8-2019

Novel Bayesian methodology in multivariate problems.

Debamita Kundu
University of Louisville

NOVEL BAYESIAN METHODOLOGY IN MULTIVARIATE PROBLEMS

By

Debamita Kundu
B.Sc., Calcutta University, 2012
M.Sc., Presidency University, 2014

A Dissertation
Submitted to the Faculty of the
School of Public Health and Information Sciences
of the University of Louisville
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy
in Biostatistics

Department of Bioinformatics and Biostatistics
University of Louisville
Louisville, Kentucky

August 2019

NOVEL BAYESIAN METHODOLOGY IN MULTIVARIATE PROBLEMS

By

Debamita Kundu
B.Sc., Calcutta University, 2012
M.Sc., Presidency University, 2014

A Dissertation Approved on

June 18, 2019

by the following Dissertation Committee:

Jeremy T. Gaskins, Ph.D., Dissertation Director

Ritendranath Mitra, Ph.D., Co-director

Maiying Kong, Ph.D.

Subhadip Pal, Ph.D.

Little Bert, Ph.D.

DEDICATION

This dissertation is dedicated to my parents

Mr. Tapan Kumar Kundu
&
Mrs. Bithi Kundu

for their endless support and blessings throughout this journey.

ACKNOWLEDGMENTS

I would like to thank my advisors Drs. Jeremy T. Gaskins and Riten Mitra for their tireless support and excellent guidance towards the completion of this dissertation. This academic journey would not have been possible without their valuable advice. Their supportive and kind behavior made my Ph.D. life extremely smooth and enjoyable. I am really grateful to have them as my advisors throughout this journey.

I would also like to thank my all dissertation committee members, Drs. Maiying Kong, Subhadip Pal and Little Bert, for their assistance and thoughtful guidance. I would like to express my gratitude to the Department of Bioinformatics and Biostatistics for supporting me financially throughout this time. I am thankful to all the faculty, students, and administrative staff of the Department of Bioinformatics and Biostatistics for their support.

Finally, I would like to express my gratitude to my parents Mr. Tapan Kumar Kundu and Mrs. Bithi Kundu and my brother Mr. Tanmoy Kundu for their tremendous support and immense encouragement. I am specially grateful to Mr. Soutik Ghosal for his support without which this journey would not be completed. I am thankful to all my friends and family members for their invaluable support.

ABSTRACT

NOVEL BAYESIAN METHODOLOGY IN MULTIVARIATE PROBLEMS

Debamita Kundu

June 18, 2019

This dissertation involves developing novel Bayesian methodology for multivariate problems. In particular, it focuses on two contexts: shrinkage based variable selection in multivariate regression and simultaneous covariance estimation of multiple groups. Both these projects are centered around fully Bayesian inference schemes based on hierarchical modeling to capture context-specific features of the data and the development of computationally efficient estimation algorithm.

Variable selection over a potentially large set of covariates in a linear model is quite popular. In the Bayesian context, common prior choices can lead to a posterior expectation of the regression coefficients that is a sparse (or nearly sparse) vector with a few non-zero components, those covariates that are most important. The first project extends the global-local shrinkage idea to a scenario where one wishes to model multiple response variables simultaneously. Here, we have developed a variable selection method for a K-outcome model (multivariate regression) that identifies the most important covariates across all outcomes. The prior for all regression coefficients is a mean zero normal with coefficient-specific variance term that consists of a predictor-specific factor (shared local shrinkage parameter) and a model-specific factor (global shrinkage term) that differs in each model. The performance of our

modeling approach is evaluated through simulation studies and a data example.

Covariance estimation for multiple groups is a key feature for drawing inference from a heterogeneous population. One should seek to share information about common features in the dependence structures across the various groups. In the second project, we introduce a novel approach for estimating the covariance matrices for multiple groups using a hierarchical latent factor model that shrinks the factor loadings across groups toward a global value. Using a spike and slab model on these loading coefficients provides a level of sparsity in the global factor structure. Parameter estimation is accomplished through a Markov chain Monte Carlo scheme, and a model selection approach is used to determine the number of factors to use. Finally, a number of simulation studies and a data application are shown to demonstrate the performance of our methodology.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Bayesian Variable Selection for Multi-Outcome Models Through Shared Shrinkage

In the context of high-dimensional data, it is critical to correctly identify a set of variables that significantly influences the responses and play an important role in prediction. Consider a set of p potential regressors $X_1, X_2, \ldots, X_p$ and a single response variable Y. In order to increase the precision of statistical estimates and prediction, we often consider a model of the form

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p + \epsilon,$$

where many of the β are exactly zero, so that only the set of $q\,(\leq p)$ regressors impact the response Y.

In the Bayesian context there are numerous approaches to the problem of variable selection. Mitchell and Beauchamp (1988) proposed the "spike and slab" approach by considering a mixture prior distribution for the regressor coefficient: a zero component (spike) and a disperse component (slab). Specifically, indicator variables were used to differentiate the important regressors from the rest. When the indicator assumes the value 0, the prior for the corresponding regression coefficient is set to follow a Gaussian with low variance. This is the zero component (spike). Otherwise, it follows a Gaussian with high variance, representing the disperse com-

ponent (slab). For this setup, George and McCulloch (1993) suggested stochastic search variable selection (SSVS) for identifying a "promising" subset. This framework was later extended to incorporate several non-conjugate and conjugate priors for prior specification (George and McCulloch, 1997). Subsequently, a related class of variable selection priors that put positive mass at 0 are based on Reversible Jump (RJ) sampling techniques (Green and Hastie, 2009). However, these selection methods require updating each regression coefficient conditionally on all others and tend to be computationally slow and display poor mixing when used for a large number of variables.

Hence, shrinkage priors have gained popularity recently as a computationally faster alternative. Rather than using a mixture prior that can set the coefficient exactly to zero, the shrinkage approach employs priors designed to pull small signals aggressively towards zero. Many of the commonly used shrinkage models fall within the global-local (GL) shrinkage framework defined by Polson and Scott (2010). In the usual multiple regression setting where the regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is believed to be sparse, the typical GL shrinkage prior for the β vector would be

$$\beta_j \sim N(0, \lambda_j^2 \tau^2),$$

$$\lambda_j \sim f(\cdot), \qquad \tau \sim g(\cdot).$$

In this model τ controls global shrinkage towards the origin, and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_p)$ are the local shrinkage parameters that allows deviation in the degree of shrinkage between predictors. The typical recommendation is that $f(\cdot)$ should have heavy tails to avoid over-shrinking large signals, and $g(\cdot)$ should have substantial mass near zero. The Normal-gamma prior (Griffin and Brown, 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015) and the horseshoe prior (Carvalho et al., 2010) are three popular methods in this framework. A review and comparison of various

variable selection methods including the shrinkage methods can be found in O'Hara and Sillanpää (2009).

Although much of the literature focuses on the situation of multiple regression with a single response variable y, the problem of variable selection when simultaneously analyzing multiple responses (multivariate regression) is much less explored. For example, multiple outcomes measuring different aspects of a patient's health (blood pressure, glucose, etc.) may be modeled using a potentially large set of risk factor predictors. In many cases, each outcome is analyzed separately with variable selection performed unique to each outcome, but this will be inefficient if each model has the same or a similar set of relevant predictors. However, borrowing strength across regression coefficients can boost the power of detecting true signals, especially if the responses share similar predictors and there is reason to believe that they exert similar influences on the responses. The gain in performance can be substantial for low to moderate sample sizes and complex noise structures. Instead of applying variable selection separately for each outcome, Brown et al. (1999, 1998) propose two approaches based on finding a common set of predictors for all models by extending the George and McCulloch's selection model (1993; 1997). However, by requiring predictors to affect either all K outcomes or none of them, their models are often overly restrictive. Hence, in this work we focus on developing a more flexible variable selection method that encourages the inclusion of similar sets of predictors in each of the K models by extending the GL shrinkage framework. Recently, Bai and Ghosh (2018) independently explored a similar setup and proposed their Multivariate Bayesian Model with Shrinkage Priors (MBSP). We will discuss differences that distinguish our work in later sections. In a frequentist setting, Turlach et al. (2005) proposed a LASSO-based approach with penalties based on the maximum absolute coefficient across all outcomes for each predictor.

1.2 A Bayesian Hierarchical Sparse Factor Model for Simultaneous Covariance Estimation

In the analysis of multivariate data, the estimation of the covariance matrix is always one prime interest. However, when data consist of multiple groups, each may be determined by its own covariance matrix. In this work, we consider data that consist of M groups, where the covariance matrix of group m is Ω_m $(m = 1, 2, \ldots, M)$. Our interest is in developing methodology to estimate this collection $\{\Omega_1, \Omega_2, \ldots, \Omega_M\}$. When faced with this scenario, it is not uncommon for the analyst to assume equality across all Ω_ms, but this will lead to erroneous inference if there are truly differences across the covariances. Conversely, estimating each Ω_m without sharing information across all groups will lead to inefficient estimation if there are common structures shared across groups. Hence, developing a reasonable method for borrowing strength across groups in the simultaneous covariance estimation problem is paramount for obtaining trustworthy inference.

In the literature of simultaneous covariance estimation, principal component methods are a well-established approach. Flury (1984) developed a method with common eigenvectors to estimate the covariance matrices by considering $\Omega_m = Q\Gamma_m Q$, where Ω_m is the $p \times p$ covariance matrix for the m^{th} group, Q is the $p \times p$ orthogonal matrix of eigenvectors that are shared across all groups and Γ_m is the diagonal matrix of positive eigenvalues specific to group m. Later, Flury (1987) extended this to the "partial common principal component model" by assuming q $(q < p)$ common eigenvectors across all Ω_ms, and the remaining eigenvectors are group-specific. Boik (2002) broadened the idea to a more general model by sharing the eigenvectors between some or all groups. Hoff (2009) also developed a hierarchical Bayesian model that shrinks the eigenvector matrix of each group across the population by using a shrinkage prior on the matrix of eigenvectors. Besides this usual spectral decomposition, Manly and

Rayner (1987) and Barnard et al. (2000) proposed decomposing the covariance matrix in terms of the standard deviation matrices (S) and correlation matrices (R), i.e: $\Omega_m = S_m R_m S_m$, and assumed R and S are independent and the correlation matrices are the same across the groups.

In the context of longitudinal data, there are additional methods based on the modified Cholesky decomposition of the covariance matrix (Pourahmadi, 1999). Pourahmadi et al. (2007) highlighted on computational advantages and fundamental differences of the unconstrained parameterization of the Cholesky decomposition for modelling several covariance matrices simultaneously in comparison to traditional eigenvalue or variance-correlation decomposition. Unlike the spectral decomposition and variance-correlation decomposition, the units that appear in the lower triangular matrix, termed as general autoregressive parameters (GARP) of the Cholesky decomposition are always unconstrained and hence involves unconstrained maximization techniques for computing maximum likelihood estimates. McNicholas and Murphy (2010) considered Gaussian mixture models in order to propose a model-based clustering framework for longitudinal data, where the modified Cholesky decompositions of the group covariance matrices are considered to have commonalities across all groups. Gaskins and Daniels (2012) proposed a family of nonparametric priors based on Dunson et al. (2008)'s matrix stick-breaking process. Their method uses the parameters from modified Cholesky decomposition which includes GARP and the innovation variances (IV) to parametrize the covariance matrix for each group. Additionally, this methodology sets some parameters of the Cholesky decomposition to zero to provide a lower-dimensional structure for the covariance matrix. Later, Gaskins and Daniels (2016) proposed a related approach that partitions the collection of groups into sets with common conditional distributions.

As an estimator for a single covariance matrix, latent factor models traditionally play an important role in modeling multivariate dependence structures in the

behavioral sciences. The essential purpose of factor analysis is to describe as the underlying covariance relationship between many variables in terms of a few unobserved random quantities, called *factors*. Consider, a situation where a researcher assembled a moderate to a large number of predictors for an analysis. In general a p-dimensional predictor variable has $p(p-1)/2$ pairwise correlation. However, when p is moderately large it is very difficult to summarize and interpret all pairwise correlations together. The factor model assumes that complex correlation structure can be explained by some latent linear combinations of fewer variables, leading to a reduction in dimension. These underlying unobserved random variables are termed as latent *factors*. This is a parsimonious model. As the number of latent factors $K \ll p$, therefore instead of p^2 terms, we need to deal with only $p(K+1)$ terms. Further, it makes interpretation simpler if variables are grouped by their underlying correlation structure. For example, if we have test scores from different subjects of a group of student, we may consider as mathematics, vocabulary, physics score as "intelligence" factor, weight, BMI, energy level as "physical fitness" factor and sociability, gregariousness, lack of shyness as "psychological" factor.

Selection of the appropriate number of factors is a key issue in such models, and traditional model selection criteria such as AIC or BIC are standard choices. In a Bayesian factor model Lopes and West (2004) considered the number of factors itself to be an unknown parameter. They introduced a customized reversible jump Markov chain Monte Carlo (RJMCMC) algorithm to sample from the model with a variable number of factors. Additionally, Ghosh and Dunson (2009) proposed an efficient parameter expansion algorithm to improve the computational efficiency of Bayesian factor models. Also, Bhattacharya and Dunson (2011) have applied a multiplicative gamma process shrinkage prior to Bayesian latent factor models to model a sparse covariance matrix for high-dimensional data by using infinite number of factors. Due to its use in several applied areas such as pattern recognition, financial time series

modeling, bioinformatics and computer vision, the theory of factor models analysis has received huge attention. However, the use of the latent factor model is relatively uncommon in the context of estimating the multiple covariance matrices.

CHAPTER 2

BAYESIAN VARIABLE SELECTION FOR MULTI-OUTCOME MODELS THROUGH SHARED SHRINKAGE

2.1 Introduction

In the context of high-dimensional data, it is critical to correctly identify a set of variables that significantly influences the responses and play an important role in prediction. Consider a set of p potential regressors $X_1, X_2, \ldots, X_p$ and a single response variable Y. In order to increase the precision of statistical estimates and prediction, we often consider a model of the form

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p + \epsilon,$$

where many of the β are exactly zero, so that only the set of $q\,(\leq p)$ regressors impact the response Y.

