2-D2(1,1)	5.9 (0.2)	0.4 (0.1)	6.3 (0.2)
		Horseshoe	5.0 (0.2)	1.4 (0.1)	6.4 (0.2)
	100	BEERS(1, 1, $\nu_{0.1}$)	5.1 (0.2)	1.0 (0.1)	6.2 (0.2)
		R2-D2(1,1)	6.9 (0.3)	0.5 (0.1)	7.4 (0.3)
		Horseshoe	6.1 (0.3)	1.9 (0.1)	8.0 (0.3)
	500 ($q = 10$)	BEERS(1, 1, $\nu_{0.1}$)	7.0 (0.3)	2.8 (0.1)	9.8 (0.3)
		R2-D2(1,1)	13.2 (0.5)	0.3 (0.1)	13.5 (0.5)
		Horseshoe	10.8 (0.5)	2.7 (0.1)	13.5 (0.6)
	500 ($q = 20$)	BEERS(1, 1, $\nu_{0.1}$)	23.7 (0.8)	4.5 (0.2)	28.2 (0.9)
		R2-D2(1,1)	48.3 (1.3)	0.6 (0.1)	48.9 (1.3)
		Horseshoe	34.1 (1.2)	5.3 (0.2)	39.3 (1.3)
0.9	50	BEERS(1, 1, $\nu_{0.1}$)	14.8 (0.6)	0.8 (0.1)	15.6 (0.6)
		R2-D2(1,1)	17.8 (0.8)	0.9 (0.1)	18.6 (0.8)
		Horseshoe	16.6 (0.7)	1.9 (0.1)	18.6 (0.8)
	100	BEERS(1, 1, $\nu_{0.1}$)	13.6 (0.5)	1.4 (0.1)	15.0 (0.5)
		R2-D2(1,1)	19.6 (0.8)	0.8 (0.1)	20.4 (0.8)
		Horseshoe	20.2 (0.8)	1.5 (0.1)	21.7 (0.8)
	500 ($q = 10$)	BEERS(1, 1, $\nu_{0.1}$)	18.5 (0.7)	4.1 (0.1)	22.6 (0.7)
		R2-D2(1,1)	25.2 (1.0)	0.5 (0.1)	25.7 (1.0)
		Horseshoe	32.5 (1.3)	1.8 (0.2)	34.3 (1.3)
	500 ($q = 20$)	BEERS(1, 1, $\nu_{0.1}$)	36.2 (1.1)	3.9 (0.1)	40.1 (1.1)
		R2-D2(1,1)	55.1 (2.1)	1.1 (0.1)	56.2 (2.1)
		Horseshoe	69.8 (2.4)	4.5 (0.3)	74.3 (2.5)

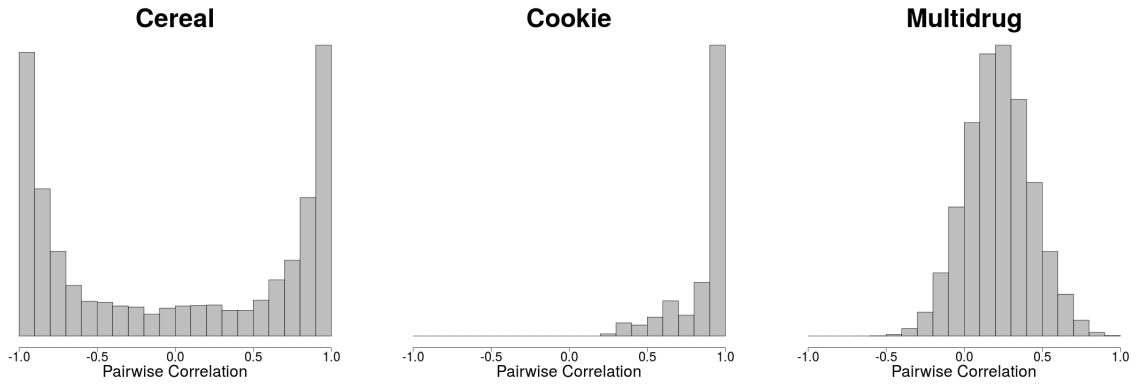


Figure 2.2: Histograms of all pairwise correlations of \mathbf{x}_j for the Cereal, Cookie and Multidrug data sets.

Table 2.5: Average root mean square error (RMSE) (and standard errors) for real data examples.

	Cereal	Cookie	Multidrug
n	15	72	60
p	145	700	853
BEERS($p/n, 0.5$)	20.8 (0.9)	1.72 (0.07)	14.9 (0.6)
BEERS(1, 1)	22.3 (0.7)	1.69 (0.07)	14.0 (0.5)
R2-D2($p/n, 0.5$)	26.7 (0.7)	2.10 (0.08)	15.4 (0.6)
R2-D2(1, 1)	22.3 (0.4)	2.07 (0.08)	11.5 (0.3)
Horseshoe	22.4 (0.8)	2.35 (0.09)	16.9 (0.8)
BLasso	30.5 (0.3)	3.20 (0.10)	15.6 (0.4)

for training, and the remaining 25% were used for estimating the square root of the mean squared error (RMSE) between the test sample and predictions. This process was repeated to create 200 data sets for each example. For the BEERS and R2-D2 priors, we considered a Uniform prior on R^2 , as well as the default choice of $a = p/n$ and $b = 0.5$ from Section 2.4. Because of the high-dimensionality, we assumed that n of the p predictors could be estimated from the sample, and therefore chose ν and μ assuming that $(100 * n/p)\%$ of the predictors account for most of the variability in the coefficients.

The average RMSE results are given in Table 2.5. The BEERS prior gives better out-of-sample predictions than the competing methods for the Cookie and Cereal examples. For the Multidrug example, it is preferred over the Horseshoe and Bayesian Lasso models, but one of the R2-D2 models is the best. Although the spectral data are highly correlated, the proposed approach works well because the important coefficients are not likely to be as sparse as the Horseshoe and R2-D2 estimates. Like the sparse simulation study, the BEERS prior uses the correlation structure to its advantage to better estimate the true signals at the expense of possibly overestimating the unimportant ones, yielding better predictions overall.

2.6 Discussion

We introduced a regression model using a prior on R^2 to easily elicit informative priors. The model was adapted for sparse and high-dimensional data sets by shrinking parameters both globally and locally. Selecting hyperparameters is a major problem and area of research for regularization models, typically relying on cross-validation, or assigning hyperpriors in the Bayesian framework. Our model performs well with default priors, and improves with minimal subjective knowledge. We demonstrated its effectiveness estimat-

ing regression parameters and making predictions through a simulation study and a real data examples.

A natural extension is to elicit priors based on summaries of the model fit for binary regression or other generalized linear models. The usual summaries for discrete responses do not have the convenient properties and interpretation of R^2 , and poses a new challenge. A possible approach to elicit priors analogous to R^2 in the linear model would be to use a pseudo- R^2 (e.g. Cox and Snell, 1989; McFadden, 1973; McKelvey and Zavoina, 1975). Although these measures are more complex, priors for generalized linear models are typically not conjugate, so this additional computational burden may be negligible.

Chapter 3

Binary Regression Using a Prior on the Model Fit

3.1 Introduction

In this chapter, we return to the problem addressed in Chapter 2: how to elicit informative priors based on a summary of the model fit. But this time for binary regression models. Although R^2 is a natural summary statistic for the linear model, there is not an equivalent measure for logistic and probit regression. There are several “pseudo- R^2 ” measures for binary regression that resemble the ordinary least squares (OLS) R^2 , but there is no clear consensus on which one is preferred (see for example Simonetti et al., 2017; Veall and Zimmermann, 1996).

McFadden’s pseudo- R^2 is one of the most popular, and measures the increase in the log-likelihood function of the full model relative to the intercept-only model (McFadden, 1973). Variations of pseudo- R^2 s using the likelihood or log-likelihood function have been considered in the literature (e.g. Aldrich and Nelson, 1984; Cox and Snell, 1989; Maddala,

1986). However, measures based on the likelihood become problematic when defining a prior distribution. The McKelvey and Zavoina (1975) pseudo- R^2 , which we denote as R_{MZ}^2 , is interpreted as the proportion of variance in the response explained by the model. Some have recommended the R_{MZ}^2 for routine use because it is better estimator for explained variance compared with other pseudo- R^2 measures, and closely resembles R^2 from OLS (DeMaris, 2002; Veall and Zimmermann, 1996). This formulation uses a latent variable representation for the binary regression model. As we will see in the next section, this representation is convenient for augmenting the tools developed for linear regression, and applying them to binary regression.

Many of the priors used for the coefficients in linear regression can also be used for logistic and probit regression. The logistic regression model lacks conjugate priors, and until recently, many of the algorithms for estimating posterior distributions either relied on approximations, Metropolis-Hastings algorithms, or point estimates of the mean. Polson et al. (2013) introduced a Gibbs sampler based on a data augmentation scheme using Pólya-Gamma random variables. This allowed for many prior distributions that can be represented as a mixture of Normals distribution to be easily used for logistic regression. Other commonly used priors include the class of global-local shrinkage priors, such as the Lasso (Park and Casella, 2008). Gelman et al. (2008) recommend independent Cauchy priors as default choice for routine applications, which give reasonable answers even under separation.

Eliciting informative priors for logistic regression is an active area of research with many applications. For example in ecology, researchers have developed elicitation programs to assist in quantifying expert knowledge (Choy et al., 2009; James et al., 2010). Hanson et al. (2014) introduce informative g -priors for logistic regression, which prescribe a fixed value for g , and a prior mean for the intercept, based on the individual success

probabilities following a particular $\text{Beta}(a_\pi, b_\pi)$ distribution. The g -prior is commonly used for linear regression, putting a Normal distribution on the regression coefficients, with covariance proportional to $(\mathbf{X}'\mathbf{X})^{-1}$, and the influence of the prior controlled by g . We propose eliciting informative priors based on the distribution of R_{MZ}^2 , and recommend default hyperparameters that perform well in various circumstances.

The remainder of the chapter is presented as follows. In Section 3.2 we describe how to induce a prior distribution for binary regression based on the distribution of R_{MZ}^2 . Following the approach we took for the linear model in Chapter 2, we develop a local shrinkage prior that can also flexibly shrink individual coefficients. We illustrate how to sample from the posteriors of the proposed distributions, building on the tools developed for the linear model. In Section 3.3, we present a simulation study to evaluate the performance of priors for binary regression models with subjective prior knowledge, and default priors for sparse regression models. We conclude the chapter with a brief discussion of the contributions, and opportunities for future research.