In the Bayesian context there are numerous approaches to the problem of variable selection. Mitchell and Beauchamp (1988) proposed the "spike and slab" approach by considering a mixture prior distribution for the regressor coefficient: a zero component (spike) and a disperse component (slab). Specifically, indicator variables were used to differentiate the important regressors from the rest. When the indicator assumes the value 0, the prior for the corresponding regression coefficient is set to follow a Gaussian with low variance. This is the zero component (spike). Otherwise, it follows a Gaussian with high variance, representing the disperse com-

ponent (slab). For this setup, George and McCulloch (1993) suggested stochastic search variable selection (SSVS) for identifying a "promising" subset. This framework was later extended to incorporate several non-conjugate and conjugate priors for prior specification (George and McCulloch, 1997). Subsequently, a related class of variable selection priors that put positive mass at 0 are based on Reversible Jump (RJ) sampling techniques (Green and Hastie, 2009). However, these selection methods require updating each regression coefficient conditionally on all others and tend to be computationally slow and display poor mixing when used for a large number of variables.

Hence, shrinkage priors have gained popularity recently as a computationally faster alternative. Rather than using a mixture prior that can set the coefficient exactly to zero, the shrinkage approach employs priors designed to pull small signals aggressively towards zero. Many of the commonly used shrinkage models fall within the global-local (GL) shrinkage framework defined by Polson and Scott (2010). In the usual multiple regression setting where the regression coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is believed to be sparse, the typical GL shrinkage prior for the β vector would be

$$\beta_j \sim N(0, \lambda_j^2 \tau^2),$$

$$\lambda_j \sim f(\cdot), \qquad \tau \sim g(\cdot).$$

In this model τ controls global shrinkage towards the origin, and $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_p)$ are the local shrinkage parameters that allows deviation in the degree of shrinkage between predictors. The typical recommendation is that $f(\cdot)$ should have heavy tails to avoid over-shrinking large signals, and $g(\cdot)$ should have substantial mass near zero. The Normal-gamma prior (Griffin and Brown, 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015) and the horseshoe prior (Carvalho et al., 2010) are three popular methods in this framework. A review and comparison of various

variable selection methods including the shrinkage methods can be found in O'Hara and Sillanpää (2009).

Although much of the literature focuses on the situation of multiple regression with a single response variable y, the problem of variable selection when simultaneously analyzing multiple responses (multivariate regression) is much less explored. For example, multiple outcomes measuring different aspects of a patient's health (blood pressure, glucose, etc.) may be modeled using a potentially large set of risk factor predictors. In many cases, each outcome is analyzed separately with variable selection performed unique to each outcome, but this will be inefficient if each model has the same or a similar set of relevant predictors. However, borrowing strength across regression coefficients can boost the power of detecting true signals, especially if the responses share similar predictors and there is reason to believe that they exert similar influences on the responses. The gain in performance can be substantial for low to moderate sample sizes and complex noise structures. Instead of applying variable selection separately for each outcome, Brown et al. (1999, 1998) propose two approaches based on finding a common set of predictors for all models by extending the George and McCulloch's selection model (1993; 1997). However, by requiring predictors to affect either all K outcomes or none of them, their models are often overly restrictive. Hence, in this work we focus on developing a more flexible variable selection method that encourages the inclusion of similar sets of predictors in each of the K models by extending the GL shrinkage framework. Recently, Bai and Ghosh (2018) independently explored a similar setup and proposed their Multivariate Bayesian Model with Shrinkage Priors (MBSP). We will discuss differences that distinguish our work in later sections. In a frequentist setting, Turlach et al. (2005) proposed a LASSO-based approach with penalties based on the maximum absolute coefficient across all outcomes for each predictor.

The layout of this manuscript is as follows. In section 2.2, we describe a general

strategy for GL shrinkage in multivariate regression. and explore details when paired with the 3 common GL models, Normal-gamma, Dirichlet-Laplace, and horseshoe, as well as relevant posterior consistency results. Section 2.3 discusses posterior sampling for each of these models, and Section 2.4 considers simulation studies to explore the performance of our model. In Section 2.5 we analyze a real data set based on the yeast cell cycle data (Chun and Keleş, 2010), and we conclude with a brief discussion in Section 2.6.

2.2 Multi-outcome Regression Coefficient Shrinkage Model

2.2.1 General Strategy

Consider a multi-outcome (multivariate) model with K outcomes/responses, p covariates and n independent observations. We write the multivariate regression model in the following form,

$$\begin{bmatrix} y_{11} & y_{12} & \dots & y_{1K} \\ y_{21} & y_{22} & \dots & y_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \dots & y_{nK} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1K} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pK} \end{bmatrix} + \varepsilon, \qquad (2.1)$$

where $Y_{i\cdot}$, the i^{th} row of the $n \times k$ matrix Y, consists of the K responses for the ith observation and $X_{i\cdot}$ is the i^{th} row of the model matrix X which contains the p predictor variables for this observation. The matrix of regression coefficients B is believed to be sparse. Further, as each row of B corresponds to the regression coefficients of predictor j on each of the K responses, we expect similar sparsity across the row. ε is the $n \times K$ residual matrix. Under the normality assumption, each row of the residual matrix follows a $N_K(0, \Psi)$ distribution independently. For simplicity, we ignore the

intercept terms for right now. Note also that throughout we assume that the columns of Y and X have been standardized. This gives a multivariate normal distribution for the vector of responses for patient i, $Y_{i\cdot} \sim MVN_K\left(X_{i\cdot}B, \Psi\right)$.

Variable selection is induced through the choice of prior on the B matrix. Our approach is to extend the global-local shrinkage framework to jointly model multiple responses. The general idea of our method is to share information about the importance of a covariate across all response models through a local-shrinkage parameter $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ and use a response-specific global shrinkage parameter $\tau = (\tau_1, \tau_2, \ldots, \tau_K)$ to allow for different scalings of the regression coefficients in the different response models. Following the usual GL framework, our prior for the coefficient matrix B comes from the following general hierarchy,

$$\begin{aligned} &\beta_{jk} \sim N(0, \lambda_j^2\tau_k^2), \quad (j = 1, 2, \cdots, p, \quad k = 1, 2, \cdots, K)\,, \\ &\lambda_j \sim f(\cdot), \\ &\tau_k \sim g(\cdot). \end{aligned} \tag{2.2}$$

The choices of the local distribution $f(\cdot)$ and the global distribution $g(\cdot)$ can be borrowed from any of the common global-local models. In particular, we focus on the utility of this approach under the following three choices: the Normal-gamma prior (Griffin and Brown, 2010), the horseshoe prior (Carvalho et al., 2010), and the Dirichlet-Laplace prior (Bhattacharya et al., 2015). The value of the local parameter λ_j will encourage similar levels of shrinkage/sparsity for all coefficients $(\beta_{j1}, \beta_{j2}, \cdots, \beta_{jK})$ of the j^{th} predictor. Following the usual GL shrinkage rules, we choose the local distribution $f(\cdot)$ to have heavy tails and $g(\cdot)$ to have substantial mass near zero (Polson and Scott, 2010). A large λ_j allows $\beta_{jk}\,(k = 1, 2, \ldots, K)$ coefficients far from zero, whereas a small λ_j will ensure all coefficients for predictor j are aggresively shrunk toward zero. Note that if there is only a single response $K = 1$, then our approach is exactly equivalent to the usual global-local framework. Finally,

note that the general framework (2.2) specifies the distributions $f(\cdot)$ and $g(\cdot)$ for the global and local parameters on the scales of the standard deviation of β_{jk}. In some cases, it may be more natural for $f(\cdot)$ and/or $g(\cdot)$ to represent the distribution for the variance contributions λ_j^2 and τ_k^2, respectively.

Despite similarities of our framework to that of Bai and Ghosh (2018), there are several key differences between our approaches. First, their MBSP model specifies a common value τ for the global τ_k parameters across all models. Further, this parameter is a priori fixed based on asymptotic arguments. Conversely, we recognize that there may be variability in the global scale of the coefficients between response models, and we allows differing τ_k which are estimated from the data. Secondly, MBSP specifies the column covariance of B to be proportional to Ψ, the residual covariance matrix. This choice facilitates additional conjugacy in their sampler, but we opt to allow the columns of B to be independent (given the τ_ks) as a more intuitive choice. As will be shown in Section 2.3, we are able to retain a high degree of conjugacy and develop an efficient posterior sampler.

Having defined our general approach, we now focus on three versions of our methodology by using common shrinkage models.

2.2.2 Multi-outcome Normal-gamma Model

First, we apply the Normal-gamma shrinkage prior from Griffin and Brown (2010) to our method. We refer this model as the Multi-outcome Normal-gamma Model

(MONG). This yields the following hierarchy:

$$\begin{aligned}
\beta_{jk} &\sim N(0, \lambda_j \tau_k^2), \qquad (j = 1, 2, \ldots, p; \quad k = 1, 2, \ldots, K)\,, \\
\lambda_j &\sim Gamma\left(c, \frac{1}{c}\right), \\
\tau_k &\sim C^+(\gamma)\,, \\
c &\sim Exp(\lambda_c).
\end{aligned} \tag{2.3}$$

In (2.3), λ_j comes from a $Gamma\left(c, \frac{1}{c}\right)$ distribution such that the prior mean is 1 and variance is $\frac{1}{c}$. Hence, small values of c will induce greater variability within the λs and more shrinkage. The tail of β_{jk} thickens with increasing c. A common special case involves setting $c = 1$ which provides the Bayesian LASSO (Park and Casella, 2008). For the prior distribution of τ, we consider a half-Cauchy distribution with density $f(x) = \frac{2\gamma}{\pi(\gamma^2 + x^2)}, x > 0$. The intuition behind considering half-Cauchy prior for global shrinkage parameter is its non-zero density near the origin with thick tails in the extremes. We recommend setting the scale parameter of this half-Cauchy to $\gamma = 0.5$ to provide a reasonably dispersed distribution for the τs, and this choice has performed well in empirical studies. For the hyper-parameter c we consider an exponential density with mean 2 to encourage slightly thicker tails in β_{jk} than the Bayesian LASSO.

2.2.3 Multi-outcome Horseshoe Model

The horseshoe prior is one of the most appealing and commonly used shrinkage priors in the literature. It became popular due to its infinitely tall spike in the density near the origin that shrinks almost everything towards zero and its flat, Cauchy-like tails that allow some parameters to escape from shrinkage. The conventional horseshoe prior places half-Cauchy priors on both the local and global contributions to the standard deviation. The Multi-outcome Horseshoe Model (MOHS) is defined by the

following hierarchy:

$$\begin{aligned} \beta_{jk} &\sim N(0, \lambda_j^2 \tau_k^2), \\ \lambda_j &\sim C^+(1), \\ \tau_k &\sim C^+(1). \end{aligned} \tag{2.4}$$

In its usual form, the model (2.4) is not conjugate, making implementation in a standard Gibbs sampling scheme difficult and time-consuming. However, Makalic and Schmidt (2016) proposed an efficient, conditionally conjugate sampling algorithm for fast updating by introducing data augmentation variables from an inverse gamma distribution. Since the marginal distribution of χ from the hierarchy $\chi^2 \mid \Upsilon \sim IG\left(\frac{1}{2}, \frac{1}{\Upsilon}\right)$ and $\Upsilon \sim IG\left(\frac{1}{2}, 1\right)$ is $C^+(1)$, we equivalently write this model as

$$\begin{aligned} \beta_{jk} &\sim N\left(0, \lambda_j^2 \tau_k^2\right), \\ \lambda_j^2 &\sim IG\left(\frac{1}{2}, \frac{1}{\nu_j}\right), \\ \tau_k^2 &\sim IG\left(\frac{1}{2}, \frac{1}{\omega_k}\right), \\ \nu_1, \nu_2, \ldots, \nu_p, \omega_1, \omega_2, \ldots, \omega_K &\sim IG\left(\frac{1}{2}, 1\right). \end{aligned} \tag{2.5}$$

Note that we define IG to have density function $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$, $x > 0$.

In both the MONG and MOHS versions, we may use the λ parameters to judge the importance of a predictor across all responses. The larger the local parameter the less shrinkage in the regression coefficients and the greater the predictive power. Hence, the estimated $\hat{\lambda}_j$ can be used as a summary of the importance of predictor j across all K models. In both cases, we may compare this value relative to 1, the prior mean for λ_j in MONG and the prior median in MOHS.

2.2.4 Multi-outcome Dirichlet-Laplace Model

In a similar manner, we also define the Multi-outcome Dirichlet-Laplace Model (MODL). Like the previous GL methods, the DL model considers the dispersion of the j^{th} coefficient to be a contribution of local and global scaling terms. However, the conditional distribution of the coefficient is Laplace (double exponential) instead of the usual normal distribution. While this may not technically fall in Polson and Scott (2010)'s GL framework, it is clearly in the same spirit, and can be paired with our multi-outcome shrinkage framework. The proposed MODL model has the following specification

$$\begin{aligned} \beta_{jk} &\sim DE\left(\phi_j \tau_k\right), \\ \tau_k &\sim Gamma\left(pa, \frac{1}{2}\right), \\ \phi = (\phi_1, \phi_2, \ldots, \phi_p) &\sim Dirichlet\left(a, a, \ldots, a\right), \end{aligned} \tag{2.6}$$

where a is concentration parameter of the Dirichlet distribution. In this model the local parameters ϕ_j sum to one, and smaller values of a will lead ϕ to be dominated by a few components. Since the majority of the DE scales $\phi_j \tau_k$ will be approximately zero, sparsity in the β_{jk} is achieved. As recommended by Bhattacharya et al. (2015), we considered $a = \frac{1}{2}$ or $a = \frac{1}{p}$ for our simulation and case study.

Similar to the HS model (2.5), we can introduce auxiliary variables to facilitate sampling. One may represent the $\beta_{jk} \sim DE\left(\phi_j \tau_k\right)$ as scale mixture of normals through $\beta_{jk} \mid \eta_{jk} \sim N\left(0, \eta_{jk} \phi_j^2 \tau_k^2\right)$ with $\eta_{jk} \sim Exp\left(\frac{1}{2}\right)$. Similar to using $\hat{\lambda}$ to evaluate predictor relevance in the MONG and MOHS models, in this MODL proposal we can compare the estimated ϕ_js to their prior mean $1/p$. Again, larger values indicate less shrinkage and greater predictor relevance across all outcomes.

Across all models for the regression coefficients the residual covariance matrix is given an inverse Wishart prior with $K+2$ degrees of freedom and the identity matrix as the prior scale matrix. This gives the prior mean for Ψ as the identity matrix. As

is common, we recommend responses and predictors be centered and scaled prior to analysis.

2.2.5 Posterior Consistency

In this section, we present a result guaranteeing posterior consistency in our model structure. For this proof, we will assume that the residual covariance matrix Ψ is fixed and known. We first state the assumptions before proving our consistency result.

Assumptions:

(A1) The prior $\pi(B)$ is continuous in B over all of $\mathbb{R}^{p\times K}$.

(A2) The vector of covariates are uniformly bounded. That is, there exist $G > 0$ such that $||X_{i\cdot}|| < G$ for all $i = 1, \ldots, n$.

(A3) The smallest eigenvalue of the design matrix is asymptotically bounded away from zero. There exists $c > 0$ such that $\liminf_{n\to\infty} \lambda_1(\frac{1}{n}X'X) > c$, where $\lambda_1(M)$ refers to the smallest eigenvalue of the matrix M.

Note that (A1) represents a much more general class of prior models than our GL shrinkage framework, although our proposal clearly falls within this assumption. Throughout, we use the Frobenius norm, $||M|| = \sqrt{\sum_{i,j}(m_{ij})^2}$. Note also that any deterministic functions of the $n \times p$ design matrix X depends on the sample size n. To avoid cumbersome notation we typically suppress the dependence on n and refer to it as simply X.

First, we state our key theorem about posterior consistency.

Theorem 1. *Assume a fixed, positive definite* Ψ *and assumptions (A1)-(A3). Let* $Y_{1\cdot}, \ldots, Y_{n\cdot}$ *be iid from model* (2.1) *under the true parameter value* B_0*. Then for any* $\epsilon > 0$*,*

$$P_{B_0}\{\, ||B - B_0|| > \epsilon \mid Y_{1\cdot}, \ldots, Y_{n\cdot}\} \to 0, \quad \textit{a.s. as } n \to \infty.$$

That is, the posterior distribution for B almost surely collapses to the true value B_0 as $n \to \infty$.

This proof along with the associated lemmas appears in the Appendix. It builds upon Schwartz's seminal proof (Schwartz, 1965), in combination with results for regression models from Amewou-Atisso et al. (2003) and Choi and Schervish (2007). The argument mainly relies on the existence of an uniformly exponentially consistent (UEC) sequence of tests and a prior positivity property. The latter in Schwartz's original proof was simply the condition that the prior mass on all Kullback–Leibler (KL) neighborhoods of the true parameter is greater than zero. However, as we show in the Appendix, this KL framework must be modified into a multi-index version for its use in models with covariates. Both of these two conditions are derived as separate lemmas that can be combined to give posterior consistency. See the Appendix for full details.

An important feature of Theorem 1 is its flexible prior condition stated in (A1). This relaxation comes at a cost, mainly assumption (A2), which essentially bounds the entries of the design matrix. In contrast, Bai and Ghosh (2018) assume upper and lower (asymptotic) bounds on the eigenvalues of the design matrix. However, the flexibility gained under our choice is significant, as we require no condition (except continuity) on the prior for B. This is much more general than the assumptions made in the consistency theorems of Bai and Ghosh (2018) and Armagan et al. (2013). Their choices require conditions on the prior with convoluted formulas involving Ψ and the eigenvalues of the design matrix, thus restricting the choice of prior on B in ways that are not straightforward.

2.3 Posterior Computation

As with most modern Bayesian models, inference is performed by approximating the posterior through Markov chain Monte Carlo (MCMC) methods. We describe the

necessary sampling steps for each of our three models below.

2.3.1 MONG Model

(i) Sample $vec(B) \mid X, \Psi, \lambda, \tau$ from $MVN_{pK}(M, W)$, where $W = \left(\left(\Psi^{-1} \otimes X^T X\right) + \Omega^{-1}\right)^{-1}$ and $M = W\left(\Psi^{-1} \otimes X^T\right) vec(Y)$. Here, $\Omega = T \otimes \Lambda$ the prior covariance matrix of $vec(B)$, $\Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$ and $T = diag(\tau_1^2, \tau_2^2, \cdots, \tau_K^2)$. Throughout, we let $\otimes$ denote the Kronecker product.

(ii) For $j = 1, 2, \cdots, p$, sample $\lambda_j \mid \beta_{jk}, \tau_k, c \sim giG\left(c - \frac{K}{2}, 2c, \sum_{k=1}^{K} \frac{\beta_{jk}^2}{\tau_k^2}\right)$, where $giG(\kappa, \chi, \rho)$ is the generalized inverse Gaussian distribution with density $f(x; \kappa, \chi, \rho) \propto x^{\rho-1} e^{-\frac{1}{2}\left(\kappa x + \frac{\chi}{x}\right)}$, $x > 0$.

(iii) The posterior density of τ_k does not have a conjugate distribution. The conditional posterior sampling distribution of τ_k is given by

$$\pi(\tau_k \mid \beta_{jk}\lambda_j) \propto \tau_k^{-p} \exp\left[-\frac{1}{\tau_k^2} \sum_{j=1}^{p} \frac{\beta_{jk}^2}{2\lambda_j}\right] \frac{\gamma^2}{(\tau_k^2 + \gamma^2)}.$$

For each $k = 1, 2, \cdots, K$, an adaptive Metropolis-Hastings (MH) step is applied to attempt an update to τ_k, based on algorithm 4 of Andrieu and Thoms (2008) applied to τ_k.

(iv) Similarly, c does not have a conjugate sampling density. The conditional posterior density of c is given by

$$\pi(c \mid \lambda_1, \lambda_2, \ldots, \lambda_p) \propto \frac{c^{cp}}{\Gamma(c)^p} \exp\left[-c\left(\lambda_c + \sum_{j=1}^{p} \lambda_j\right) + (c-1) \sum_{j=1}^{p} \log \lambda_j\right], \; c > 0.$$

An adaptive MH step based on the Andrieu and Thoms (2008) algorithm is also performed here.

(v) Ψ is drawn from $Inv{-}Wishart(\upsilon_0 + n, S_0 + S)$, where $S = (Y - XB)^T (Y - XB)$.

2.3.2 MOHS Model

Sampling steps for MOHS model are described below.

(i) Sampling distribution for $vec\,(B) \mid X, \Psi, \lambda, \tau$ is the same as in MONG step (i), except $\Lambda = diag(\lambda_1^2, \lambda_2^2, \cdots, \lambda_p^2)$ here.

(ii) For $j = 1, 2, \cdots, p$, sample $\lambda_j^2 \mid \beta_{jk}, \tau_k, \nu_j \sim IG\left(\frac{K+1}{2}, \frac{1}{\nu_j} + \sum_{k=1}^{K} \frac{\beta_{jk}^2}{2\tau_k^2}\right)$.

(iii) For $k = 1, 2, \cdots, K$, sample $\tau_k^2 \mid \beta_{jk}, \lambda_j, \epsilon_k \sim IG\left(\frac{p+1}{2}, \frac{1}{\omega_k} + \sum_{j=1}^{p} \frac{\beta_{jk}^2}{2\lambda_j^2}\right)$.

(iv) For $j = 1, 2, \cdots, p$, sample $\nu_j \mid \lambda_j \sim IG\left(1, 1 + \frac{1}{\lambda_j^2}\right)$.

(v) For $k = 1, 2, \cdots, K$, sample $\omega_k \mid \tau_k \sim IG\left(1, 1 + \frac{1}{\tau_k^2}\right)$.

(vi) Sample $\Psi \mid B \sim Inv - Wishart\,(v_0 + n, S_0 + S)$.

2.3.3 MODL Model

For the original DL specification, (Bhattacharya et al., 2015) propose a block sampler that involves marginalizations over different sets of parameters. Due to sharing ϕ_js across multiple outcome models, this is no longer feasible in our MODL model (2.6), and we require (adaptive) Metropolis-Hasting to jointly sample the vector $(\phi_1, \ldots, \phi_p)$ of local parameters. Sampling steps are as follows:

(i) First sample $vec(B) \mid X, \Psi, \phi, \eta, \tau$ from $MVN_{pK}\,(M, W)$. The conditional posterior distribution of $vec(B)$ is as in the case of NG prior except $\Omega = diag\left(\eta_{jk}\phi_j^2\tau_k^2\right)$.

(ii) For $k = 1, 2, \cdots, K$, sample $\tau_k \mid \beta_{jk}, \phi$ (marginalizing over η) from a generalized inverse Gaussian distribution $giG\left(pa - p, 1, 2\sum_{j=1}^{p} \frac{\beta_{jk}}{\phi_j}\right)$.

(iii) The conditional posterior density of $\phi|B,\tau$ (marginalizing over η) is proportional to

$$\pi\left(\phi_1,\phi_2,\ldots,\phi_p \mid B,\tau\right) \propto \prod_{j=1}^{p} \phi_j^{a-K-1} \exp\left[-\frac{1}{\phi_j}\sum_{k=1}^{K}\frac{\mid \beta_{jk} \mid}{\tau_k}\right], \tag{2.7}$$

where ϕ resides in the $(p-1)$-dimensional simplex. We have used an adaptive MH algorithm by extending algorithm 4 of Andrieu and Thoms (2008) for sampling ϕ. We sample from distribution (2.7) as described below:

- At the t^{th} iteration, sample the proposed move by

$$\left(\phi_1^*,\phi_2^*,\ldots,\phi_p^*\right) \sim Dirichlet\left(\zeta^{(t)}\phi_1,\zeta^{(t)}\phi_2,\ldots,\zeta^{(t)}\phi_p\right). \tag{2.8}$$

  The $\zeta^{(t)}$ is a positive tuning parameter that controls the dispersion of the proposal distribution. Note that this choice behaves similarly to a random walk with $E\phi_j^* = \phi_j$ and $Var\left(\phi_j^*\right) = \frac{\phi_j(1-\phi_j)}{1+\zeta^{(t)}}$. The variance of our candidate is inversely related to $\zeta^{(t)}$.

- Calculate the MH probability $\alpha = min\left(1, \frac{\pi(\phi^*|B,\tau)\, g\left(\phi_1,\phi_2,\ldots,\phi_p|\phi_1^*,\phi_2^*,\ldots,\phi_p^*\right)}{\pi(\phi|B,\tau)\, g\left(\phi_1^*,\phi_2^*,\ldots,\phi_p^*|\phi_1,\phi_2,\ldots,\phi_p\right)}\right)$, where $g(\cdot)$ is the proposal distribution (2.8). With probability α, we accept the proposed value $\phi^* = \left(\phi_1^*,\phi_2^*,\ldots,\phi_p^*\right)$, and otherwise, we retain the current $\phi = (\phi_1,\phi_2,\ldots,\phi_p)$.

- Updating the tuning parameter ζ :

$$\log(\zeta^{(t+1)}) = \log(\zeta^{(t)}) - \gamma^{(t+1)}\left(\alpha - \alpha^*\right),$$

  where $\alpha^* = 0.24$ is the ideal acceptance probability and the step size is $\gamma^{(t)} = \min\left(500^{-\frac{1}{2}}, t^{-\frac{1}{2}}\right)$.

(iv) Sample $\eta_{jk}^{-1} \mid \beta_{jk},\phi,\tau$ independently from $Inv-Gaussian\left(1, \frac{\phi_j\tau_k}{|\beta_{jk}|}\right)$. The Inverse Gaussian distribution is defined by the density function $f\left(x;\mu,\theta\right) =$

$\left(\frac{\theta}{2\pi x^3}\right)^{\frac{1}{2}} \exp\left[-\frac{\theta(x-\mu)^2}{2\mu^2 x}\right],\ x > 0.$

(v) Sample $\Psi \mid B \sim Inv - Wishart\,(v_0 + n, S_0 + S)$.

2.4 Simulation study

Here we implement simulation studies to evaluate the performance of our methodology. In addition to our MONG, MOHS, and MODL methods, we consider the following competitors:

- **Naive Normal-gamma Model:** To assess the utility of sharing the local parameters across all response variables, we consider an approach that fails to make use of this information by independently placing a NG prior on the vector of regression coefficients $(\beta_{1k}, \beta_{2k}, \cdots, \beta_{pk})$ for each model k. This naive model is unable to borrow strength across models to inform the shared level of sparsity. To that end, $\beta_{jk} \sim N\,(0, \lambda_{jk}\tau_k^2)$, where all λ_{jk} are independent from $Gamma\left(c, \frac{1}{c}\right)$. The rest of the model is unaffected.

- **Naive Horseshoe Model:** Similar to the naive NG model, we consider applying a horseshoe prior independently for each response. In this case, $\beta_{jk} \sim N\left(0, \lambda_{jk}^2\tau_k^2\right)$, with all λ_{jk} independently from $C^+(1)$.

- **Naive Dirichlet-Laplace Model:** We also consider a naive version of DL prior. To that end, we let $\beta_{jk} \sim DE\,(\phi_{jk}\tau_k)$. Here, independent local shrinkage parameters are drawn for each response model k: $\phi_k = (\phi_{1k}, \phi_{2k}, \cdots, \phi_{pk}) \sim Dir(a, a, \ldots, a)$.