3.2 Model Specification

3.2.1 Prior Distribution on Pseudo- R^2

We begin by considering a latent variable formulation for the binary regression model:

$$y^* = \mathbf{x}'\boldsymbol{\beta} + \varepsilon \tag{3.1}$$

where y^* is the latent response, \mathbf{x} is the vector of covariates, $\boldsymbol{\beta}$ is the p -dimensional coefficient vector including an intercept, and ε is a random error component. The observed binary response is determined by the sign of y^* . If $y^* > 0$ then $y = 1$, and $y = 0$ otherwise.

If ε has a Logistic(0,1) distribution, then $P(y = 1) = \text{logit}^{-1}(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})/(1 + \exp(\mathbf{x}'\boldsymbol{\beta}))$, the logistic regression model. If ε has a Normal(0,1) distribution, then $P(y = 1) = \Phi(\mathbf{x}'\boldsymbol{\beta})$, the probit regression model, where $\Phi(\cdot)$ is the cumulative distribution function of a standard Normal distribution.

The McKelvey and Zavoina (1975) pseudo- R^2 , is defined as the proportion of variance of the latent variable, explained by the variance of the model:

$$R_{MZ}^2 = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta})}{\text{Var}(\mathbf{x}'\boldsymbol{\beta} + \varepsilon)} = \frac{\text{Var}(\mathbf{x}'\boldsymbol{\beta})}{\text{Var}(\mathbf{x}'\boldsymbol{\beta}) + \sigma^2} \quad (3.2)$$

where $\sigma^2 = \text{Var}(\varepsilon)$. Because the latent variable is not observed, the error variance of the assumed model is plugged in for σ^2 ($\pi^2/3$ for logistic regression, or 1 for probit regression). Similar to Chapter 2, if we write R_{MZ}^2 as a function of $\boldsymbol{\beta}$, and plug in the sample estimate for $\text{Var}(\mathbf{x}) \approx \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{S}_X/n \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{0} & \mathbf{S}_X \end{pmatrix}$, after centering the predictors, we get the following expression for R_{MZ}^2 :

$$R_{MZ}^2(\boldsymbol{\beta}) = \frac{\boldsymbol{\beta}'\text{Var}(\mathbf{x})\boldsymbol{\beta}}{\boldsymbol{\beta}'\text{Var}(\mathbf{x})\boldsymbol{\beta} + \sigma^2} \approx \frac{\boldsymbol{\beta}'_1\mathbf{S}_X\boldsymbol{\beta}_1}{\boldsymbol{\beta}'_1\mathbf{S}_X\boldsymbol{\beta}_1 + \sigma^2} \quad (3.3)$$

Note that a Beta prior on R_{MZ}^2 induces a prior on $\boldsymbol{\beta}_1$, the regression coefficients without the intercept, because the predictors are centered and the first diagonal element of $\text{Var}(\mathbf{x})$ corresponding to the intercept is assumed to be 0.

Eq. 3.3 defines $\boldsymbol{\beta}_1$ to lay on the surface of the ellipsoid $\{\boldsymbol{\beta}_1 : \boldsymbol{\beta}'_1\mathbf{S}_X\boldsymbol{\beta}_1 = \sigma^2 R_{MZ}^2/(1 - R_{MZ}^2)\}$, so we can let $\boldsymbol{\beta}_1$ be uniformly distributed on the ellipsoid. As we did in Chapter 2, we call this prior the Beta Ellipsoid R-Squared prior (BEERS), and is equivalent to the elliptical distribution in Proposition 2.1, with σ^2 replaced by a constant. That is, if $R_{MZ}^2 \sim$

Beta(a, b), then

$$p(\boldsymbol{\beta}_1) = \frac{\Gamma(p/2) |\mathbf{S}_X|^{1/2}}{B(a, b) \pi^{p/2}} (\sigma^2)^{-a} (\boldsymbol{\beta}'_1 \mathbf{S}_X \boldsymbol{\beta}_1)^{a-p/2} (1 + \boldsymbol{\beta}'_1 \mathbf{S}_X \boldsymbol{\beta}_1 / \sigma^2)^{-(a+b)}, \quad (3.4)$$

Unlike the prior for the linear model we cannot just center the predictors to remove the intercept. Because R^2_{MZ} doesn't give prior information about the intercept, β_0 , we simply place a diffuse Cauchy prior on the intercept as recommended for weakly informative priors by Gelman et al. (2008).

A Beta prior on R^2_{MZ} shrinks coefficients in a global sense by shrinking $\boldsymbol{\beta}'_1 \mathbf{X}' \mathbf{X} \boldsymbol{\beta}_1 / n$, but cannot flexibly shrink individual components. We take the same approach as in Chapter 2 and put a Normal-Gamma prior on $\boldsymbol{\beta}_1$, restricted to the surface of the ellipsoid. The prior is identical to that for the linear model, but again with σ^2 replaced by $\pi^2/3$ or 1. For simplicity, we make the transformation $\theta = R^2_{MZ} / (1 - R^2_{MZ})$. The full prior for local shrinkage BEERS priors is as follows.

$$\beta_0 \sim \text{Cauchy}(0, 10) \quad (3.5)$$

$$\boldsymbol{\beta}_1 | (\theta, \boldsymbol{\Lambda}) \sim N_{p-1}(\mathbf{0}, \sigma^2 \theta \boldsymbol{\Lambda}) \mathbb{1}\{\boldsymbol{\beta}'_1 \mathbf{S}_X \boldsymbol{\beta}_1 = \sigma^2 \theta\} \quad (3.6)$$

$$\lambda_j \sim \text{Gamma}(\nu, \mu), \text{ for } j = 1, \dots, p-1, \quad (3.7)$$

$$\theta \sim \text{Beta-Prime}(a, b) \quad (3.8)$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \dots, \lambda_{p-1}\}$ and $\mathbb{1}\{\cdot\}$ is the indicator function. Throughout this chapter we distinguish these two BEERS priors by calling them “uniform-on-ellipsoid” (UOE) and “local shrinkage” (LS) priors.

3.2.2 Posterior Computation

We can estimate the posterior distributions for both models using MCMC methods. The algorithms are very similar to those introduced in Chapter 3, so we provide the details in Appendix B.

Polson et al. (2013) show that in logistic regression, the likelihood for β can be equivalently represented as a Normal distribution conditional on n latent Pólya-Gamma (PG) responses. We follow their approach and introduce latent PG variables to sample from the posterior with a Gibbs sampler. Efficient sampling of PG random variables is provided in the R package `BayesLogit`. The samplers for the UOE and LS models are the same as those for the linear model, but include the extra step of sampling PG variables. In fact, the Metropolis-Hastings step in the LS model is simplified because only θ is updated since σ^2 is fixed.

The BEERS prior is easily implemented for the probit as well. We introduce the latent variables described in Eq. 3.1. Conditional on all other parameters, y^* has $N(\mathbf{x}'\beta, 1)$ distribution, truncated to be positive if $y = 1$, and negative otherwise. The samplers of Chapter 2 can be used with y^* as the “response variable” and σ^2 fixed at 1. For the remainder of the chapter we exclusively turn our attention to logistic regression, but note that a probit model can be used instead.

3.3 Simulations

3.3.1 Using Informative Priors

We conduct a simulation study to compare the proposed BEERS prior with alternative prior distributions for logistic regression. For the the BEERS prior, we consider informa-

tive priors that include prior knowledge about the distribution of R_{MZ}^2 , as well as default parameter choices. The first simulation is similar to the experiments designed by Hanson et al. (2014) to demonstrate the effectiveness of informative g -priors. We evaluate the models in terms of estimation of the true regression coefficients under varying degrees of correlation in the covariates, as well as different dimensions.

We generate data covariates from a standard Normal distribution with correlation $(x_{ij}, x_{ij'}) = r^{|j-j'|}$, where r is either 0, .5, or .9. The intercept is set to 1, and the true coefficients are generated independently and componentwise from a t_3 distribution, giving parameters that are centered around zero with some possibly large values. The dimension of $\boldsymbol{\beta}$ is either 3, 5 or 10, and the sample size is 100. The responses are generated from a Bernoulli distribution with the success probability, $\pi_i = \text{logit}^{-1}(\mathbf{x}'_i\boldsymbol{\beta}) = \exp(\mathbf{x}'_i\boldsymbol{\beta})/\exp(1+\mathbf{x}'_i\boldsymbol{\beta})$.

To simulate the scenario of having prior knowledge we use the true distribution of R_{MZ}^2 under this data generation process. We first estimate this distribution by generating a very large number of data sets (e.g 10,000) from the process described above. For each observation in a data set, we generate an error term from logistic distribution, and calculated the $R_{MZ}^2 = \text{Var}(\mathbf{x}'\boldsymbol{\beta}) / \text{Var}(\mathbf{x}'\boldsymbol{\beta} + \varepsilon)$. We approximate the distribution of R_{MZ}^2 with a Beta(a^*, b^*) distribution using the method-of-moments estimators on this large sample of R_{MZ}^2 . As a potential default prior, we choose a Uniform(0,1) distribution for the R_{MZ}^2 like we did for the linear model. For the local shrinkage BEERS priors, we also choose $\nu = \mu = 1$, equivalent to the Bayesian Lasso using an Exponential(1) mixing distribution for the variance.

We compared the proposed priors with other commonly used priors for logistic regression. Gelman et al. (2008) recommend independent Cauchy priors with scale parameter 10 for the intercept, and 2.5 for the other coefficients as a default, weakly informative

prior distribution. We also consider the Bayesian Lasso with variance parameter equal to 1 as a default shrinkage prior. Finally, we implement the g -prior using both an informative prior and a default Beta(0.5, 0.5) prior on the success probabilities as suggested by Hanson et al. (2014). The informative prior is “elicited” by generating many observations from the data generating process, and matching the true success probabilities to a Beta(a_π, b_π) distribution using the method-of-moments estimators.

We compare performance of the models using the root mean squared error (RMSE) between the true coefficients, and their posterior mean estimates from 10,000 MCMC iterations after a burn in period of 1,000 iterations. Each model was implemented with a Gibbs sampler, using the R packages `tgglm` for Cauchy priors, and `bayesreg` for the Bayesian Lasso. Tables 3.1-3.3 show the average RMSE from 200 randomly generated data sets for each p and r . For the examples considered here, the informative g -priors and BEERS priors do not have a clear advantage over their default counterparts. Since n is much larger than p , the rather diffuse prior information offers little help. These examples demonstrate however, that the default BEERS priors are preferred over the default Cauchy and Bayesian Lasso priors. Because of its flexibility over the g -priors, the local shrinkage BEERS prior is preferred in the larger p and high-correlation cases because of its flexibility to handle varying magnitudes of coefficients.

3.3.2 Sparse Regression Example

We ran an additional simulation to evaluate the performance of the default priors when the true regression coefficients are sparse. 200 data sets are generated in the same way as the first simulation, but $p = 20$, and $\beta = (1, \mathbf{0}_8, B_3, \mathbf{0}_8)'$, where B_3 is a vector of 3 independent t_3 distributed random variables. This mimics applications with a few

Table 3.1: Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 3$ and correlation $(x_{ij}, x_{ij'}) = r^{|j-j'|}$.

Model	$r = 0.0$	$r = 0.5$	$r = 0.9$
BEERS (UOE), default	3.2 (0.2)	3.5 (0.2)	5.6 (0.3)
BEERS (UOE), informative	3.2 (0.2)	3.3 (0.2)	5.5 (0.3)
BEERS (LS), default	4.0 (0.3)	4.0 (0.2)	5.2 (0.3)
BEERS (LS), informative	4.2 (0.3)	4.1 (0.2)	5.2 (0.3)
g -prior, default	3.3 (0.2)	3.4 (0.2)	5.6 (0.3)
g -prior, informative	3.4 (0.3)	3.5 (0.3)	5.4 (0.3)
Cauchy	3.7 (0.2)	4.1 (0.4)	6.0 (0.4)
B. Lasso	3.6 (0.2)	4.0 (0.4)	5.5 (0.4)

Table 3.2: Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 5$ and correlation $(x_{ij}, x_{ij'}) = r^{|j-j'|}$.

Model	$r = 0.0$	$r = 0.5$	$r = 0.9$
BEERS (UOE), default	4.1 (0.3)	4.3 (0.3)	7.0 (0.3)
BEERS (UOE), informative	4.1 (0.3)	4.2 (0.3)	7.0 (0.3)
BEERS (LS), default	5.6 (0.4)	5.7 (0.4)	7.4 (0.4)
BEERS (LS), informative	5.5 (0.4)	5.6 (0.4)	7.2 (0.4)
g -prior, default	4.7 (0.5)	4.9 (0.5)	7.6 (0.5)
g -prior, informative	4.7 (0.5)	5.0 (0.5)	7.6 (0.5)
Cauchy	4.4 (0.2)	6.9 (1.3)	7.9 (0.9)
B. Lasso	4.1 (0.2)	5.7 (0.7)	11.2 (4.2)

Table 3.3: Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients, multiplied by 10 for readability. The dimension is $p = 10$ and correlation $(x_{ij}, x_{ij'}) = r^{|j-j'|}$.

Model	$r = 0.0$	$r = 0.5$	$r = 0.9$
BEERS (UOE), default	5.3 (0.3)	5.9 (0.3)	11.1 (0.4)
BEERS (UOE), informative	5.4 (0.3)	6.0 (0.3)	11.6 (0.5)
BEERS (LS), default	6.6 (0.5)	6.8 (0.5)	9.0 (0.5)
BEERS (LS), informative	6.0 (0.5)	6.3 (0.5)	8.9 (0.4)
g -prior, default	6.2 (0.6)	6.5 (0.5)	10.4 (0.5)
g -prior, informative	5.6 (0.5)	6.3 (0.5)	11.5 (0.5)
Cauchy	10.7 (0.9)	12.2 (1.1)	14.3 (0.9)
B. Lasso	12.3 (4.0)	18.8 (5.8)	29.5 (11.5)

important predictors, but many weak or unimportant ones.

Table 3.4 displays the RMSE for each model and correlation setting. Although the g -prior is better than the uniform-on-ellipsoid BEERS prior, the local shrinkage BEERS prior estimates the true coefficients significantly better than the other models. The Bayesian Lasso shrinks parameters in a global sense, but lacks the flexibility for situations like this that have many small (or zero) coefficients and a few large.

Table 3.4: Average RMSE (and standard errors) between estimated coefficient by posteriors means and the true coefficients for the sparse regression example. All errors are multiplied by 10 for readability. The dimension is $p = 20$ and correlation $(x_{ij}, x_{ij'}) = r^{|j-j'|}$. Both BEERS put a Uniform distribution on R^2 .

Model	$r = 0.0$	$r = 0.5$	$r = 0.9$
BEERS (UOE)	4.01 (0.26)	4.38 (0.18)	9.61 (0.43)
BEERS (LS)	3.06 (0.13)	3.20 (0.12)	4.60 (0.18)
g -prior	3.58 (0.13)	4.19 (0.12)	9.10 (0.17)
Cauchy	16.92 (2.20)	17.59 (2.49)	15.24 (1.70)
B. Lasso	9.84 (4.22)	5.47 (1.07)	5.48 (0.99)

3.4 Discussion

We introduced a new prior for logistic and probit regression to easily include subjective prior knowledge into the analysis based on the distribution of a pseudo- R^2 . The simulation study showed that informative priors are useful for some situations. Even in the absence of prior knowledge the default priors work quite well compared to other recommended default priors, and the BEERS prior can compete with other shrinkage priors.

Future work should consider the applicability of the proposed priors to high-dimensional data. When the responses are completely separable in binary regression given the predictors, the parameters are not identifiable. Ghosh et al. (2017) showed that the posterior means exist under certain conditions for independent Cauchy priors, but not for multivariate Cauchy priors, which is a special case of the BEER (UOE) prior. Regularization methods can help with these issues, but it is still unknown whether the proposed local shrinkage priors will give sensible results with separation.

Since the McKelvey-Zavoina Pseudo- R^2 is only one of many choices for goodness-of-fit in binary regression, eliciting priors from other pseudo- R^2 measures should be considered. A possible approach would be to consider the squared correlation between the responses and the estimated probabilities, similar to how a prior for R^2 was derived for the linear model.

Chapter 4

An Efficient Bayesian Rank-Ordered Probit Model

4.1 Introduction

Collecting rank-ordered data through in surveys is an effective way to learn about individual preferences. More information is obtained from each observation, because the relative ordering of choices are given, as opposed to just the most preferred one. Statistical models for rank-ordered data have found applications in transportation studies (Beggs et al., 1981; Hausman and Ruud, 1987), marketing (Chapman and Staelin, 1982; Fok et al., 2012), and voter preferences (Koop and Poirier, 1994) to name a few. To understand the relationship between observed rankings and attributes of individuals, researchers often turn to the rank-ordered logit (ROL) and rank-ordered probit (ROP) models. While most applications consider only a few ranking choices (usually less than 10), these methods pose a challenge when considering many more choices.

In this chapter we consider the rank-ordered probit model, which can be viewed as a

generalization of the multinomial probit (MNP). Like the probit model for binary data, the MNP assumes a vector of underlying latent variables from a multivariate Normal distribution. These variables, also called random utilities, indicate the preferred choice among a set of choices through its largest element. Because the underlying distribution is Normal, a linear function with covariates can easily model individual preferences. The ROP extends this formulation by assuming the latent utilities determine the preferred ordering of the choices, as opposed to just the top choice.

Maximum likelihood estimation for the ROP requires evaluating a $p - 1$ dimensional integral, where p is the number of choices. Although numerical approximations are satisfactory when p is small, they break down for larger p , say more than 15 (Alvo and Philip, 2014). On the other hand, the latent variable formulation makes a Bayesian ROP model easy to implement with MCMC methods. McCulloch and Rossi (1994) first introduced a Gibbs sampler for the Bayesian MNP model, which was later improved by McCulloch et al. (2000) to place a prior distribution directly on the identifiable parameters. These methods were the basis for the marginal data augmentation methods used by Imai and van Dyk (2005) for the MNP, and implemented in the popular R package `MNP`. Imai et al. (2005) demonstrate how to extend the MNP for ordered preferences leading to the ROP. Yu (2000) and Yao and Böckenholt (1999) independently developed Bayesian ROP models, but their approaches are improved by the advances of Imai and van Dyke.

Rank-ordered data is often modeled using the rank-ordered logit model, which is an extension of the multinomial logit model (McFadden, 1973). Beggs et al. (1981) introduced the ROL model to analyze consumer demand for electric vehicles, and Koop and Poirier (1994) develop Bayesian methods to analyze voter preference data. The likelihood can be expressed as the product of MNL likelihoods, making maximum likelihood estimation for the ROL simpler than the for the ROP. However, MNL models suffer from the

“independence of irrelevant alternatives” property, which states that the relative preference between two choices does not depend on other available choices (see for example Alvo and Philip (2014)). This assumption is unrealistic in practice, and therefore the MNP is often favored over the MNL because it allows for correlations between preferences.

Although the MNP is preferred to the MNL model, researchers tend to use the logit model more often for rank-ordered data. Computationally, the ROP becomes more difficult to work with as p grows, because of the need to model covariance matrix of the latent utilities. We propose a new sampling method and ROP model to better handle data sets with large number of choices, improving both the speed and mixing of the MCMC samplers.

The rest of this chapter is presented as follows. In Section 4.2 we review the details of the rank-ordered probit model, and the sampling algorithms of Imai and van Dyk (2005). We then introduce an improved algorithm, taking advantage of advancements in sampling from truncated multivariate Normal distributions. Section 4.3 proposes a new ROP model based on the multiple shrinkage MNP of Burgette and Reiter (2013) that scales efficiently as the number of choices increase. Several simulated and real data examples illustrate the effectiveness of the new methods in Section 4.4. We conclude the chapter with a summary of the contributions, and a discussion of future research in this topic.

4.2 Bayesian Rank-Ordered Probit Model

4.2.1 Preliminaries

We consider the rank-ordered probit (ROP) model with p choices. Each observation, $\mathbf{y}_i = (y_{i1}, \dots, y_{ip})$, is a permutation of the integers $1, \dots, p$, and represents a ranking of items by judge i , for $i = 1, \dots, n$. For example, $y_{ij} = k$ specifies that the i -th judge ranks the j -th item in k -th place. We model \mathbf{y}_i through a latent multivariate Normal distribution:

$$\mathbf{u}_i \sim N_p(\mathbf{X}_i^0 \boldsymbol{\beta}^0, \mathbf{V}). \quad (4.1)$$

where the ordering of the elements of \mathbf{u}_i indicate the ranking in \mathbf{y}_i . That is, the ordering $\{u_{i,j_1} > u_{i,j_2} > \dots > u_{i,j_p}\}$ implies $\{y_{i,j_1} = 1, y_{i,j_2} = 2, \dots, y_{i,j_p} = p\}$. \mathbf{X}_i^0 is a matrix of covariates for judge i , $\boldsymbol{\beta}^0$ is a coefficient vector to be estimated, and \mathbf{V} is a positive-definite covariance matrix. If each individual has a set of q attributes, $\mathbf{x}_i = (1, x_{i,1}, \dots, x_{i,q-1})$, then $\mathbf{X}_i^0 = (\mathbf{I}_p, x_{i,1}\mathbf{I}_p, \dots, x_{i,q-1}\mathbf{I}_p)$, allowing for the average utilities of each item to depend on the attributes of the judge. Choice-specific covariates can be incorporated by appending columns to each \mathbf{X}_i^0 .