- **No Shrinkage Model:** As a baseline that does not favor any variable selection, we consider a basic conjugate prior model. For all j, k, $\beta_{jk} \sim N\,(0, 10)$ to provide minimal shrinkage towards zero.

- **Selection Prior (Brown et al., 1998) Model:** As noted in the introduction, this approach constrains each predictor to either be in the model of all K responses or to be excluded from all.

- **MBSP Model (Bai and Ghosh, 2018):** As previously noted, this approach is similar to our MOHS model where the global parameter τ is common across all responses and fixed by asymptotic arguments, rather than estimated from the data. The performance of this model is obtained using their available R package MBSP.

Data are generated from the multi-response linear regression model (2.1) using a design matrix $X^{n\times p}$ whose elements are independently drawn from a standard normal distribution. Then, rows of the response matrix $Y^{n\times p}$ are independently generated from $N_K(X_{i\cdot}B, \Psi)$, where $\Psi_{ij} = 0.5$ if $i \neq j$, and 1 otherwise. We consider $p = 20$ predictors, $K = 10$ response variables, and a sample size of $n = 500$. We generate 100 datasets, and for each dataset and model choice we run the MCMC chain for 90,000 iterations with a burn-in of 10,000 iterations. We measure predictive performance by computing the mean square prediction error (MSPE) using the posterior mean regression coefficients $\hat{B}$ and an independently generated test data set. To assess the accuracy of the regression coefficient estimation, we consider the sum of square errors (SSE). To distinguish between error of over-shrinking relevant signals and under-shrinking non-signals, we partition this SSE into the SSE over the true non-zero β_{jk}s and the SSE for the β_{jk}s that are true zeros. These quantities are determined by the following formulas:

$$MSPE = \frac{1}{Kn_{test}} \sum_{k=1}^{K} \sum_{i=1}^{n_{test}} \left(X_{i\cdot}^{(t)} \hat{B}_{\cdot k} - Y_{ik}^{(t)} \right)^2 \qquad SSE = \sum_{k=1}^{K} \sum_{j=1}^{p} \left(\hat{\beta}_{jk} - \beta_{jk} \right)^2,$$

where n_{test} is the number of observations in the test dataset ($n_{test} = 500$), and $Y^{(t)}$ and $X^{(t)}$ denote respectively the response and design matrices for the test set. We

consider two scenarios for choosing the true regression coefficient structure. First, we consider a simple sparse $B^{(0)}$ matrix (Table 2.1), where each covariate is either important for all responses or has no contribution to the mean of any response. Table 2.3 presents the results for this case.

Comparing each of our multi-outcome models to their respective naive versions, we find reduced MSPE in all cases. While the difference in MSPE between models are relatively minor, there are large improvements in the coefficient estimation. Our shared shrinkage models lead to reduction in total SSE of around 50% when compared to the respective naive version. When looking at the two components of SSE, we see clear improvement in the estimation of the coefficients that are truly zero. That is, by sharing the local parameters across the K outcome models, our model is able to better identify those coefficients that should be aggressively shrunk toward zero. Our proposed model also yields similar level of predictive performance with the selection prior approach (Brown et al., 1998), which is perfectly suited to this choice of $B^{(0)}$.

We note that the model without shrinkage is not competitive due to its large SSE in the zero coefficients. Also the naive DL with $a = \frac{1}{p}$ performs poorly in estimating the non-zero coefficients. Setting $a = \frac{1}{p}$ provides a much stronger level of shrinkage than the $a = 0.5$ case. For the naive DL model, we do see more shrinkage under $a = \frac{1}{p}$ than $a = 0.5$, but by sharing shrinkage information across multiple responses, our MODL model is able to find an acceptable balance in the amount of shrinkage under both choices of a.

Next, we consider a situation that does not have the exact same sparse structure for each response model. There are two important considerations for such a choice. First, in light of our original motivation, we are interested in a more flexible model than those require the same subset of predictors for all responses. We wish to assess the performance of our model in such a case where there are variations in the relevant predictors across models. An alternative motivation is to understand the

impact of misspecification for models that assume the exact same subset of relevant predictors across all outcomes. To that end, the new true coefficient matrix $B^{(1)}$ in Table 2.2 is created by perturbing $B^{(0)}$ so that the true model no longer has exact sparsity across all models. We switch three of the zero coefficients from $B^{(0)}$ to non-zero β_{jk} and also change three non-zero coefficients in $B^{(0)}$ to zero (as denoted in bold). This potentially represents a more realistic scenario where a small subset of predictors impact all responses, but there are some minor deviations from this general rule.

The results for this simulation settings are reported in the Table 2.4 and are generally similar to the previous analysis. As would be expected, the gap between the shared shrinkage and the naive approaches is somewhat narrowed, but the proposed approaches continue to show lower MSPE and lower SSE than their naive counterparts in all cases. Hence, even if there are some differences in which predictors are relevant across models, sharing shrinkage information through our common local parameter structure can continue to improve estimation. The selection prior approach and MBSP model also show similar prediction performance, although both have poorer performance in the coefficient estimation relative to our approach. Of particular note, the MBSP has fairly large SSE for the zero signals, indicating a lower level of shrinkage than our proposals. Our model estimates the global parameters from the data to adjust the amount of shrinkage, whereas MBSP fixes τ and is unable to correct for undershrinkage in this data.

In conclusion, our three multi-outcome models perform well in those simulation studies. Using $a = \frac{1}{p}$ in the MODL model may lead to overshrinking, so we typically prefer $a = 0.5$. While the differences between methods are relatively minor, MONG tends to perform best among our proposals.

2.5 Application

We now demonstrate our methodology with the yeast cell cycle data set (Spellman et al., 1998) from the spls package in R. The data was first analyzed by Chun and Keleş (2010) and also by Bai and Ghosh (2018). In this dataset, the response matrix Y contains gene expression data for $n = 542$ genes from an α factor based experiment. Each column of Y corresponds to mRNA levels measured at 7 minute intervals across 2 hours providing a total of $K = 18$ responses. The covariate matrix X contains the binding information for $p = 106$ transcription factors (TFs). In molecular biology, transcription factors are a diverse family of proteins which are involved in the process of transcribing, DNA into RNA. Hence, it is of common interest to identify the most significant TFs that play an important role in gene regulations.

We applied our method to capture those TFs that affect the expression levels across all time points. We perform the analysis using our proposed MONG, MODL, and MOHS models, followed by the three naive models, the no shrinkage model, the selection model (Brown et al., 1998) and the MBSP model (Bai and Ghosh, 2018). Due to over-shrinkage observed in the MODL$\left(a = \frac{1}{p}\right)$ model, we do not consider its performance here. For each case, we run a burn-in for 1000 iterations followed by another 30,000 iterations. We report the MSPE by performing cross validation on 50 data sets for each model to assess the predictive power of each method. For cross validation we randomly assign 80% of observations to the training set to estimate B, and then measured the MSPE using the remaining 20%. We also analyze the full dataset and compute the deviance information criteria DIC as a model comparison measure (Spiegelhalter et al., 2002). DIC is calculated by $DIC = D + 2p_D$, where D is the deviance at the posterior expectation of the parameter values and p_D is the effective number of parameters, and smaller DICs are favored. p_D is calculated as $p_D = E\left\{D\left(B, \Psi|Y\right)\right\} - D(\hat{\Psi}, \hat{B})$. Table 2.5 shows the MSPE, the deviance at the

posterior expectation of the parameter values (D), the effective number of parameters (p_D), and the deviance information criteria (DIC) of the yeast cycle data for each model.

The MODL($a = 0.5$) choice yields the lowest prediction error among our models. Consistent with the simulation study, each of the multi-outcome approaches have smaller MSPE than their respective naive counterparts. The MONG and MOHS model also yield a lower mean square prediction error by slightly outperforming the selection prior model.

When using the competitor MBSP model, the prediction error is 0.786, scoring lowest among all approaches. It appears that for this particular data application, using a fixed value of τ performs slightly better than our methods which require estimating $K = 18$ global parameters. However, as noted in the simulation study, this is not always the case, and worse performance may result. Finally, we note that the R package of MBSP model only produces model estimates and not the full set of posterior samples. So we were unable to compute DIC estimates for the MBSP model.

The DIC criteria favors the MONG and MOHS models. When considering the effective number of parameters, we see that these models estimate a much sparser regression coefficient matrix than MODL. When comparing DIC between the shared shrinkage and naive models, we again see that our proposals consistently dominate their counterparts that fail to share variable selection information between responses. The selection approach from Brown et al. (1998), which requires a common set of predictors for all models performs poorly with respect to DIC. This model places the majority of the posterior probability on models with only 2 or 3 predictors. This excessive sparsity leads to high prediction error, poor model fit, and large DIC.

Based on the results from fitting the full data set, we consider the use of the local parameters as a marker of variable importance. Figure 1 graphically displays

these parameters for each of the multi-outcome models. Based on the MONG results, we would consider those covariates with $\hat{\lambda}_j > 1$ as evidence of a strong effect across all response models. This criterion selects 8 important TFs: SWI5, SWI6, NDD1, ACE2, STE12, HIR1, GAT3, MBP1. The 8 predictors with the largest $\hat{\lambda}_j$ in the MOHS model corresponds to the same 8 TFs, indicating robustness in the predictor weights across the model variations. Consistent with its large p_D indicating less sparsity, the MODL choice demonstrates much less separation between large and small ϕ_j and consequently less shrinkage/sparsity in the $\hat{B}$ matrix. For this MODL case, distinguishing important predictors based on the value of the local parameters will not be effective.

2.6 Conclusion and discussion

In this paper, we have proposed a general strategy of variable selection in the multivariate regression model by sharing common local parameters across all of the response variables. We have demonstrated our approach using the Normal-gamma, Dirichlet-Laplace and horseshoe priors. Based on the results from simulation studies and the analysis of data from an mRNA experiment, we have demonstrated the utility of our approach in comparison to alternatives. Our approaches are found to be superior in terms of both predictive performance and parameter estimation. In general, we recommend the use of the MONG version of our model as it displayed consistently strong behavior across all empirical experiments, although the MODL and MOHS also performed well.

Regarding computational comparisons between our methods, the MOHS model tends to run fastest as all of its sampling distributions are conditionally conjugate. While slightly slower, MONG has comparable computational time for a fixed number of iterations. However, the MODL model tends to be computationally slower due to the sampling of pK data augmentation parameters η_{jk}. Moreover, as noted in

Section 2.3, the mixing in this algorithm tends to be slower due to the multivariate MH sampling of $\phi = (\phi_1, \ldots, \phi_p)$. While our adaptive step is generally effective here, further algorithmic improvements may be possible here in future research.

2.7 Tables and Figures

Table 2.1: True $B^{(0)}$ regression coefficient matrix in first simulation study.

$$B^{(0)} = \begin{pmatrix} 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 \\ -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 \\ 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Table 2.2: True $B^{(1)}$ regression coefficient matrix in second simulation study.

$$B^{(1)} = \begin{pmatrix} 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 & 2.0 \\ -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 & -3.0 \\ 1.0 & 1.0 & \mathbf{0} & 1.0 & \mathbf{0} & 1.0 & 1.0 & 1.0 & 1.0 & 1.0 \\ 0.0 & 0.0 & \mathbf{0.5} & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \mathbf{0.3} & 0.0 & 0.0 & 0.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & \mathbf{1.5} & 0.0 & 0.0 & 0.0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.3 & 0.3 & 0.3 & 0.3 & \mathbf{0} & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{pmatrix}$$

Table 2.3: Prediction and estimation results from simulation study with $B^{(0)}$.

Models	MSPE	SSE		
		All β	$\beta \neq 0$	$\beta = 0$
MONG	1.028	0.097	0.095	0.002
MODL($a = 0.5$)	1.029	0.098	0.089	0.009
MODL($a = 1/p$)	1.032	0.125	0.112	0.013
MOHS	1.030	0.113	0.099	0.014
Naive NG	1.040	0.205	0.167	0.038
Naive DL($a = 0.5$)	1.036	0.176	0.104	0.073
Naive DL($a = 1/p$)	1.079	0.564	0.547	0.016
Naive Horseshoe	1.040	0.203	0.111	0.092
No shrinkage	1.059	0.416	0.080	0.337
Selection prior	1.028	0.134	0.134	0.000
MBSP model	1.029	0.104	0.092	0.012

Table 2.4: Prediction and estimation results from simulation study with $B^{(1)}$.

Models	MSPE	SSE		
		All β	$\beta \neq 0$	$\beta = 0$
MONG	0.976	0.113	0.095	0.028
MODL($a = 0.5$)	0.976	0.118	0.085	0.033
MODL($a = 1/p$)	0.978	0.127	0.097	0.030
MOHS	0.978	0.132	0.098	0.034
Naive NG	0.978	0.131	0.091	0.040
Naive DL($a = 0.5$)	0.980	0.158	0.083	0.075
Naive DL($a = 1/p$)	0.994	0.283	0.251	0.032
Naive Horseshoe	0.982	0.177	0.083	0.094
No shrinkage	1.005	0.416	0.080	0.336
Selection prior	0.979	0.139	0.090	0.050
MBSP model	0.978	0.146	0.080	0.066

Table 2.5: Cross-validation prediction error and model comparison statistics for yeast cell cycle data.