The parameters in Eq. 4.1 are not identifiable because location and scale shifts to \mathbf{u}_i will result in the same ranking. Following the approach of McCulloch and Rossi (1994), we specify a difference of random utilities model to account for the location shift. That is, we treat the last category as a “base” category, and comparing all the others to it by setting its coefficients to zero, and thus $u_{ip} = 0$. The latent variables become

$$\mathbf{w}_i = (\mathbf{I}_{p-1}, -\mathbf{1})\mathbf{u}_i \sim N_{p-1}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}), \quad (4.2)$$

where \mathbf{X}_i and $\boldsymbol{\beta}$ are the covariates and coefficients for the first $p - 1$ categories relative to the base category, and $\boldsymbol{\Sigma} = (\mathbf{I}_{p-1}, -\mathbf{1})\mathbf{V}(\mathbf{I}_{p-1}, -\mathbf{1})'$. The covariance, $\boldsymbol{\Sigma}$, must also be constrained to make it identifiable. Imai and van Dyk (2005) set the first diagonal element to 1, but Burgette and Nordheim (2012) show that this can negatively affect estimation depending on the base category. Following Burgette and Nordheim (2012), we impose the restriction that the trace of $\boldsymbol{\Sigma}$ be set to $p - 1$ for more flexibility. This is possible by multiplying to both side of Eq. 4.2 a parameter α , so that

$$\tilde{\mathbf{w}}_i = \alpha \mathbf{w}_i \sim N_{p-1}(\mathbf{X}_i \tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}), \quad (4.3)$$

where $\tilde{\boldsymbol{\Sigma}} = \alpha^2 \boldsymbol{\Sigma}$, and $\tilde{\boldsymbol{\beta}} = \alpha \boldsymbol{\beta}$. Introducing this “working parameter” α is the key to the marginal data augmentation algorithm described in the next section. We give $\boldsymbol{\beta}$ a $N_{q(p-1)}(\mathbf{0}, \mathbf{B})$ prior, and $\tilde{\boldsymbol{\Sigma}}$ an inverse-Wishart(ν, \mathbf{S}) distribution.

4.2.2 Marginal Data Augmentation Algorithm

Imai and van Dyk (2005) introduced marginal data augmentation (MDA) for the MNP model to improve the mixing of previous algorithms, and make it easier to sample from standard distributions without requiring a Metropolis step. The working parameter, α , makes sampling from the conditional posterior distributions easier, but conditioning on this unidentifiable parameter causes MCMC chains to converge slowly. The general idea of the MDA algorithm is to marginalize out the working parameter, giving a more diffuse posterior distribution than conditioning on it. Specifically, instead of repeatedly sampling from $p(\mathbf{W}|\boldsymbol{\Sigma}, \boldsymbol{\beta}, \alpha), p(\boldsymbol{\Sigma}|\mathbf{W}, \boldsymbol{\beta}, \alpha), p(\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{W}, \alpha)$ as usual in a Gibbs sampler, we sample $p(\mathbf{W}|\boldsymbol{\Sigma}, \boldsymbol{\beta}), p(\boldsymbol{\Sigma}|\mathbf{W}, \boldsymbol{\beta}), p(\boldsymbol{\beta}|\boldsymbol{\Sigma}, \mathbf{W})$, where $\mathbf{W} = (\mathbf{w}'_1, \dots, \mathbf{w}'_n)'$. This is possible at each step because we can sample each of the identifiable parameters jointly with α , achieving

the same marginal distribution.

The complete Gibbs sampler using MDA for the ROP is adapted from the algorithms of Imai and van Dyk (2005) and Burgette and Nordheim (2012):

- (a) Set initial values for $\mathbf{W}, \alpha^2, \boldsymbol{\beta}, \Sigma$.
- (b) Sample $(\mathbf{W}, \alpha^2 | \boldsymbol{\beta}, \Sigma, \mathbf{Y}) = (\mathbf{y}'_1, \dots, \mathbf{y}'_n)$
 - (i) Sample each $\mathbf{w}_i \sim N(\mathbf{X}_i \boldsymbol{\beta}, \Sigma)$ subject to the the ordering required by \mathbf{y}_i . This is a sample from a truncated multivariate Normal distribution (TMVN).
 - (ii) Sample $\alpha^2 | (\boldsymbol{\beta}, \Sigma, \mathbf{Y}) = \alpha^2 | \Sigma \sim tr(\mathbf{S} \Sigma^{-1}) / \chi^2_{\nu p}$.
 - (iii) Set $\tilde{\mathbf{W}} = \alpha \mathbf{W}$.
- (c) Sample $(\tilde{\boldsymbol{\beta}}, \alpha^2 | \Sigma, \mathbf{W})$.
 - (i) Sample, $\alpha^2 | (\mathbf{W}, \Sigma) \sim C / \chi^2_{(n+\nu)p}$,
where $C = \sum_{i=1}^n \left(\tilde{\mathbf{W}} - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right)' \Sigma^{-1} \left(\tilde{\mathbf{W}} - \mathbf{X}_i \hat{\boldsymbol{\beta}} \right) + \hat{\boldsymbol{\beta}}' \mathbf{B} \hat{\boldsymbol{\beta}} + tr(\mathbf{S} \Sigma^{-1})$, and
 $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \mathbf{X}_i' \Sigma^{-1} \mathbf{X}_i + \mathbf{B} \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i' \Sigma^{-1} \tilde{\mathbf{w}}_i \right)$.
 - (ii) Sample $\tilde{\boldsymbol{\beta}} \sim N(\hat{\boldsymbol{\beta}}, \alpha^2 \left(\left(\sum_{i=1}^n \mathbf{X}_i' \Sigma^{-1} \mathbf{X}_i + \mathbf{B} \right)^{-1} \right))$.
 - (iii) Set $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} / \alpha$.
- (d) Sample $(\Sigma, \alpha^2 | \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{W}}) = (\tilde{\Sigma} | \tilde{\boldsymbol{\beta}}, \tilde{\mathbf{W}})$.
 - (i) Sample $\tilde{\Sigma} \sim \text{inverse-Wishart} \left(n + \nu, \mathbf{S} + \sum_{i=1}^n \left(\tilde{\mathbf{w}}_i - \mathbf{X}_i \tilde{\boldsymbol{\beta}} \right) \left(\tilde{\mathbf{w}}_i - \mathbf{X}_i \tilde{\boldsymbol{\beta}} \right)' \right)$
- (e) Set $\alpha^2 = tr(\tilde{\Sigma}) / p$, $\Sigma = \tilde{\Sigma} / \alpha^2$, $\mathbf{W} = \tilde{\mathbf{W}} / \alpha$.

The difference between this proposed algorithm and the one for the MNP is sampling from the TMVN distribution in step (b)(i), and is discussed in the next section.

4.2.3 An Efficient MCMC Sampler

We now turn our attention to sampling the latent variables from the truncated multivariate Normal distribution in step (b)(i) of the MDA algorithm. Straightforward rejection sampling from the multivariate Normal distribution is infeasible when dealing with more than a few dimensions, since very few draws would be accepted. The more common approach, and the one implemented by Imai et al. (2005) in the MNP package, is to use Gibbs sampling to sample each variable conditional on the others $(w_{i,j}|\mathbf{w}_{i,-j})$ using standard formulas for conditional Normal distributions truncated to be between the next highest and lowest variables. However, if Σ exhibits strong correlation between $w_{i,j}$, then the MCMC sampler will move slowly through the posterior distribution. Additionally, at each iteration of the sampler the conditional covariance matrix of a multivariate Normal has to be updated, and will be quite cumbersome as the number of choices increases. We propose using a more efficient algorithm to sample from the TMVN distribution, improving posterior sampling under high-correlation with many choices.

Our approach to sampling is related to the work of Rodriguez-Yam et al. (2004) and Li and Ghosh (2015). Given a ranking, \mathbf{y}_i , we create a matrix of contrasts $\tilde{\mathbf{R}}$, such that $\mathbf{c} \leq \tilde{\mathbf{R}}\mathbf{w}_i \leq \mathbf{d}$, where the inequalities are defined elementwise. We first make the transformation $\mathbf{z}_i = \Sigma^{-1/2}(\mathbf{w}_i - \mathbf{X}_i\boldsymbol{\beta})$, so that z_{ij} are independent standard Normals, and we arrive at an updated truncation region for \mathbf{z}_i : $\mathbf{a} \leq \mathbf{R}\mathbf{z}_i \leq \mathbf{b}$, where $\mathbf{a} = \mathbf{c} - \tilde{\mathbf{R}}\mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{b} = \mathbf{d} - \tilde{\mathbf{R}}\mathbf{X}_i\boldsymbol{\beta}$, $\mathbf{R} = \Sigma^{1/2}\tilde{\mathbf{R}}$, and $\Sigma^{1/2}$ is the Cholesky decomposition. The main difficulty is determining the new bounds of the transformed variables, but Li and Ghosh (2015) introduce a simple algorithm to find them quickly. The Cholesky decomposition and its inverse only need to be calculated at each iteration of the sampler, offering additional computational efficiency over the standard method, which requires $p-1$ matrix inversions

to sample from the conditional Normal. This improved sampler for TMVN distributions replaces step (b)(i) in the algorithm of Section 4.2.2.

4.3 Multiple-Shrinkage Rank-Ordered Probit Model

A disadvantage of the ROP model described in Section 4.2 is its inability to scale as the number of choices becomes large. This is because, as p increases, the covariance parameters in Σ grow at rate of p^2 , making the ROP troublesome for even moderately sized p , e.g. $p > 10$. Additionally, each additional judge-specific covariate introduces an additional $p - 1$ regression parameters. To address this problem, we introduce a shrinkage ROP model which assumes the individual latent variables are conditionally independent, while still allowing for model flexibility via a truncated Dirichlet Process prior on the regression coefficients with Laplace kernels. In addition to the reducing the number of parameters to estimate, the proposed model benefits from the even more efficient sampling of the latent variables in the truncated MVN distribution.

Traditional shrinkage priors for regression are not suited for rank-ordered or multinomial data. Shrinking one parameter to zero has the unintended consequence of affecting the probability distributions of the other choices. Additionally, shrinking a parameter to zero does not mean that it is unimportant, only that it is the same as the base case. Instead of shrinking to zero, we propose the Multiple-Shrinkage ROP (MSROP) model, based on the MNP model (Burgette and Reiter, 2013). This shrinks parameters to small number of learned locations by finding groups of parameters that are indistinguishable from each other.