Models	MSPE	D	p_D	DIC
MONG	0.833	15580	370	16321
MODL($a = 0.5$)	0.814	14077	1299	16676
MOHS	0.841	15683	318	16320
Naive NG	0.987	16594	148	16890
Naive DL($a = 0.5$)	0.907	13990	1430	16851
Naive DL($a = 1/p$)	0.872	15117	733	16584
Naive HS	0.864	14706	827	16361
No shrinkage	0.971	13453	2131	17716
Selection prior	0.845	17425	257	17940

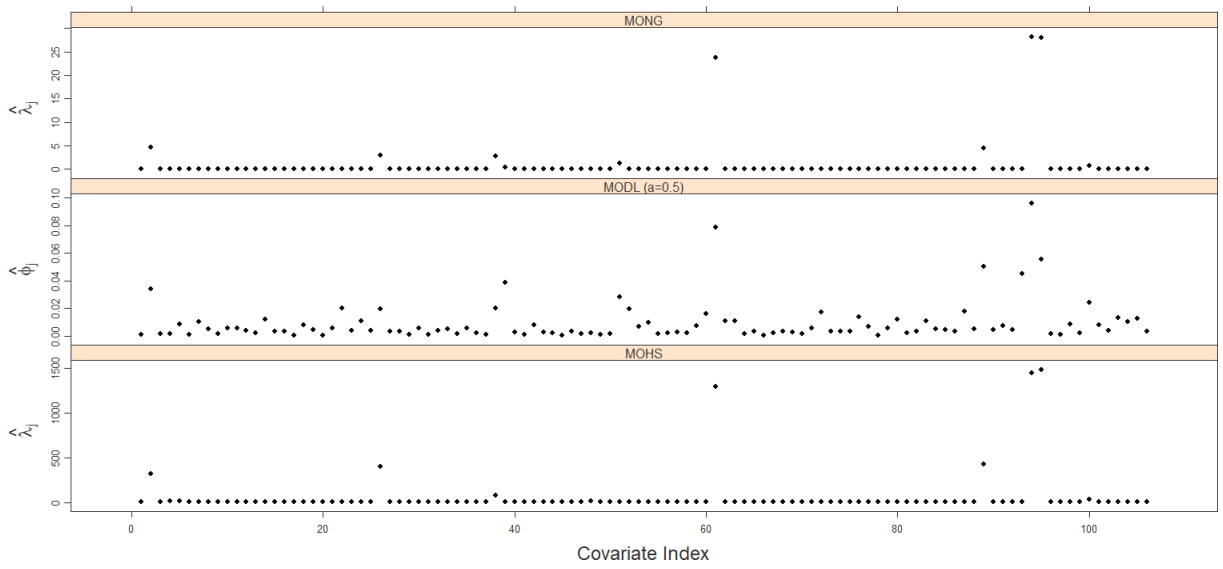

Figure 2.1: Estimated local parameter $\left(\hat{\lambda}_j \text{ or } \hat{\phi}_j\right)$ across all predictors in the three multi-outcome regression analyses for the yeast cell data.

CHAPTER 3

A BAYESIAN HIERARCHICAL SPARSE FACTOR MODEL FOR SIMULTANEOUS COVARIANCE ESTIMATION

3.1 Introduction

In the analysis of multivariate data, the estimation of the covariance matrix is always one prime interest. However, when data consist of multiple groups, each may be determined by its own covariance matrix. In this work, we consider data that consist of M groups, where the covariance matrix of group m is Ω_m $(m = 1, 2, \ldots, M)$. Our interest is in developing methodology to estimate this collection $\{\Omega_1, \Omega_2, \ldots, \Omega_M\}$. When faced with this scenario, it is not uncommon for the analyst to assume equality across all Ω_ms, but this will lead to erroneous inference if there are truly differences across the covariances. Conversely, estimating each Ω_m without sharing information across all groups will lead to inefficient estimation if there are common structures shared across groups. Hence, developing a reasonable method for borrowing strength across groups in the simultaneous covariance estimation problem is paramount for obtaining trustworthy inference.

In the literature of simultaneous covariance estimation, principal component methods are a well-established approach. Flury (1984) developed a method with common eigenvectors to estimate the covariance matrices by considering $\Omega_m = Q\Gamma_m Q$, where Ω_m is the $p \times p$ covariance matrix for the m^{th} group, Q is the $p \times p$ orthogonal matrix of eigenvectors that are shared across all groups and Γ_m is the diagonal matrix of positive eigenvalues specific to group m. Later, Flury (1987) extended this to the

"partial common principal component model" by assuming q ($q < p$) common eigenvectors across all Ω_ms, and the remaining eigenvectors are group-specific. Boik (2002) broadened the idea to a more general model by sharing the eigenvectors between some or all groups. Hoff (2009) also developed a hierarchical Bayesian model that shrinks the eigenvector matrix of each group across the population by using a shrinkage prior on the matrix of eigenvectors. Besides this usual spectral decomposition, Manly and Rayner (1987) and Barnard et al. (2000) proposed decomposing the covariance matrix in terms of the standard deviation matrices (S) and correlation matrices (R), i.e: $\Omega_m = S_m R_m S_m$, and assumed R and S are independent and the correlation matrices are the same across the groups.

In the context of longitudinal data, there are additional methods based on the modified Cholesky decomposition of the covariance matrix (Pourahmadi, 1999). Pourahmadi et al. (2007) highlighted on computational advantages and fundamental differences of the unconstrained parameterization of the Cholesky decomposition for modelling several covariance matrices simultaneously in comparison to traditional eigenvalue or variance-correlation decomposition. Unlike the spectral decomposition and variance-correlation decomposition, the units that appear in the lower triangular matrix, termed as general autoregressive parameters (GARP) of the Cholesky decomposition are always unconstrained and hence involves unconstrained maximization techniques for computing maximum likelihood estimates. McNicholas and Murphy (2010) considered Gaussian mixture models in order to propose a model-based clustering framework for longitudinal data, where the modified Cholesky decompositions of the group covariance matrices are considered to have commonalities across all groups. Gaskins and Daniels (2012) proposed a family of nonparametric priors based on Dunson et al. (2008)'s matrix stick-breaking process. Their method uses the parameters from modified Cholesky decomposition which includes GARP and the innovation variances (IV) to parametrize the covariance matrix for each group. Additionally, this

methodology sets some parameters of the Cholesky decomposition to zero to provide a lower-dimensional structure for the covariance matrix. Later, Gaskins and Daniels (2016) proposed a related approach that partitions the collection of groups into sets with common conditional distributions.

As an estimator for a single covariance matrix, latent factor models traditionally play an important role in modeling multivariate dependence structures in the behavioral sciences. The essential purpose of factor analysis is to describe as the underlying covariance relationship between many variables in terms of a few unobserved random quantities, called *factors*. Consider, a situation where a researcher assembled a moderate to a large number of predictors for an analysis. In general a p-dimensional predictor variable has $p(p-1)/2$ pairwise correlation. However, when p is moderately large it is very difficult to summarize and interpret all pairwise correlations together. The factor model assumes that complex correlation structure can be explained by some latent linear combinations of fewer variables, leading to a reduction in dimension. These underlying unobserved random variables are termed as latent *factors*. This is a parsimonious model. As the number of latent factors $K \ll p$, therefore instead of p^2 terms, we need to deal with only $p(K+1)$ terms. Further, it makes interpretation simpler if variables are grouped by their underlying correlation structure. For example, if we have test scores from different subjects of a group of student, we may consider as mathematics, vocabulary, physics score as "intelligence" factor, weight, BMI, energy level as "physical fitness" factor and sociability, gregariousness, lack of shyness as "psychological" factor.

Selection of the appropriate number of factors is a key issue in such models, and traditional model selection criteria such as AIC or BIC are standard choices. In a Bayesian factor model Lopes and West (2004) considered the number of factors itself to be an unknown parameter. They introduced a customized reversible jump Markov chain Monte Carlo (RJMCMC) algorithm to sample from the model with a variable

number of factors. Additionally, Ghosh and Dunson (2009) proposed an efficient parameter expansion algorithm to improve the computational efficiency of Bayesian factor models. Also, Bhattacharya and Dunson (2011) have applied a multiplicative gamma process shrinkage prior to Bayesian latent factor models to model a sparse covariance matrix for high-dimensional data by using infinite number of factors. Due to its use in several applied areas such as pattern recognition, financial time series modeling, bioinformatics and computer vision, the theory of factor models analysis has received huge attention. However, the use of the latent factor model is relatively uncommon in the context of estimating the multiple covariance matrices.

In this article, we introduce a novel approach for the estimation of multiple covariance matrices using a hierarchical Bayesian latent factor model. In section 3.2 we explain our methodology including a full model specification and a discussion of our computational estimation procedure. Section 3.3 describes a number of simulation studies to explore the performance of our model. In section 3.4, we have applied our method on Letter recognition data and compared the performance with other competitor models. We conclude with a brief discussion in section 3.5.

3.2 Bayesian Sparse Hierarchical Factor (BaSH-F) Model

3.2.1 Model & Prior specification

Consider M groups containing n_m observations in group m, and let $N = \sum_m n_m$ be the total number of observations. We also let $Y_{mi} = (Y_{mi1}, Y_{mi2}, \cdots, Y_{mip})$ represents the p-dimensional sample for the i^{th} observation $(i = 1, 2, \cdots, n_m)$ of the m^{th} group $(m = 1, 2, \cdots, M)$. Without loss of generality, we let the mean vector for each group be zero. We assume that Y_{mi} is multivariate normally distributed: $Y_{mi} \sim MVN_p(0, \Omega_m)$, $i = 1, 2, \cdots, n_m; m = 1, 2, \cdots, M$. We further assume that each covarinace matrix can be decomposed using the usual factor model $\Omega_m = \Lambda_m \Lambda_m^T + \Sigma_m$.

Here Λ_m is a $p \times K$ matrix with (j,k) element λ_{mjk} and $\Sigma_m = diag(\sigma^2_{m1}, \sigma^2_{m2}, \cdots, \sigma^2_{mp})$, where all $\sigma^2_{im} > 0$.

This model may be equivalently motivated by introducing a vector of K latent factor values for each observation. To that end we let $\eta_{mi} = (\eta_{mi1}, \eta_{mi2}, \cdots, \eta_{miK})^T \sim MVN_K(0, I_K)$ be the K factor scores of observation i in group m, and consider

$$Y_{mi} = \Lambda_m \eta_{mi} + \epsilon_{mi}, \tag{3.1}$$

where $\epsilon_{mi} = (\epsilon_{mi1}, \epsilon_{mi2}, \cdots, \epsilon_{mip}) \sim MVN(0, \Sigma_m)$ is a vector of error terms. Marginally over η_{mi} and ϵ_{mi}, we again obtain $Y_{mi} \sim MVN(0, \Omega_m)$. The benefit of this approach is that we may consider Λ_m as a matrix of regression coefficients (with η_{mi} as predictors) and the σ^2_{mj}s as the regression variances, which facilitates posterior sampling. In this work, when we refer to the factor loadings we mean the regression coefficients λ_{mjk}, not the correlation between Y_{mij} and η_{mik}.

The general idea of our methodology is to consider the commonalities between the factor loading matrices Λ_m across the M groups by shrinking λ_{mjk}, the $(j,k)^{th}$ element of Λ_m, towards a global value ω_{jk} $(j = 1, 2, \cdots, p; k = 1, 2, \cdots, K)$ shared across all groups. The $\mathcal{W} = (\omega_{jk})$ matrix can be thought of as representing the overall relationship/factor loadings across all groups in the population. To help control the complexity of the model and improve interpretation, it is common to assume sparsity in the factor loading matrix (Carvalho et al., 2008). For instance, if $\lambda_{mjk} = 0$, this implies that the k^{th} factor is not associated with the j^{th} response. Here, we assume that the sparsity in Λ_m is a feature shared across all groups and the $\mathcal{W}$ matrix. To that end, we introduce the parameter Z_{jk}. If $Z_{jk} = 0$, then response j is unaffected by factor k in all groups, and $0 = \omega_{jk} = \lambda_{1jk} = \lambda_{2jk} = \cdots = \lambda_{mjk}$. If $Z_{jk} = 1$, then the factor k loads on response j, and ω_{jk} and λ_{mjk}s $(m = 1, 2, \cdots, M)$ are non-zero. The hierarchy that describes the distribution of $(Z_{jk}, \omega_{jk}, \lambda_{mjk})$ is

$$\begin{aligned} \lambda_{mjk} &\sim N(\omega_{jk}, Z_{jk}\nu_m^2), \\ Z_{jk} &\sim Ber(\pi), \\ \omega_{jk} &\sim N(0, Z_{jk}\tau^2). \end{aligned} \tag{3.2}$$

Note that the variance parameter ν_m^2 determines how similar Λ_m is to the global $\mathcal{W}$ matrix. This shrinkage parameter is group-specific allowing some groups to be less similar to the overall structure. We assume the distribution of the standard deviation ν_m to be $HC(\theta)$, where $HC(\theta)$ represents the half-Cauchy distribution with scale of θ and density $f(x) = \frac{2\theta}{\pi(\theta^2+x^2)}, \quad x > 0$. This prior encourages shrinkage towards zero by imposing substantial mass near zero while its thick tail simultaneously captures the strong signals (Gelman et al., 2006). Further, we consider an inverse gamma $IG(1,1)$ prior on θ, the median of the ν_ms. The pdf for the $IG(\alpha, \beta)$ distribution is $f(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}}$, $x > 0$.

The parameter π in the distribution of the Z_{jk}s controls the overall level of sparsity in the factor loading matrices. Values near zero will produce highly sparse Λ_m. The prior for π is $Beta(a_\pi, b_\pi)$ with $a_\pi = b_\pi = 1$ as default choices. We place an $IG(\alpha, \beta)$ prior on τ^2, the variance of the non-zero ω_{jk}s. We use $\alpha = \beta = 1$ in our analysis. For the error variance terms in the Σ_m matrices, we take $\sigma_{mj}^2 \sim IG(c, d)$ with $c = d = 0.1$ as default choices for the hyperparameters. We refer to our approach as the Bayesian Sparse Hierarchical Factor Model (BaSH-F) for simultaneous covariance estimation.

3.2.2 Posterior Computation

We adapt the usual Markov Chain Monte Carlo(MCMC) methods for factor model to develop a posterior computation scheme. The necessary sampling steps are described below.

(i) First sample the probability $\pi|Z$ from $Beta(a_\pi + \sum_{j,k} Z_{jk}, b_\pi + pK - \sum_{j,k} Z_{jk})$.

(ii) For each (j,k) pair, we update Z_{jk}, ω_{jk} and λ_{mjk} blockwise. That is, we update $Z_{j,k}$ marginally over ω_{jk} and the λ_{mjk}s. Then we update ω_{jk} conditionally on the new Z_{jk} and marginally over the λ_{mjk}s. Finally, we sample each of λ_{mjk}s given the ω_{jk} and Z_{jk}.

- First update Z_{jk} from $Ber(p^*)$, where $p^* = \frac{B}{A+B}$, where

$$A = \pi \prod_{m=1}^{M} \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk} = 0\right),$$

$$B = (1-\pi) \prod_{m=1}^{M} \left[\int \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk}\right) f\left(\lambda_{mjk}|\omega_{jk}\right) d\lambda_{mjk} \right] f\left(\omega_{jk}|Z_{jk} = 1\right) d\omega_{jk}.$$

A is the likelihood when j^{th} response is not loaded in the k^{th} factor across all groups. That means $\omega_{jk} = 0$ and hence $\lambda_{mjk} = 0$, for all m. In a similar way, B defines the likelihood when j^{th} response is loaded in the k^{th} factor, i.e., $Z_{jk} = 1$ and hence $\omega_{jk} = 1$, for all m. p^* defines the posterior probability of j^{th} response being loaded in the k^{th} factor.

Simplification yields $p^* = (1 + e^c)^{-1}$ where

$$\begin{aligned} c = {} & \log(1-\pi) - \frac{1}{2}\log(2\pi\tau^2) - \log(\pi) \\ & - \frac{1}{2}\left\{ \log\left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2 + \sigma_{mj}^2 / \sum_{i=1}^{n_m} \eta_{mik}^2} \right) \right\} \\ & + \frac{1}{2} \sum_{m=1}^{M} \left\{ \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} \right)^2 \right\} + \\ & \frac{1}{2} \left\{ \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2 + \sigma_{mj}^2 / \sum_{i=1}^{n_m} \eta_{mik}^2} \right)^{-1} \left(\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2 \nu_m^2} \right)^2 \right\} \\ & - \frac{1}{2} \sum_{m=1}^{M} \left\{ 2\log\nu_m + \log\left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right) - \log(2\pi) \right\}, \end{aligned} \tag{3.3}$$

with $e_{mij} = Y_{mij} - \sum_{\substack{l=1 \\ l \neq k}}^{K} \lambda_{mjl}\eta_{mil}$, the residual of the j^{th} response for the i^{th} observation excluding the role of factor k. The necessary derivations are shown in Appendix B.

- If $Z_{jk} = 0$, then $\omega_{jk} = 0$. Otherwise, sample ω_{jk} from $N(\mu_w^*, \sigma_w^{*\,2})$:

$$\mu_\omega^* = \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2 + \sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right)^{-1} \left\{\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2}\right)\right\}$$

$$\sigma_\omega^{2\,*} = \left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\left\{\frac{1}{\nu_m^2 + \sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right\}\right)^{-1}.$$

- For each $m = 1, 2, \cdots, M$, we set $\lambda_{mjk} = 0$ if $\omega_{jk} = 0$, otherwise update λ_{mjk} from $N\left(\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2}\right), \left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\right)$.

(iii) For all $i = 1, 2, \cdots, n_m; m = 1, 2, \cdots, M$, update η_{mi} from $MVN_K\left(\left(I_K + \Lambda_m^T\Sigma_m^{-1}\Lambda_m\right)^{-1}\Lambda_m^T\Sigma_m^{-1}Y_{mi}, \left(I_K + \Lambda_m^T\Sigma_m^{-1}\Lambda_m\right)^{-1}\right)$.

(iv) To obtain an efficient conditionally conjugate sampling distribution for ν_m, we adapt Makalic and Schmidt (2016)'s sampling algorithm by introducing data augmentation variables from an inverse gamma distribution. Marginally $\nu_m \sim HC(\theta)$, and we can equivalently write hierarchically as $\nu_m^2|a_m \sim IG\left(\frac{1}{2}, \frac{1}{\theta a_m}\right)$ and $a_m \sim IG\left(\frac{1}{2}, 1\right)$. The conditional sampling density of ν_m^2 is $IG\left(\frac{\sum_{j,k} Z_{jk}}{2} + \frac{1}{2}, \frac{1}{\theta a_m} + \frac{1}{2}\sum_{j,k}\left(\lambda_{mjk} - \omega_{jk}\right)^2\right)$ and sample a_m from $IG\left(\frac{M+1}{2}, 1 + \sum_{m=1}^{M}\frac{\theta^{-1}}{\nu_m^2}\right)$.

(v) Sample θ from $IG\left(\frac{M+1}{2}, 1 + \sum_{m=1}^{M}\frac{a_m^{-1}}{\nu_m^2}\right)$.

(vi) Update τ^2 from $IG\left(\alpha + \frac{\sum Z_{jk}}{2}, \beta + \sum_{j,k}\frac{Z_{jk}\omega_{jk}^2}{2}\right)$.

(vii) For $j = 1, 2, \cdots, p$ and $m = 1, 2, \cdots, M$, we sample σ_{mj}^2 from $IG\left(\frac{n_m}{2} + c, d + \frac{1}{2}\sum_{i=1}^{n_m}\left(Y_{mij} - \sum_{k=1}^{K}\lambda_{mjk}\eta_{mik}\right)^2\right)$.

In latent factor model, often MCMC samples get stuck in local modes due to model complexity and do not mix very well. To resolve this complications, we

consider running multiple MCMC chains and a swapping step inside the sampling algorithm. After completing step (ii) in the sampling algorithm, we randomly choose a j^{th} response and fix that j. For this fixed j^{th} response we propose a swap between the positions of two elements λ_{mjk_1} and λ_{mjk_2} for all m, where $k_1 \in \{k : Z_{jk} = 0\}$ and $k_2 \in \{k : Z_{jk} = 1, \}$, i.e we load the j^{th} response in the k_1^{th} factor if this is not loaded originally and unload it from the ${k_2}^{th}$ factor. Next we compute the MH probability using the following equation 3.4

$$\phi = \frac{L\left(Y|\Lambda^*, \Sigma\right)}{L\left(Y|\Lambda, \Sigma\right)}, \tag{3.4}$$

where Λ^* is the updated Λ matrix after swapping the position of $\lambda_{mjk_1}^{th}$ and $\lambda_{mjk_2}^{th}$ position for all m, j. We update $\Lambda = \Lambda^*$ and the corresponding Z, $\mathcal{W}$ matrix with probability ϕ or retain at the current Λ matrix. To ensure a better mixing we attempt this swapping steps 5 times within each MCMC chain.

3.2.3 Determination of K

The choice of K, the number of factors to use, can have a large result on the effectiveness of our method. Using too few factors will lead to inconsistent estimation, while estimating with too large of K will produce inefficient estimators. Following the approach of Akaike (1987), we apply a model selection approach to determine the value of K. To this end, we run the factor model for a small number of choices of K and calculate the deviance information criteria (DIC) to compare the fits for each choice. In the Bayesian context, DIC is a more natural approach than AIC or BIC as it automatically determines the model complexity without counting the number of parameters. The deviance of the K-factor model at the parameter value Ω is given by $D_K\left(\Omega\right) = -2L_K\left(\Omega|Y\right)$, where $L_K\left(\Omega|Y\right)$ denotes the likelihood function using the usual $Y_{mi} \sim MVN_p(0, \Omega_m)$ with $\Omega_m = \Lambda_m\Lambda_m^T + \Sigma_m$. The posterior expected deviance

is the average value of the deviance with respect to the posterior. We use the Bayes estimator of the covariance matrix of group m, $\hat{\Omega}_m = E\left[\Omega_m^{-1}\right]^{-1}$ (Yang and Berger, 1994), and the DIC for K-factor model is given by $DIC_K = D_K\left(\hat{\Omega}\right) + 2p_K$. Here, p_K is the effective number of parameters and calculated from $p_K = E\{D_K(\Omega)|Y\} - D_K(\hat{\Omega})$ (Spiegelhalter et al., 2002). The selected value for the number of factors $\hat{K}$ is chosen as the K with the smallest DIC. The estimates $\{\hat{\Omega}_1, \hat{\Omega}_2, \cdots, \hat{\Omega}_M\}$ are taken to be the posterior estimators for the MCMC chain with $\hat{K}$ factors.

We considered yet an another approach using log pseudo marginal likelihood ($LPML$) as the model selection statistic. $LPML$ is based on considering the predictive distributions $p\left(Y_{mi}|Y_{-mi}\right) = \int p\left(Y_{mi}|\Omega\right)\pi\left(\Omega|Y_{-mi}\right)d\Omega$ for all (m, j). We combine each of the predictive densities to form $LPML_k$ for the K-factor model as $LPML_K = \sum_{m,i} \log\, p(Y_{mi}|Y_{-mi})$ (Gelfand and Dey, 1994). To avoid generating posterior samples from $\pi\left(\Omega|Y_{-mi}\right)$ for each (m, i) pair, we use an importance sampling approach using the following equation

$$p(Y_{mi}|Y_{-mi}) = \left[\frac{1}{G}\sum_{g=1}^{G} p(Y_{mi}|\Omega^{(g)})^{-1}\right]^{-1} \quad i = 1, 2, \cdots, N; m = 1, 2, \cdots, M,$$

where G is the total number of posterior samples from the full posterior with all observations. Using the $LPML$ criteria, we select $\hat{K}$ to be the value with the largest $LPML$ and take estimators from this model.

3.3 Simulation

We have implemented a number of simulation studies to evaluate the performance of our methodology. In addition to our BaSH-F model, we have also considered some competing models to check the performance of our model. The first two competitors are simplifications of the BaSH-F model:

- **No Shrinkage Model:** To access the utility of sharing information across all

groups in our model, we consider a setting where no information is shared, but sparsity is still included. To this end, there are no ω_{jk} global parameters in our model and each group has its own set of sparsity indicators Z_{mjk}. We replace equation (3.2) with $Z_{mjk} \sim Ber(\pi_m)$ and $\lambda_{mjk} \sim N(0, Z_{mjk}\nu_m^2)$.

- **No Sparsity Model:** To test the utility of the spike-and-slab model on ω_{jk}, we consider a version of our hierarchical factor model without sparsity. We share information about an underlying factor structure across groups, but constrain all $Z_{jk} = 1$. Equivalently, we swap equation (3.2) with $\omega_{jk} \sim N(0, \tau^2)$ and $\lambda_{mjk} \sim N(\omega_{jk}, \nu_m^2)$. The rest of the model is unaffected.

- **Hoff (2009) Model:** In this method, the covariance matrix is decomposed through the eigenvalue decomposition, i.e. $\Omega_m = U_m V_m U_m^T$, where U_m is the eigenvector matrix and V_m be the eigenvalue for the m^{th} group. Then a shrinkage prior is applied on the eigenvector matrices to pull the information across all groups for estimating covariance.

- **Hierarchical Inverse-Wishart (IWH) Model:** As a slightly more sophisticated competitor, we consider a hierarchical model based on conjugate inverse Wishart distribution.

$$\Omega_m \sim InvW(\delta, \delta\Psi)$$

$$\Psi \sim Wishart(p, \frac{1}{p} I_p) \quad \delta \sim Unif(p, N).$$

  Note that since $E(\Omega_m) = \frac{\delta\Psi}{\delta - p - 1}$, all Ω_m are pulled toward a common/overall covariance matrix based on Ψ and δ. The amount of shrinkage is determined by the degrees of freedom δ and a higher δ indicates more shrinkage.

- **Independent Inverse-Wishart (IW) Model:** This is a naive approach where each covariance matrices comes independently from the conjugate $InvW(p+2, I_p)$ prior. Hence, we do not allow any sparsity or sharing of information across

groups in the model. This competitor represents the naive assumption that all covariance matrices are independent and does not share any information between groups.

We will consider a variety of data generating models, and for each parameter specification we generate 200 datasets. For each model, we run the Gibbs sampler for 3 different chains with 50,000 iterations for each chain. After the first 10,000 iterations, we retain every 10^{th} iteration, providing 4000 iterations from each chain to use for inference. To measure the accuracy of our estimators, we consider the loss function from Gaskins and Daniels (2016) that uses a weighted average of the log-likelihood loss for each group, with weights proportional to group's sample size n_m. The formula is given by

$$\mathcal{L}\left(\Omega, \hat{\Omega}\right) = \sum_{m=1}^{M} \frac{n_m}{N} \left[tr(\Omega_m^{-1}\hat{\Omega}_m) - \log|\Omega_m^{-1}\hat{\Omega}_m| - p \right],$$

where $\hat{\Omega}_m = \left[E_{\Omega|y}\Omega_m^{-1}\right]^{-1}$. We calculate the risk estimates as the average loss $\mathcal{L}\left(\Omega, \hat{\Omega}\right)$ over the 200 datasets. For the factor models, we consider 3 methods of choosing the K parameter: DIC, $LPML$ and an oracle estimates that uses the true value of K. We consider the following data generating models:

- **Case 1:**

  First, we generate data consistent with our model specification. We consider the number of factors $K = 5$, the response dimension $p = 12$, number of groups $M = 3$, and total number of observations $N = 300$ with 100, 50 and 150 observations in 1^{st}, 2^{nd} and 3^{rd} group respectively. We consider $\pi = 0.5$, $\tau^2 = 1$, $\nu_m^2 = 0.2$ for all m and $\sigma_{mj}^2 = 1$ for all (j, m). These values are used to generate one set of true covariance matrices $\{\Omega_1, \Omega_2, \cdots, \Omega_M\}$, and all 200 generated datasets are simulated from this set of parameters. When estimating the true number of factors in the BaSH-F model, the no sparsity model and the

no shrinkage model, we run the model from $K = 3$ to $K = 8$.