Consider the same data generation model described in Section 4.2, and for simplicity assume there are only judge-specific covariates. For each covariate, x_k there are $p - 1$

coefficients, $\boldsymbol{\beta}_k = (\beta_{k,1}, \dots, \beta_{k,p-1})$. We introduce additional parameters, $\boldsymbol{\mu}_k$, $\boldsymbol{\lambda}_k$, and $\boldsymbol{\tau}_k$, corresponding to the same coefficients in $\boldsymbol{\beta}_k$. The full model is parameterized as follows for $k = 1, \dots, q$; $j = 1, \dots, p - 1$; and $i = 1, \dots, n$.

$$D_{0k} = N(c_k, d_k) \times \text{Gamma}(a_k, b_k) \quad (4.4)$$

$$(\mu_{jk}, \tau_{jk}) \sim \text{truncated-DP}(\alpha, D_{0k}; p - 1) \quad (4.5)$$

$$\lambda_{jk} \sim \text{Exponential}(2/\tau_{jk}) \quad (4.6)$$

$$\beta_{jk} \sim N(\mu_{jk}, \lambda_{jk}) \quad (4.7)$$

$$\mathbf{w}_i \sim N_{p-1}(\mathbf{X}_i \boldsymbol{\beta}, \mathbf{I}_{p-1}) \mathbb{1} \{ \text{order}(\mathbf{w}_i) = \mathbf{y}_i \} \quad (4.8)$$

where D_{0k} is the base measure for the k -th covariate.

The variances for $\boldsymbol{\beta}_k$ are given Exponential distributions, inducing a Laplace distributions on $\boldsymbol{\beta}_k$, similar to the Bayesian lasso. We truncate the DP priors after $(p - 1)$ terms since β_{jk} can take on at most $p - 1$ unique values. However, we observe in the examples that the number of clusters is far less than this. We choose the same default hyperparameters for a_k, b_k, c_k, d_k and α , suggested by MacLehose and Dunson (2010) and again by Burgette and Reiter (2013). These choices compromise between having strong shrinkage, but still encouraging few clusters. Similar to the MNP, we implement the blocked Gibbs sampler by Ishwaran and James (2002) to sample from the posterior. The details of the MCMC algorithm follow directly from MacLehose and Dunson (2010) and Burgette and Reiter (2013), so we omit them here. Again, the efficient step to sample from the TMVN distribution is an important modification.

4.4 Examples

4.4.1 MCMC Sampling Comparison

We first investigate the performance of the MCMC sampler of the improved ROP model compared with the standard model by Imai and van Dyk (2005) through a simulation study with varying degrees of correlations and number of choices. A data set of voter preferences for political parties in from Japan is considered, which is publicly available in the MNP package Imai et al. (2005).

In the first simulation we choose the sample size to be $n = 400$, $p \in \{5, 10, 20\}$, and $r \in \{0, 0.5, 0.9\}$. The true coefficients, β_{jk} , and the covariates, x_{ij} , are generated independently from a $N(0,1)$ distribution. The covariance of latent utilities, Σ , is generated from an inverse-Wishart distribution with $p+10$ degrees of freedom, and mean covariance chosen to have compound symmetry equal to r . This generates latent variables with an average pairwise correlation of r . True latent variables are generated from the ROP model, and induce an ordering for each individual. We repeat the process 10 times for each setting. To compare the samplers for timing purposes, we implement both in R, so that the only difference between them is the method used to sample from the TMVN distribution. The results refer to the sampler of Imai and van Dyk (2005) as the ROP(UV), and the proposed method as ROP(MV) to reflect whether the latent variables are sampled from the univariate or multivariate truncated Normal distributions. Each sampler ran for 6,000 iterations, discarding the first 1,000 for a burn in period.

We evaluate the efficiency of each sampler using the effective sample size for each β_j . The effective sample size is an estimate of the number “independent” samples generated from a chain. Figures 4.1 through 4.3 compare the distributions of the effective sample size for chains each of length 5,000. Figures 4.4 through 4.6 make the same comparison,

but display the distribution of the effective sample rate (ESR), which we define as the effective sample size per minute. The boxplots show that the improved sampler is most effective when there is high-correlation among latent variables. Further, the new sampler produces more independent samples faster because the conditional covariance matrix does not need to be calculated for each p and at each step. Instead only the Cholesky decomposition of Σ and its inverse are required to generate draws from the TMVN.

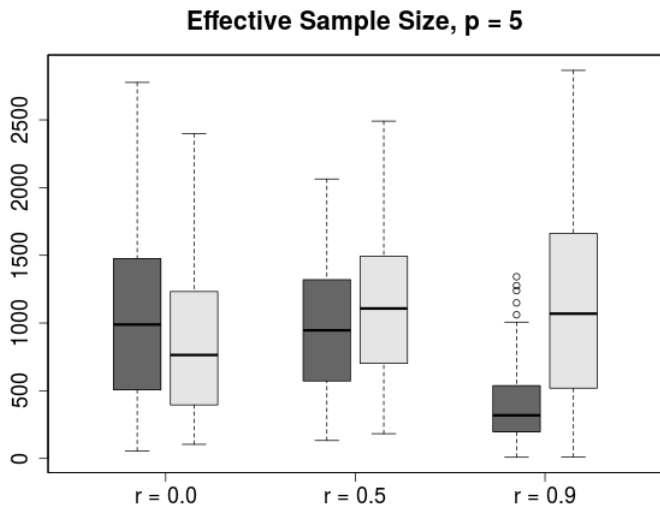


Figure 4.1: Distributions of effective sample size from 5,000 MCMC iterations for $p = 5$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

We also compare the sampling performance on survey data of voter choice preferences for political parties in Japan. The data includes 418 participants giving numerical preferences for political parties on a scale of 0-100, which are converted to rankings. There are three covariates for the judges: gender, education level, and age. Figure 4.7 displays the mean of the autocorrelation function across the 12 coefficients. This example shows

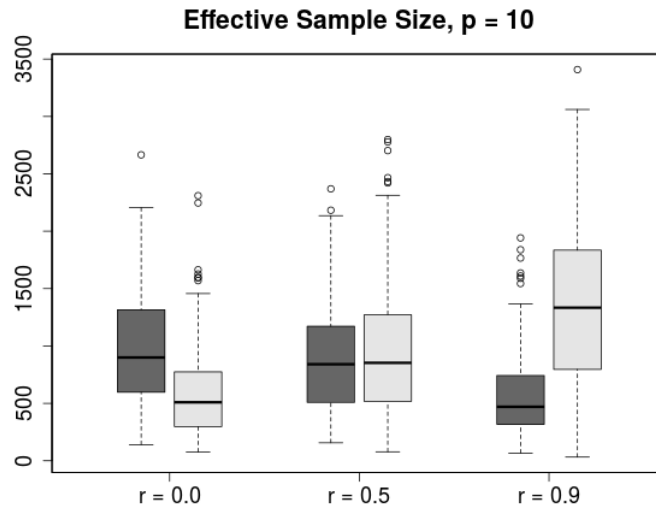


Figure 4.2: Distributions of effective sample size from 5,000 MCMC iterations for $p = 10$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

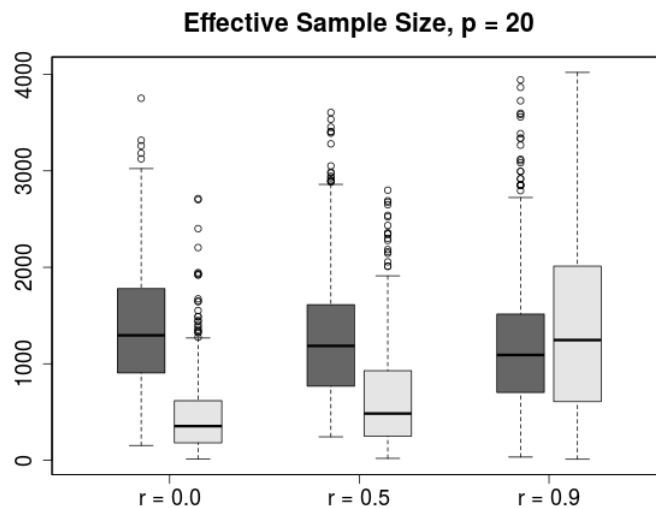


Figure 4.3: Distributions of effective sample size from 5,000 MCMC iterations for $p = 20$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

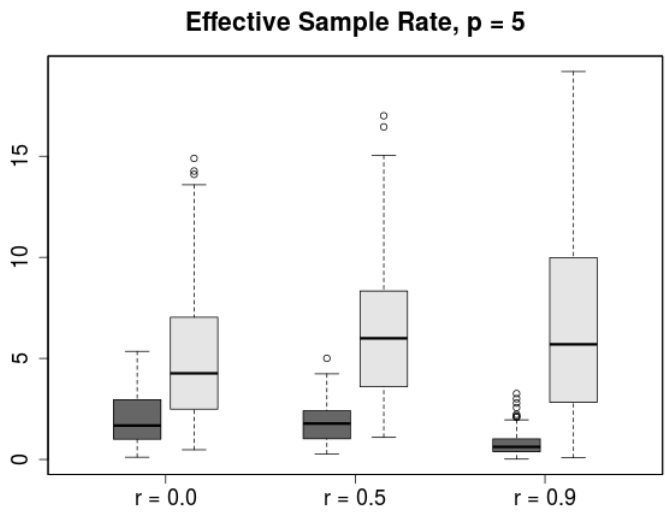


Figure 4.4: Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 5$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

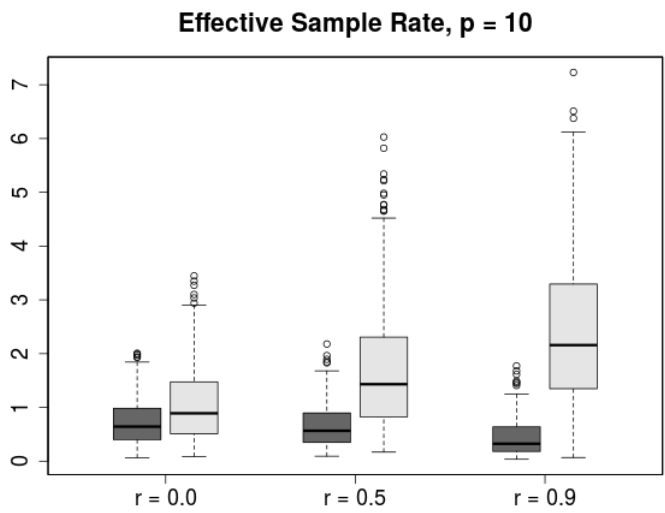


Figure 4.5: Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 10$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

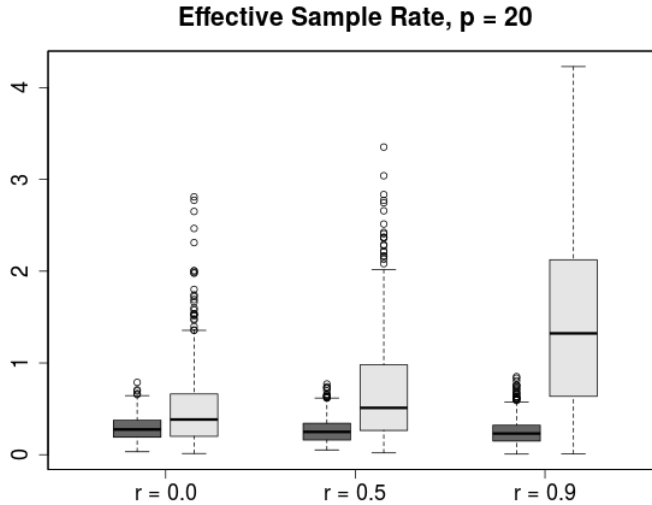


Figure 4.6: Distributions of effective sample rate (ESS per minute) from 5,000 MCMC iterations for $p = 20$. The current sampler, ROP(UV) is on the left (dark grey), and the proposed sampler, ROP(MV), is on the right (light grey).

substantial improvement in convergence for the new sampler. The effective samples sizes for each β_j were 4-8 times higher using the new sampling method. This example shows substantial improvement in convergence for the new sampler, likely because the latent utilities have strong correlations between them. The estimated posterior mean of Σ imply that the pairwise correlations are between 0.77 and 0.89.

4.4.2 Model Comparison

Next we compare the predictive performance of the MSROP with that of the traditional ROP model through a simulation study and real data examples with a large number of choices. The MSROP is easier to apply to situations with many choices, because it estimates far fewer parameters by assuming the latent variables are conditionally independent, and by also reducing the number of unique parameters due to the multiple

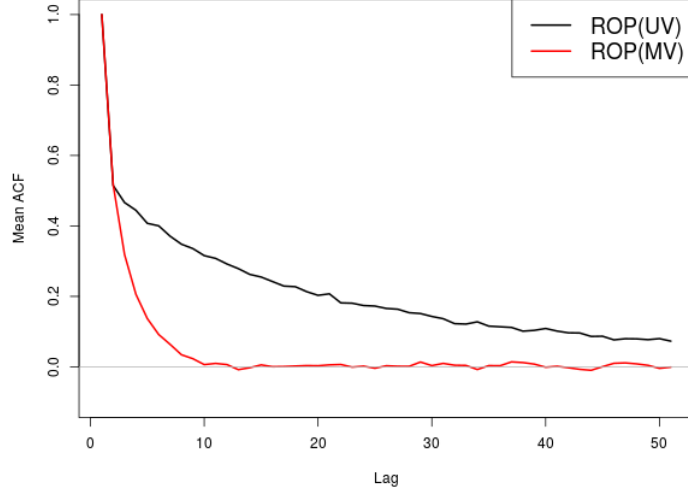


Figure 4.7: Mean autocorrelation functions (ACF) across samples of the 12 β_j in the Voters' Preference for Political Parties in Japan data.

shrinkage prior. The simulation study is identical to that of Section 4.4.1, except the correlation is fixed at $r = 0.5$. We also generate a test set of 100 observations from each true model, and repeat this for 50 total data sets. Posterior means of the parameters are calculated, and predictions are made for the test set from the estimated regression function.

We also consider out-of-sample predictive performance on several real data sets with many choices. Kamishima et al. (2010) collected survey data of Sushi preferences with many covariates from the judges and choices. The first data set includes rankings of 10 different sushi menu items from 5,000 participants in Japan. The second data set includes rankings from the same participants, but they were asked to rank 10 items which were randomly chosen from a larger set of 100 choices. Because of the size of this data set, we only include an intercept for each choice and a binary variable indicating whether the participant grew up in east or west Japan. Even though we include only two covariates,

these make β a 198-dimensional vector, and Σ a 198*198 covariance matrix for the larger data set. The data are subsequently referred to as the ‘‘Sushi10’’ and ‘‘Sushi100’’ data sets.

Performance is evaluated by comparing Kendall’s (tau) and Spearman’s (rho) rank correlation coefficients between the predicted and observed rankings. The rank correlations were calculated on the test set for the simulations. For the real data analyses, the entire set was randomly partitioned into 10 subsets, and each subset was used to train models and predict on one of the other subsets, similar to 10-fold cross validation. Tables 4.1 and 4.2 show the medians and IQRs of the rank correlation coefficients for the simulated and real data, respectively. Table 4.2 also shows the ESR for the real data to compare the speed of the different models.

Table 4.1: Median rank correlation coefficients (and IQRs) for the simulation study with $p \in \{5, 10, 20\}$.

p	Model	Kendall’s τ	Spearman’s ρ
5	ROP	0.694 (0.133)	0.779 (0.122)
	MSROP	0.694 (0.128)	0.781 (0.119)
10	ROP	0.696 (0.068)	0.823 (0.060)
	MSROP	0.696 (0.068)	0.823 (0.060)
20	ROP	0.698 (0.074)	0.847 (0.061)
	MSROP	0.698 (0.072)	0.847 (0.061)

The results from both the simulation and real data analysis show that there is no difference in predicted rankings between the ROP and MSROP models. Although the ROP is the true model for the simulations, the MSROP demonstrates that it can give the same results from a model that is much simpler as p become large. Additionally, the

Table 4.2: Median rank correlation coefficients (and IQRs), and the median effective sampling rate (ESS per minute) for the Sushi data sets.

Data	Model	Kendall's τ	Spearman's ρ	ESR
Sushi10	ROP	0.333 (0.36)	0.467 (0.41)	42.9
	MSROP	0.333 (0.36)	0.467 (0.40)	204.1
Sushi100	ROP	0.333 (0.36)	0.455 (0.44)	0.1
	MSROP	0.333 (0.36)	0.467 (0.44)	3.1

MSROP model not only gives the same rankings for the Sushi100 data, but does so much quicker than the traditional model.

4.5 Discussion

In this chapter, we introduced a new sampling method for the ROP model that improves convergence of the sampler under high-correlation, and faster sampling in general. A multiple shrinkage version of the ROP model proved to be tractable and efficient for large number of choice categories. Their usefulness and efficiency was demonstrated through both simulated and real data examples.

Burgette and Hahn (2010) show that the base category can influence the estimated distributions, and propose a Symmetric MNP model to address the issue. This is likely to be worse in large p cases where little information may be available for the base category. We could instead modify the TMVN algorithm to allow for a sum-to-zero constraints on the latent variables. Sampling β from a Normal distribution subject to sum-to-zero constraints would be needed to make the model identifiable, but Burgette and Reiter (2013) note that in practice the sum-to-zero constraints on w_i almost translate to the same constraints on β . Future research should demonstrate that the multiple shrinkage

prior is not only faster, but can give better predictions than the traditional model in high-dimensional settings.

REFERENCES

- John H Aldrich and Forrest D Nelson. *Linear probability, logit, and probit models*, volume 45. Sage, 1984.
- Mayer Alvo and LH Philip. *Statistical methods for ranking data*. Springer, 2014.
- Steven Beggs, Scott Cardell, and Jerry Hausman. Assessing the potential demand for electric cars. *Journal of econometrics*, 17(1):1–19, 1981.
- Anirban Bhattacharya, Debdeep Pati, Natesh S Pillai, and David B Dunson. Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.
- Christopher Bingham. An antipodally symmetric distribution on the sphere. *The Annals of Statistics*, pages 1201–1225, 1974.
- Howard D Bondell and Brian J Reich. Consistent high-dimensional bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association*, 107(500):1610–1624, 2012.
- Philip J Brown, T Fearn, and Marina Vannucci. Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398–408, 2001.
- Lane F Burgette and P Richard Hahn. Symmetric bayesian multinomial probit models. *Duke University Statistical Science Technical Report*, pages 1–20, 2010.
- Lane F Burgette and Erik V Nordheim. The trace restriction: An alternative identification

- strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410, 2012.
- Lane F Burgette and Jerome P Reiter. Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian analysis (Online)*, 8(2), 2013.
- Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, page asq017, 2010.
- Randall G Chapman and Richard Staelin. Exploiting rank ordered choice set data within the stochastic utility model. *Journal of marketing research*, pages 288–301, 1982.
- Samantha Low Choy, Rebecca O’Leary, and Kerrie Mengersen. Elicitation by design in ecology: using expert opinion to inform priors for bayesian statistical models. *Ecology*, 90(1):265–277, 2009.
- David Roxbee Cox and E Joyce Snell. *Analysis of binary data*, volume 32. CRC Press, 1989.
- Wen Cui and Edward I George. Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.
- Alfred DeMaris. Explained variance in logistic regression: A monte carlo study of proposed measures. *Sociological Methods & Research*, 31(1):27–74, 2002.
- Robert Denham and Kerrie Mengersen. Geographically assisted elicitation of expert opinion for regression models. *Bayesian Analysis*, 2(1):99–135, 2007.
- Christopher J Fallaize and Theodore Kypraios. Exact bayesian inference for the bingham distribution. *Statistics and Computing*, 26(1-2):349–360, 2016.

- Dennis Fok, Richard Paap, and Bram Van Dijk. A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of applied econometrics*, 27(5): 831–846, 2012.
- Paul H Garthwaite and James M Dickey. Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 462–474, 1988.
- Paul H Garthwaite and James M Dickey. Elicitation of prior distributions for variable-selection problems in regression. *The Annals of Statistics*, pages 1697–1719, 1992.
- Paul H Garthwaite, Joseph B Kadane, and Anthony O’Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–701, 2005.
- Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, pages 1360–1383, 2008.
- Joyee Ghosh and Andrew E Ghattas. Bayesian variable selection under collinearity. *The American Statistician*, 69(3):165–173, 2015.
- Joyee Ghosh, Yingbo Li, Robin Mitra, et al. On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, 2017.
- Jeff Gill and Lee D Walker. Elicited priors for bayesian model specifications in political science research. *The Journal of Politics*, 67(3):841–872, 2005.
- Michael Goldstein. Subjective bayesian analysis: principles and practice. *Bayesian Analysis*, 1(3):403–420, 2006.

- Jim E Griffin and Philip J Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.
- Jim E Griffin and Philip J Brown. Some priors for sparse regression modelling. *Bayesian Analysis*, 8(3):691–702, 2013.
- Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.
- Timothy E Hanson, Adam J Branscum, Wesley O Johnson, et al. Informative g -priors for logistic regression. *Bayesian Analysis*, 9(3):597–612, 2014.
- Jerry A Hausman and Paul A Ruud. Specifying and testing econometric models for rank-ordered data. *Journal of econometrics*, 34(1-2):83–104, 1987.
- Kosuke Imai and David A van Dyk. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2):311–334, 2005.
- Kosuke Imai, David A Van Dyk, et al. Mnp: R package for fitting the multinomial probit model. *Journal of Statistical Software*, 14(3):1–32, 2005.
- Hemant Ishwaran and Lancelot F James. Approximate dirichlet process computing in finite normal mixtures: smoothing and prior information. *Journal of computational and graphical statistics*, 11(3):508–532, 2002.
- Allan James, Samantha Low Choy, and Kerrie Mengersen. Elicitor: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25(1):129–145, 2010.