- **Case 2: Bigger sample size**

  In this case, we double the sample size to study the performance of our model for a larger sample. The total number of observations is $N = 600$ with 200, 100 and 300 observations in 1^{st}, 2^{nd} and 3^{rd} group respectively. All other settings are the same as in case 1.

- **Case 3: Much larger sample size**

  In this case, we considered a larger sample size than case 2. The total number of observations is $N = 1200$ with 400, 200 and 600 observations in 1^{st}, 2^{nd} and 3^{rd} group respectively. All other settings are the same as in case 1.

Table 3.1 and Figure 3.1 shows the results for case 1 and case 2. Note that we re-scaled the risk in each case so that Independent IW has risk 1. In both case 1 and case 2 our BaSH-F model outperforms all other competing models in both the *Oracle* version and when K is chosen by DIC and $LPML$. In fact, the DIC does equivalently well as the *Oracle* model. This indicates that it effectively finds a choice of K that produces good estimators. In both cases, the Hoff model yields better estimation in comparison to no sparse and no shrinkage model. Independent IW model shows poor performance as no information is shared between all groups. But the IWH model yields better result as δ determines the amount of shrinkage from the data.

In the latent factor model, we also need to inspect the model selection accuracy. We have studied both DIC and $LPML$ model selection criteria for estimating the number of latent factors. Table 3.2 shows that for case 1 (N=300), DIC captures the true number of factors in most cases. It tends to choose a higher factor model as N increases (case 2 and case 3). Turning to K selection based on $LPML$, this approach shows bad results in all cases and performs worse as N increases. Also parameter

estimation using $LPML$ is not as good as DIC. Hence, thereafter we only consider DIC model selection criteria for estimation.

These results seem to suggest that as the sample size increase DIC may tend to select models that are slightly overly complex (too many parameters). Despite this behavior our models are still quite parsimonious, and more importantly, DIC BaSH-F has equivalent estimation performance on $\{\Omega_1, \Omega_2, \cdots, \Omega_M\}$ as the BaSH-F version with the true K. As estimation of these covariance matrices is our goal (not the selection of the number of factors), small levels of inconsistency in selection of K is not a concern as long as it does not impact estimation accuracy.

We now consider 3 scenarios to investigate the impact of varying levels of sparsity.

- **Case 4: Larger p, K, N**

  In this case, we increase the response dimension and the number of factors in the model. We set $p = 30$ and $K = 10$ for this simulation settings. We also increase the number of groups $M = 5$, a total of $N = 750$ observations and 100, 125, 150, 175, 200 observations in each group respectively. We induce a moderate sparsity in this settings with $\pi = 0.4$. For estimating the true number of factors, we run the model from $K = 8$ to $K = 13$.

- **Case 5: Less sparsity $\pi = 0.7$**

  Under the previous settings as in case 4, we set $\pi = 0.7$, i.e. we induce less sparsity in the data. For estimating the true number of factors, we run the model from $K = 8$ to $K = 13$.

- **Case 6: High sparsity $\pi = 0.2$, higher K**

  In this settings we set the $K = 20$ and $\pi = 0.2$, i.e a model with higher number of factors along with higher sparsity. p and N are set as in case 4. For estimation of the number of factors we run the model from $K = 18$ to $K = 23$.

In high-dimensional scenarios (case 4, case 5 & case 6), BaSH-F model continues to perform better than all competitors. The difference between the risk for the BaSH-F model and the no shrinkage model increases from around 4% (case 1) to around 12% (case 6), indicating the increasing benefit of sharing information as p grows (Table 3.3 and Figure 3.2). In case 4 and case 6, when there is moderate or high sparsity in the model, BaSH-F model performs very well in comparison to other models. The no sparse model has equivalent risk to BaSH-F in case 5 (low sparsity scenario). Hence, BaSH-F can effectively adopt to non-sparse scenarios when needed. Naive independent inverse-Wishart model performs poor throughout all cases. Failing to either share info across group or to incorporate sparsity in Ω_ms, leads to much worse. The Hoff model performs well while dealing with low sparsity scenario (case 5).

- **Case 7: Less similarity** $\nu_m^2 = 0.5$

  In this case, we study how the simulation results behave with more variability between groups. Under our standard simulation settings, we consider $\nu_m^2 = 0.5$, and to maintain the marginal variance in λ_{mjk}, we set $\tau^2 = 0.7$. All other parameters are as in case 1.

- **Case 8: Very low similarity** $\nu_m^2 = 1$

  In this case, we set much more variability between groups by setting $\nu_m^2 = 1$ and set $\tau^2 = 0.2$ to maintain the overall marginal variance in λ_{mjk}. All other parameters are as in case 1.

Table 3.4 and Figure 3.3 shows that, in the situations where the covariance matrices are less similar (case 7 & case 8), BaSH-F continues to perform well and outperforms all other competitive models. The no shrinkage estimates, while we might expect to be doing better, continues to have around 4% higher risk than BaSH-F. This indicates our approach is able to determine how much shrinkage to apply in

each scenarios. Hoff model also have 3-5% higher risk than BaSH-F model in these cases. For both case 7 and case 8, IWH model yields around 10% increase in loss than its performance in case 1. Naive IW model continues to show poor performance as previous cases.

Finally, we explore the performance where the data generating model is not a factor structure.

- **Case 9: Data are generated from a non-factor model**

  We consider a different data generation procedure to see how our model is performs if the underlying data do not belong to any factor model. Here our data come from a hierarchical inverse Wishart model. We generate Y_{mi}, for all m, i from $MVN_p(0, \Omega_m)$, where $\Omega_m \sim IW(p+50, \Psi), m = 1, 2, \cdots, M$. Ψ is chosen as a block diagonal matrix, $\Psi = \left[\begin{array}{c|c|c} A & B & 0^{4\times 4} \\ \hline B & A & B \\ \hline 0^{4\times 4} & B & A \end{array}\right]$, where A is a 4×4 equi-correlation matrix with $\rho = 0.8$ and B is a 4×4 matrix with all elements equal to 0.3.

- **Case 10: Same covariance matrix across all groups**

  Lastly, we consider a situation where the covariance matrix are equal for all groups and the common covariance is not a factor model. We generate data Y_{mi} from $MVN_p(0, \Omega_m)$, where $\Omega_m = \Psi$ (from case 9), $m = 1, 2, \cdots, M$.

Table 3.5 and figure 3.4 shows the results for case 9 and case 10. For both cases, our BaSH-F model outperforms no sparse and no shrinkage model. Unsurprisingly, the IWH model performs best in case 9 when it is the correct model. In this scenario, the Hoff model also does a good shrinking of Ω_m towards the common structure. In case 10 where all covariances are the same, Hoff, IWH and BaSH-F all do well. Independent IW continues to perform poorly in both cases.

3.4 Letter Image Recognition Data Application

3.4.1 Data and Model specification

We consider the letter image recognition data from mlbench package in R to demonstrate our methodology. This data consist of character images based on 20 different fonts. The fonts represent five different stroke styles (simplex, duplex, triplex, complex, and Gothic) and six different letter styles (block, script, italic, English, Italian, and German). Each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli (Frey and Slate, 1991). Each of these stimuli was converted into $p = 16$ primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15. The objective was to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet.

We consider $M = 26$ groups defined by each letter A to Z to perform our methodology. Both in training and test data set, we consider equal number of observation in each group. We conducted our study with three different choices of $n_m = 20$ 40 and 100. We standardize all observations and consider a group specific mean parameter $\mu_m \sim MNV_p(0, 100I_p)$. We perform the analysis using our proposed BaSH-F model and the other competitive models. We do not consider the no sparse model and the no shrinkage model as these were special cases of BaSH-F model. For each case, we run 3 chains with 50,000 iterations in each chain. After the first 10,000 iterations, we retain every 10^{th} providing 4000 iteration from each chain. Based on these 12,000 final samples from the training set, we calculate the true classification rate for the test data set.

Let $Y_i^* = (Y_{i1}^*, Y_{i2}^*, \cdots, Y_{ip}^*)$ represents the p-dimensional test data set for the i^{th} sample ($i = 1, 2, \cdots, N_{test}$). To use the covariance models for classification, we slightly augment the model hierarchy by including a unknown class membership variable $\mathcal{C}_i$

that indicates the letter group that generates observation i. Thus $\mathcal{C}_i \in \{1, 2, \cdots, M\}$ and we assume a discrete uniform prior for $\mathcal{C}_i$. For estimating the true classification rate for test data set, we calculate the probability (3.5) of each sample being in group m using the posterior samples and assign to the group having the highest probability.

$$P\left(\mathcal{C}_i = m | Y_i^*\right) = \frac{1}{G} \sum_{g=1}^{G} \frac{f\left(Y_i^* | \mu_m^{(g)}, \Omega_m^{(g)}\right)}{\sum_{m=1}^{M} f\left(Y_i^* | \mu_m^{(g)}, \Omega_m^{(g)}\right)}, \tag{3.5}$$

Here, $\mu_m^{(g)}$ and $\Omega_m^{(g)}$ are the mean and covariances in the g^{th} MCMC imputation from the analysis of the training set. $\mathcal{C}_i$ is the class membership variable associated with the test observation Y_i^*. It can be easily verified that this is a MCMC estimate of the posterior predictive probability. Since the group sizes in the training data are all equal, $1/M$ is a reasonable estimate of the class probabilities.

The loss for assigning the i^{th} sample to the m^{th} group is estimated using the following formula.

$$\mathcal{L}_{\mathcal{C}} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \sum_{m=1}^{M} \left[I\left(\mathcal{C}_i = m\right) - P\left(\mathcal{C}_i = m | Y_i^*\right)\right]^2.$$

Lower value of $\mathcal{L}_{\mathcal{C}}$ indicates better performance. Finally we compare the predictive accuracy of different models using the log-score of test data Y_i^* (Gaskins, 2019). We define,

$$\mathbb{LS}\left(Y_i^*\right) = \frac{1}{G} \sum_{g=1}^{G} \log f\left(Y_i^* | \mu_{\mathcal{C}_i}^{(g)}, \Omega_{\mathcal{C}_i}^{(g)}\right)$$

For each i^{th} observation in the test set, we calculate the $\mathbb{LS}\left(Y_i^*\right)$ averaging over all posterior samples. Then we take the sum over total number of observations in the test set N_{test} to obtain the log-score values for the test data. This value is basically a log-likelihood of the test data using the predictive distribution from the training data. Consequently, a model that produces a higher log-score is more consistent with respect to the out-of-sample test data. Larger values (less negative) indicate better

models.

Finally we note that there are a variety of other classification methods available in the literature (e.g., neural networks, support vector machines, etc.). However, our main interest is on methodology for modeling covariance matrices between multiple groups, not in the development of classification algorithms. In this example, we seek to understand the impact of various models on $\{\Omega_1, \Omega_2, \cdots, \Omega_M\}$, and so we consider only approaches with multivariate normal models for each group.

3.4.2 Modeling Results

We apply our methodology and other competitor models to this data set with 6 choices of the number of factors $K = 3$ to $K = 8$. We implement DIC model selection criteria to find the number of factors K in the training data set and choose one having the lowest DIC. Table 3.6 contains the predictive accuracy and model selection result for the data application. Throughout the analysis, we see that our BaSH-F model outperforms all other competitors both in terms of classification accuracy and loss estimation for each choice of n_m in the training data set. Hoff model shows 2-7% lower classification accuracy than BaSH-F model. The IWH model yields a lower classification accuracy and higher risk in all cases. Naive IW model performs poorly throughout all cases. Overall, with the increasing number of observation per group, all models tend to perform better.

3.5 Conclusion and discussion

In this project, we proposed a novel approach for simultaneous covariance estimation based on sparse Bayesian factor models. The sparsity pattern was shared across the groups while borrowing strength across non-zero factors. The model has the flavor of a covariance analogue of multilevel mean models in the sense that it also estimates a global (across-group) covariance structure. The number of factors were chosen using

established model selection criteria like the DIC. The simulation studies clearly demonstrated the superiority of our model with respect to a metric that quantifies discrepancy between covariance matrices.

The Bayesian hierarchy allows enough flexibility to adapt to non-sparse scenarios and include the right amount of shrinkage. We applied the model to a classification problem on real data where the group-specific covariance based models were used to discriminate the groups. Here too, BaSH-F proved itself to be competitive and even outperformed classical algorithms like Linear Discriminant Analysis (LDA) in prediction. This clearly points to the necessity of sharing information across groups in the presence of moderate and low sample sizes. For future work, we plan an extension to non-Gaussian response models. Such models can be applied to inferring shared biological networks, a problem of growing importance in current genomics applications. Also we are contemplating an automated way to choose factors by integrating the current model with flexible priors on the number of factors.

3.6 Tables and Figures

Table 3.1: Risk Estimates for Case 1, Case 2 and Case 3

Model	Model selection criterion	$\mathcal{L}\left(\Omega_m, \hat{\Omega}_m\right)$		
		Case 1	Case 2	Case 3
BaSH-F	DIC	0.478	0.526	0.585
	LPML	0.497	0.543	0.601
	Oracle	0.473	0.522	0.576
No shrinkage	DIC	0.521	0.603	0.641
	LPML	0.537	0.616	0.662
	Oracle	0.516	0.602	0.635
No sparse	DIC	0.599	0.699	0.759
	LPML	0.623	2.048	0.791
	Oracle	0.614	0.697	0.758
Hoff Model		0.517	0.588	0.682
IWH		0.632	0.759	0.861
Independent IW [1]		1.000	1.000	1.000

[1] Risk re-scaled, so Independent IW has value 1.0.