- Sindhu R Johnson, George A Tomlinson, Gillian A Hawker, John T Granton, and Brian M Feldman. Methods to elicit beliefs for bayesian priors: a systematic review. *Journal of clinical epidemiology*, 63(4):355–369, 2010.
- Joseph Kadane and Lara J Wolfson. Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(1):3–19, 1998.
- Joseph B Kadane, James M Dickey, Robert L Winkler, Wayne S Smith, and Stephen C Peters. Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, 75(372):845–854, 1980.
- Toshihiro Kamishima, Hideto Kazawa, and Shotaro Akaho. A survey and empirical comparison of object ranking methods. In *Preference learning*, pages 181–201. Springer, 2010.
- John T Kent. The fisher-bingham distribution on the sphere. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 71–80, 1982.
- John T Kent, Asaad M Ganeiber, and Kanti V Mardia. A new method to simulate the bingham and related distributions in directional data analysis with applications. *arXiv preprint arXiv:1310.8110*, 2013.
- Gary Koop and DJ Poirier. Rank-ordered logit models: An empirical analysis of ontario voter preferences. *Journal of Applied Econometrics*, 9(4):369–388, 1994.
- Petra M Kuhnert, Tara G Martin, and Shane P Griffiths. A guide to eliciting and using expert knowledge in bayesian ecological models. *Ecology letters*, 13(7):900–914, 2010.
- Minjung Kyung, Jeff Gill, Malay Ghosh, George Casella, et al. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

- Yifang Li and Sujit K Ghosh. Efficient sampling methods for truncated multivariate normal and student-t distributions subject to linear inequality constraints. *Journal of Statistical Theory and Practice*, 9(4):712–732, 2015.
- Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008. doi: 10.1198/016214507000001337.
- Richard F MacLehose and David B Dunson. Bayesian semiparametric multiple shrinkage. *Biometrics*, 66(2):455–462, 2010.
- Gangadharrao S Maddala. *Limited-dependent and qualitative variables in econometrics*. Number 3. Cambridge university press, 1986.
- Robert McCulloch and Peter E Rossi. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240, 1994.
- Robert E McCulloch, Nicholas G Polson, and Peter E Rossi. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193, 2000.
- D McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1973.
- Richard D McKelvey and William Zavoina. A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, 4(1):103–120, 1975.
- Robb J Muirhead. *Aspects of multivariate statistical theory*. John Wiley & Sons, 1982.

- Iain Murray, Zoubin Ghahramani, and David MacKay. Mcmc for doubly-intractable distributions. In *Proceedings of the 22nd annual conference on uncertainty in artificial intelligence*, pages 359–366. AUAI Press, 2006.
- Brian G Osborne, Thomas Fearn, Andrew R Miller, and Stuart Douglas. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35(1):99–105, 1984.
- Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.
- Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.
- Nicholas G Polson and James G Scott. Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012.
- Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.
- Gabriel Rodriguez-Yam, Richard A Davis, and Louis L Scharf. Efficient gibbs sampling of truncated multivariate normal with application to constrained linear regression. *Technical Report*, 2004.
- Steven L Scott and Hal R Varian. Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23, 2014.

- Biagio Simonetti, Pasquale Sarnacchiaro, and M Rosario González Rodríguez. Goodness of fit measures for logistic regression model: an application for students evaluations of university teaching. *Quality & Quantity*, pages 1–10, 2017.
- Gergely Szakács, Jean-Philippe Annereau, Samir Lababidi, Uma Shankavaram, Angela Arciello, Kimberly J Bussey, William Reinhold, Yanping Guo, Gary D Kruh, Mark Reimers, et al. Predicting drug sensitivity and resistance: profiling abc transporter genes in cancer cells. *Cancer cell*, 6(2):129–137, 2004.
- Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- Michael R Veall and Klaus F Zimmermann. Pseudo-r² measures for some common limited dependent variable models. *Journal of Economic surveys*, 10(3):241–259, 1996.
- Grace Yao and Ulf Böckenholt. Bayesian estimation of thurstonian ranking models based on the gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52(1):79–92, 1999.
- Philip LH Yu. Bayesian analysis of order-statistics models for ranking data. *Psychometrika*, 65(3):281–299, 2000.
- Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243, 1986.
- Arnold Zellner and Aloysius Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de estadística y de investigación operativa*, 31(1):585–603, 1980.

Yan Zhang and Howard D Bondell. High-dimensional variable selection via penalized credible regions with global-local shrinkage priors. *arXiv preprint arXiv:1602.01160*, 2016.

Yan Zhang, Brian J Reich, and Howard D Bondell. High dimensional linear regression via the r2-d2 shrinkage prior. *arXiv preprint arXiv:1609.00046v2*, 2016.

APPENDICES

Appendix A

Chapter 2 Proofs

A.1 Proposition 2.1

Derivation of Equation (2.5): Let γ be uniformly distributed on the $p - 1$ dimensional unit sphere. That is,

$$p(\gamma) = \frac{\Gamma(p/2)}{2\pi^{p/2}} \mathbf{1}\{\gamma'\gamma = 1\}.$$

Define $\theta = \frac{R^2}{1-R^2} \sim \text{beta-prime}(a, b)$. Make the transformation $r = \sqrt{\theta}$:

$$\begin{aligned} p(\gamma, \theta) &= p(\gamma)p(\theta) = \frac{\Gamma(p/2)}{2\pi^{p/2}B(a, b)} \theta^{a-1} (1 + \theta)^{-a-b} \mathbf{1}\{\gamma'\gamma = 1\} \\ p(\gamma, r) &= \frac{\Gamma(p/2)}{2\pi^{p/2}B(a, b)} r^{2a-2} (1 + r^2)^{-a-b} \mathbf{1}\{\gamma'\gamma = 1\} |2r| \\ &= \frac{\Gamma(p/2)}{\pi^{p/2}B(a, b)} r^{2a-1} (1 + r^2)^{-a-b} \mathbf{1}\{\gamma'\gamma = 1\} \end{aligned}$$

Make the transformation $\mathbf{z} = r\gamma$. The Jacobian is r^{p-1} when decomposing a vector (supported on \mathbb{R}^p) to a radius (supported in \mathbb{R}_+) and a unit direction (supported on the

$p - 1$ unit sphere, \mathbb{S}_{p-1}), so the reciprocal Jacobian is $|(\mathbf{z}'\mathbf{z})^{\frac{p-1}{2}}|^{-1} = (\mathbf{z}'\mathbf{z})^{\frac{1-p}{2}}$. Thus,

$$\begin{aligned} p(\mathbf{z}) &= \frac{\Gamma(p/2)}{\pi^{p/2}B(a, b)} (\mathbf{z}'\mathbf{z})^{a-1/2} (1 + \mathbf{z}'\mathbf{z})^{-a-b} \times (\mathbf{z}'\mathbf{z})^{\frac{1-p}{2}} \\ &= \frac{\Gamma(p/2)}{\pi^{p/2}B(a, b)} (\mathbf{z}'\mathbf{z})^{a-p/2} (1 + \mathbf{z}'\mathbf{z})^{-a-b} \end{aligned}$$

Finally, make the transformation $\boldsymbol{\beta} = \mathbf{V}\mathbf{D}^{-1/2}\mathbf{z}\sigma$, so that $\mathbf{z}'\mathbf{z} = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(\sigma^2n)$, where $\mathbf{V}\mathbf{D}\mathbf{V}'$ is the eigendecomposition of $\mathbf{X}'\mathbf{X}/n = \boldsymbol{\Sigma}_X$:

$$\begin{aligned} p(\boldsymbol{\beta}|\sigma^2) &= \frac{\Gamma(p/2)}{\pi^{p/2}B(a, b)} (\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(\sigma^2n))^{a-p/2} (1 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}/(\sigma^2n))^{-a-b} |\mathbf{V}\mathbf{D}^{-1/2}\sigma|^{-1} \\ &= \frac{\Gamma(p/2) |\boldsymbol{\Sigma}_X|^{1/2}}{B(a, b) \pi^{p/2}} (\sigma^2)^{-a} (\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta})^{a-p/2} (1 + \boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}/\sigma^2)^{-(a+b)}. \end{aligned}$$

A.2 Proposition 2.2

Mixture of normals representation for $a \leq p/2$.

$$\text{Let } \boldsymbol{\beta} | z, w, \boldsymbol{\Sigma} \sim N\left(0, zw\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right),$$

$$z \sim \text{inverse-gamma}(b, n/2), \text{ and}$$

$$w \sim \text{beta}(a, p/2 - a).$$

Define $\theta = \frac{1-w}{w}$, so $\theta \sim \text{beta-prime}(p/2 - a, a)$, and to simplify notation let $\boldsymbol{\Sigma} =$

$\mathbf{X}'\mathbf{X}/\sigma^2$.

$$\begin{aligned}
p(\boldsymbol{\beta}, \theta, z|\boldsymbol{\Sigma}) &= (2\pi)^{-p/2} z^{-p/2} (1 + \theta)^{p/2} |\boldsymbol{\Sigma}|^{1/2} \exp \left\{ -\frac{1 + \theta}{2z} \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} \right\} \\
&\quad \times \frac{n^b 2^{-b}}{\Gamma(b)} z^{-b-1} \exp \{-n/2z\} \\
&\quad \times \frac{\Gamma(p/2)}{\Gamma(a)\Gamma(p/2 - a)} \theta^{p/2-a-1} (1 + \theta)^{-p/2} \\
&= \frac{2^{-p/2-b} n^b |\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} \Gamma(a)\Gamma(b)\Gamma(p/2 - a)} z^{-p/2-b-1} \theta^{p/2-a-1} \exp \left\{ -\frac{n + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta} + \theta \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2z} \right\}
\end{aligned}$$

$$\begin{aligned}
p(\boldsymbol{\beta}, z|\boldsymbol{\Sigma}) &= \frac{2^{-p/2-b} n^b |\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} \Gamma(a)\Gamma(b)\Gamma(p/2 - a)} z^{-p/2-b-1} \exp \left\{ -\frac{n + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2z} \right\} \int \theta^{p/2-a-1} \exp \left\{ -\theta \frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2z} \right\} d\theta \\
&= \frac{2^{-a-b} n^b |\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} \Gamma(a)\Gamma(b)} z^{-a-b-1} \exp \left\{ -\frac{n + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2z} \right\} (\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})^{a-p/2}
\end{aligned}$$

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\Sigma}) &= \frac{2^{-a-b} n^b |\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} \Gamma(a)\Gamma(b)} (\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})^{a-p/2} \int z^{-a-b-1} \exp \left\{ -\frac{n + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}}{2z} \right\} dz \\
&= \frac{2^{-a-b} n^b |\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} \Gamma(a)\Gamma(b)} (\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})^{a-p/2} (n + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})^{-a-b} 2^{a+b} \Gamma(a + b) \\
&= \frac{|n\boldsymbol{\Sigma}|^{1/2} \Gamma(p/2)}{\pi^{p/2} B(a, b)} (\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}/n)^{a-p/2} (1 + \boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta}/n)^{-(a+b)} \\
&= \frac{\Gamma(p/2) |\boldsymbol{\Sigma}_X|^{1/2}}{B(a, b) \pi^{p/2}} (\sigma^2)^{-a} (\boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta})^{a-p/2} (1 + \boldsymbol{\beta}'\boldsymbol{\Sigma}_X\boldsymbol{\beta}/\sigma^2)^{-(a+b)},
\end{aligned}$$

where $\boldsymbol{\Sigma}_X = \mathbf{X}'\mathbf{X}/n$.

A.3 Proposition 2.3

If we let $\mathbf{\Lambda}^{-1} = \mathbf{X}'\mathbf{X} = n\mathbf{V}\mathbf{D}\mathbf{V}'$, then

$$p(\boldsymbol{\gamma}) = C_{X,\Lambda} \exp\{-\boldsymbol{\gamma}'\mathbf{D}^{-1/2}\mathbf{V}'\mathbf{V}\mathbf{D}\mathbf{V}'\mathbf{V}\mathbf{D}^{-1/2}\boldsymbol{\gamma}/2\} \mathbb{1}\{\boldsymbol{\gamma}'\boldsymbol{\gamma} = 1\} = \frac{1}{C} \mathbb{1}\{\boldsymbol{\gamma}'\boldsymbol{\gamma} = 1\}$$

where the constant, $C = \int \mathbb{1}\{\boldsymbol{\gamma}'\boldsymbol{\gamma} = 1\} d\boldsymbol{\gamma} = \frac{2\pi^{p/2}}{\Gamma(p/2)}$, the surface area of the a $p - 1$ dimensional unit sphere. The rest of the proof is identical to that of Proposition 1 deriving the distribution in (2.5). It is clear that $\boldsymbol{\beta}$ is uniform given the ellipsoid, because it's an elliptical distribution, but it also comes from putting a uniform distribution on $\boldsymbol{\gamma}$.

A.4 Choosing Hyperparameters ν and μ

Assume that q of the p variance components (λ_j) account for $(1 - \varepsilon)$ proportion of the variability. That is,

$$\frac{\sum_{j=1}^q \lambda_{(j)}}{\sum_{j=1}^p \lambda_{(j)}} = 1 - \varepsilon, \quad (\text{A.1})$$

where $\lambda_{(j)}$ is the j -th order statistic (i.e. $\lambda_{(1)} > \lambda_{(2)} > \dots > \lambda_{(p)}$). If $\lambda_j \stackrel{iid}{\sim} \text{Gamma}(\nu, \mu)$ for $j = 1, \dots, p$, then we wish to find ν such that the median of the left side of the (A.1) is equal to $1 - \varepsilon$. Note that for a given ν , the distribution of the left side of (A.1) is the same for any μ . We repeatedly sample from a $\text{Gamma}(\nu, 1)$ distribution until the median of $\left(\sum_{j=1}^q \lambda_{(j)} / \sum_{j=1}^p \lambda_{(j)}\right)$ is within a small tolerance of $1 - \varepsilon$, by increasing or decreasing the current value of ν .

After eliciting a value for ν , we use a similar simulation to find an appropriate value for μ . We want the unrestricted Normal distribution of $\boldsymbol{\beta}$ to be close to zero near $p - q$ of the axes. Since the Normal distribution is flat in the tails, we want most of the distribution

to be inside the ellipsoid for $p - q$ directions. The ellipsoid restriction is $\boldsymbol{\beta}'\Sigma_X\boldsymbol{\beta} = \theta\sigma^2$, so we want to find μ such that probability of β_j being inside the ellipsoid at the axis is $1 - q/p$. Without the ellipsoid restriction this is $P(\beta_j^2 < \theta\sigma^2)$ since Σ_X is scaled to have 1's on the diagonal.

$$\begin{aligned} P(\beta_j^2 < \theta\sigma^2) &= P\left(w_j < \frac{1}{\lambda_j}\right) = \int_0^\infty \int_0^{1/n\lambda_j} f_w(w_j) f_\lambda(\lambda_j) dw_j d\lambda_j \\ &= \int_0^\infty \frac{1}{\Gamma(1/2)} \gamma\left(\frac{1}{2}, \frac{1}{2\lambda_j}\right) f_\lambda(\lambda_j) d\lambda_j \\ &= \frac{1}{\Gamma(1/2)} E_\lambda \left[\gamma\left(\frac{1}{2}, \frac{1}{2\lambda_j}\right) \right], \end{aligned}$$

where w_j is the a χ_1^2 random variable, and γ is the lower incomplete gamma function. We can estimate this quantity by simulating samples of λ_j from a $\text{Gamma}(\nu, \mu)$ distribution until the sample average of the above expression is approximately $(1 - q/p)$. Note, that because μ is a rate parameter, we actually only need to take one large sample from a $\text{Gamma}(\nu, 1)$ distribution, and just vary ν in the second parameter of the incomplete gamma function.

Appendix B

MCMC Samplers for Logistic Regression

B.1 Gibbs sampler for Uniform-on-ellipsoid Model

This Gibbs sampler of the BEERS (UOE) model for logistic regression is similar to the Gibbs sampler described for the linear model in Section 2.3.1. Proposition 2.2 showed that $\beta|\sigma^2$ has a mixture of normals representation for $a \leq p/2$. We change the notation here to not confuse the other parameters with latent Pólya–Gamma (PG) parameters. For the logistic regression model, $\sigma^2 = \pi^2/3$, the variance of a Logistic(0,1) distribution.

Assume the intercept, β_0 has a Cauchy(0,10) distribution. If $R_{MZ}^2 \sim \text{Beta}(a, b)$ and

$a \leq p/2$, then we have the equivalent representation for the prior of $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}_1|u, v \sim N_p(\mathbf{0}, \sigma^2 uv (\mathbf{X}'\mathbf{X})^{-1}) \quad (\text{B.1})$$

$$v \sim \text{inverse-Gamma}(b, n/2) \quad (\text{B.2})$$

$$u \sim \text{Beta}(a, p/2 - a) \quad (\text{B.3})$$

$$\beta_0 \sim N(0, \gamma) \quad (\text{B.4})$$

$$\gamma \sim \text{inverse-Gamma}(1/2, 10/2) \quad (\text{B.5})$$

The Gibbs sampler proceeds as follows.

- (a) Set initial values for $\boldsymbol{\beta}$, u , v and γ .
- (b) Sample the latent PG variables $\mathbf{w} = (w_1, \dots, w_n)'$. For $i = 1, \dots, n : w_i \sim PG(1, \mathbf{x}'_i \boldsymbol{\beta})$.
- (c) Sample $\boldsymbol{\beta}|u, v, \mathbf{w} \sim \text{Normal}(\boldsymbol{\mu}_\beta, \mathbf{V}_\beta)$,
 where $\mathbf{V}_\beta = \mathbf{X}'\Omega\mathbf{X} + \begin{pmatrix} 1/\gamma & \mathbf{0}' \\ \mathbf{0} & \mathbf{X}'_1\mathbf{X}_1/\sigma^2 uv \end{pmatrix}$ and $\boldsymbol{\mu}_\beta = \mathbf{V}_\beta^{-1}(\boldsymbol{\kappa}'\mathbf{X}\boldsymbol{\beta})$, $\Omega = \text{diag}(w_1, \dots, w_n)$ and $\kappa_i = y_i - 1/2$. \mathbf{X}_1 and $\boldsymbol{\beta}_1$ are the parts of \mathbf{X} and $\boldsymbol{\beta}$ with the intercept removed.
- (d) Sample $u|\boldsymbol{\beta}, z$, by first sampling $\tau \sim \text{Gamma}(p/2 - a, \boldsymbol{\beta}'_1\mathbf{X}'_1\mathbf{X}_1\boldsymbol{\beta}_1 / (2v\sigma^2))$, and setting $u = 1/(1 + \tau)$.
- (e) Sample $v|\boldsymbol{\beta}, u \sim \text{inverse-Gamma}(p/2 + b, \boldsymbol{\beta}'_1\mathbf{X}'_1\mathbf{X}_1\boldsymbol{\beta}_1 / (2u\sigma^2) + n/2)$.
- (f) Sample $\gamma|\beta_0 \sim \text{inverse-Gamma}(1, (\sigma^2 + \beta_0^2)/2)$.
- (g) Repeat steps b-e until convergence.

B.2 MCMC sampler for Local Shrinkage Model

The MCMC sampler for the local shrinkage BEERS priors follows closely the algorithm described in Section 2.3.2 as well as some of the changes in Appendix B.1. We make the following changes to the algorithm in Section 2.3.2:

- Introduce a latent PG variable.

At each iteration we sample $w_i \sim PG(1, \mathbf{x}'_i \boldsymbol{\beta})$, for $i = 1, \dots, n$.

- Sample the intercept: $\beta_0 \sim N(\mu_0, v_0)$, where

$$v_0 = (\sum_{i=1}^n w_i + 1/\tau)^{-1}, \text{ and } \mu_0 = v_0(\sum_{i=1}^n \kappa_i + \mathbf{w}' \mathbf{X}_1 \boldsymbol{\beta}_1).$$

- Sample the coefficient other than intercept from a Fisher-Bingham distribution.

$$\boldsymbol{\beta}_1 \sim FB(\boldsymbol{\mu}, \mathbf{A}), \text{ where } \boldsymbol{\mu} = (\boldsymbol{\kappa} + \beta_0 \mathbf{w})' \mathbf{X}_1 \mathbf{V} \mathbf{D}^{-1/2} \sqrt{\theta \sigma^2}, \text{ and}$$

$$\mathbf{A} = D^{-1/2} \mathbf{V}' (\mathbf{X}'_1 \Omega \mathbf{X}_1 \theta \sigma^2 + \boldsymbol{\Lambda}^{-1}) \mathbf{X}_1 \mathbf{V} \mathbf{D}^{-1/2} / 2.$$

- Sample $\gamma | \beta_0 \sim \text{inverse-Gamma}(1, (\sigma^2 + \beta_0^2)/2)$.

The rest of the sampler proceeds identically to that for the linear model, plugging in $\pi^2/3$ for σ^2 everywhere.