Table 3.2: Model Selection Results

	DIC						LPML					
Number of factors	K=3	K=4	K=5	K=6	K=7	K=8	K=3	K=4	K=5	K=6	K=7	K=8
Case 1 (N=300)		49	116	32	3				5	37	63	95
Case 2 (N=600)			58	82	49	11			4	7	44	145
Case 3 (N=1200)			55	79	53	13			3	11	55	131

Table 3.3: Risk Estimates for Case 4, Case 5 & Case 6

Model	Model selection criterion	$\mathcal{L}\left(\Omega_m, \hat{\Omega}_m\right)$		
		Case 4	Case 5	Case 6
BaSH-F	DIC	0.300	0.431	0.495
	Oracle	0.305	0.431	0.498
No shrinkage	DIC	0.364	0.483	0.614
	Oracle	0.364	0.481	0.613
No sparse	DIC	0.434	0.436	0.606
	Oracle	0.434	0.436	0.612
Hoff Model		0.384	0.399	0.508
IWH		0.584	0.636	0.600
Independent IW[1]		1.000	1.000	1.000

Table 3.4: Risk Estimates for Case 1, Case 7 & Case 8

Model	Model selection criterion	$\mathcal{L}\left(\Omega_m, \hat{\Omega}_m\right)$		
		Case 1	Case 7	Case 8
BaSH-F	DIC	0.478	0.525	0.552
	Oracle	0.473	0.521	0.548
No shrinkage	DIC	0.521	0.567	0.593
	Oracle	0.516	0.808	0.591
No sparse	DIC	0.599	0.657	0.663
	Oracle	0.614	0.633	0.663
Hoff Model		0.517	0.583	0.589
IWH		0.632	0.733	0.772
Independent IW[1]		1.000	1.000	1.000

Table 3.5: Risk Estimates for Case 9 & Case 10

Model	Model selection criterion	$\mathcal{L}\left(\Omega_m, \hat{\Omega}_m\right)$	
		Case 9	Case 10
BaSH-F	DIC	0.639	0.358
No shrinkage	DIC	0.696	0.433
No sparse	DIC	0.769	0.460
Hoff Model		0.601	0.245
IWH		0.538	0.321
Independent IW[1]		1.000	1.000

Table 3.6: Model comparison statistics and true classification rate for Letter recognition data

Sample Size	Model Specification	True Classification	Loss	Log-score
$n_m = 20$	BaSH-F ($K = 4$)	0.74	0.376	-18,446
	Hoff	0.69	0.428	-21,652
	IWH	0.71	0.451	-22,455
	Independent IW	0.51	0.694	-87,083
$n_m = 40$	BaSH-F ($K = 7$)	0.81	0.276	-15,031
	Hoff	0.79	0.299	-17,497
	IWH	0.75	0.464	-26,381
	Independent IW	0.72	0.402	-23,739
$n_m = 100$	BaSH-F ($K = 8$)	0.85	0.214	-12,370
	Hoff	0.77	0.273	-18,943
	IWH	0.80	0.293	-15,511
	Independent IW	0.79	0.300	-15,924

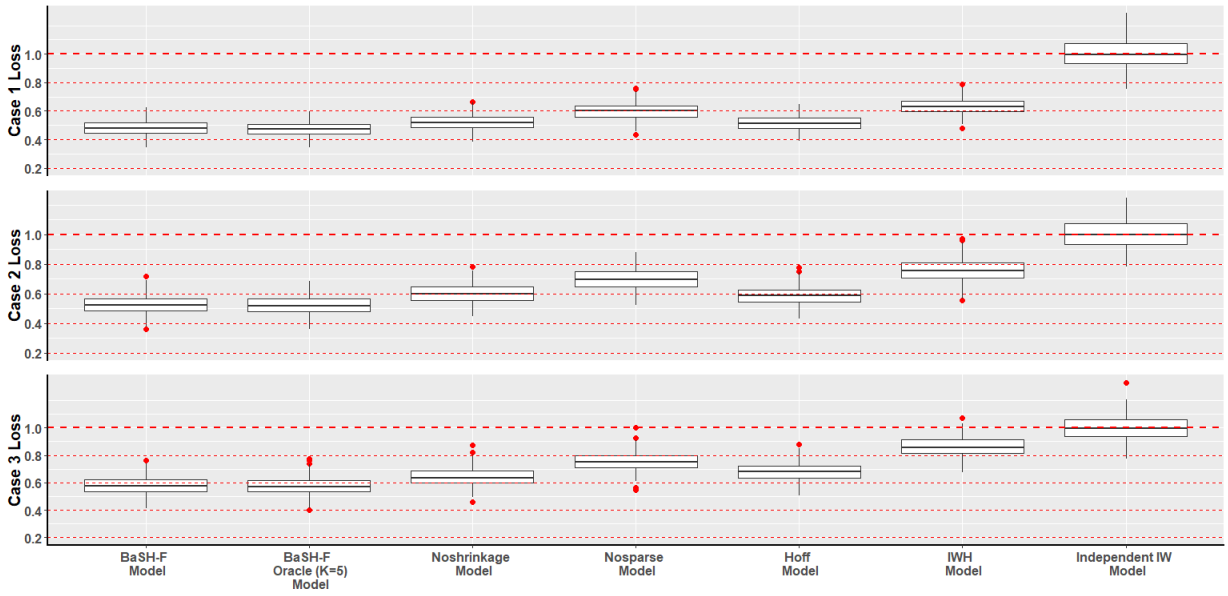

Figure 3.1: Estimated loss for Case 1, Case 2 & Case 3

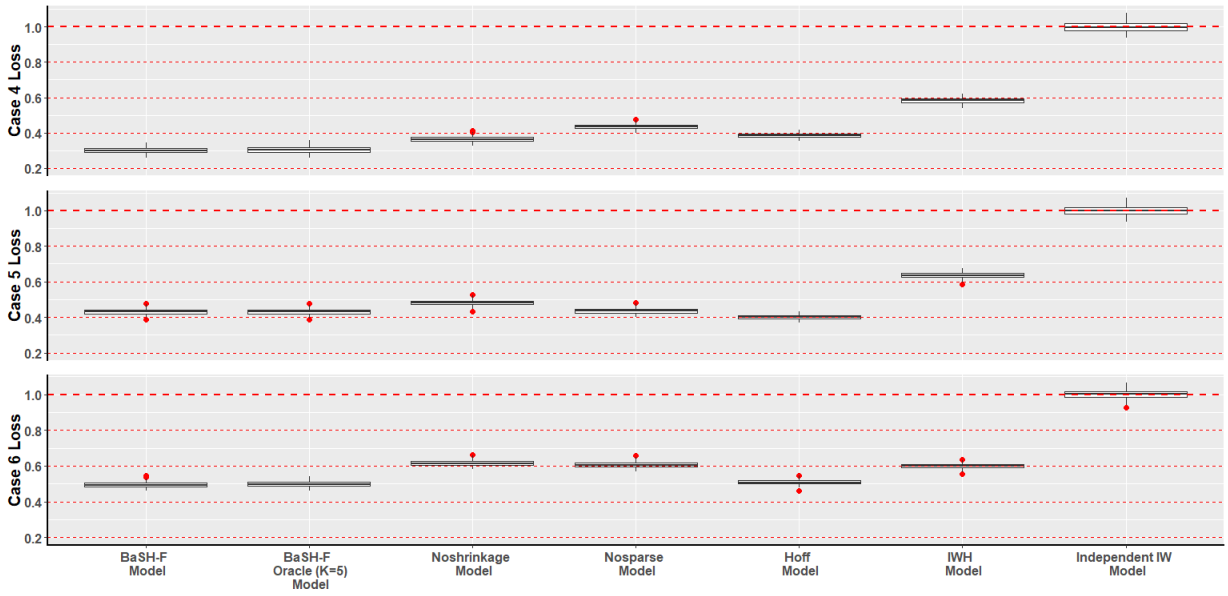

Figure 3.2: Estimated loss for Case 4, Case 5 & Case 6

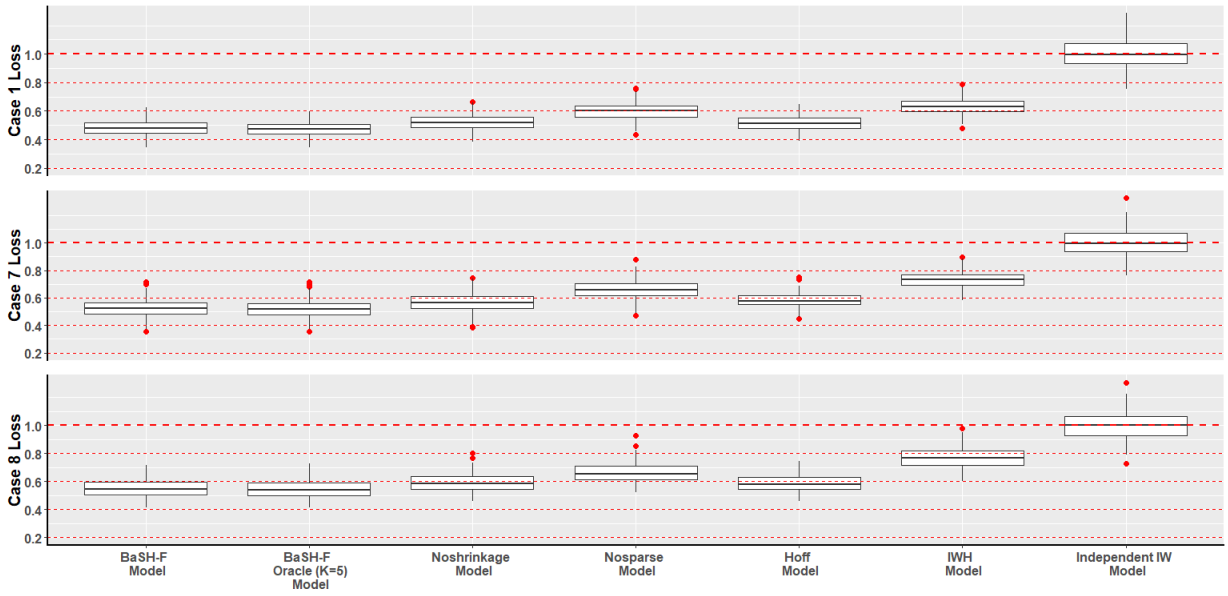

Figure 3.3: Estimated loss for Case 1, Case 7 & Case 8

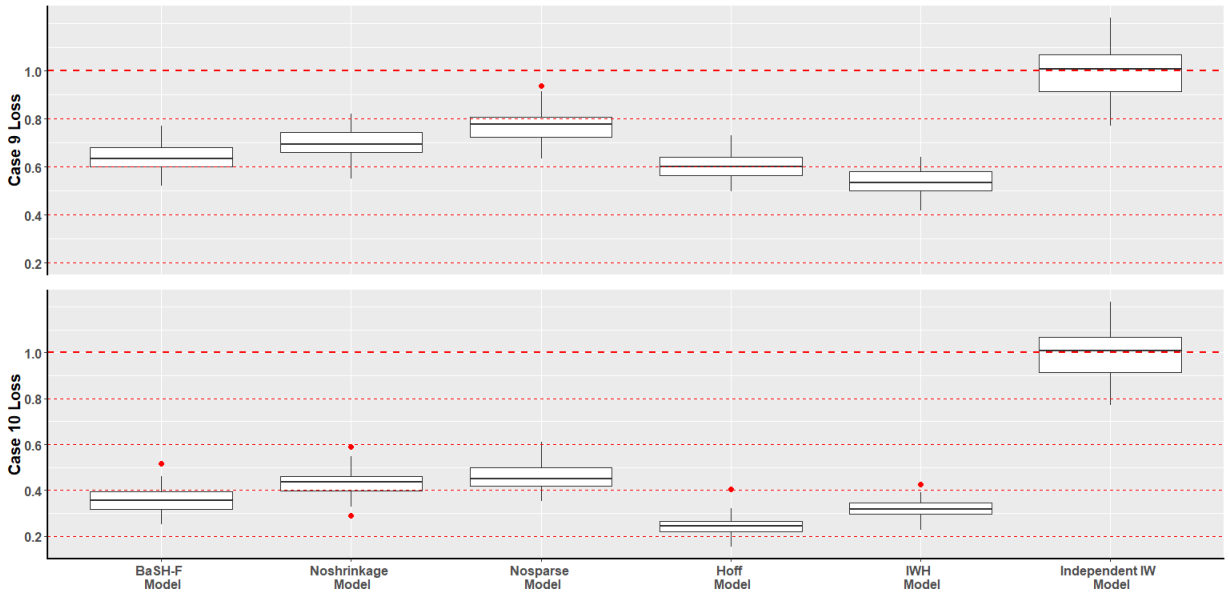

Figure 3.4: Estimated loss for Case 9 & Case 10

CHAPTER 4

DISCUSSION

In this dissertation, we have proposed two novel Bayesian approaches in the field of multivariate analysis. In Chapter 2 we have developed a general strategy of variable selection in the multivariate regression model by sharing common local parameters across all of the response variables. We have demonstrated the utility of our approach in comparison to alternatives. Our approaches are found to be superior in terms of both predictive performance and parameter estimation.

In Chapter 3, we have developed a novel technique for simultaneous covariance estimation based on sparse Bayesian factor model. We have also established the prediction accuracy of our proposed method in compare to other competitors through simulation results and data applications.

Both these projects are centered around fully Bayesian inference schemes based on Gibbs sampling and teasing out theoretically challenging posterior conditionals. The next layer of challenges involved devising computationally scalable algorithms to implement these schemes for high dimensional datasets. These often require considerable care in tuning the MCMC schemes. While some of these issues have included a careful choice of hyper-parameters, others involved employing matrix inversion techniques while some others cleverly incorporating adaptive sampling schemes from the existing literature. The computational success of these algorithms is borne out in extensive simulation studies that have been conducted in R for validating our models.

This dissertation project represents a first step at the problem, and there are

many further extensions and developments worth considering. For instance, in the first project, some possible application of this methodology could be in binary outcome data, hurdle models, causal-inference models and generalized linear models. Also, working on these projects have stirred a couple of ideas for natural extensions that we have set aside for our future work. These include extending the response distributions to non-Gaussian settings with an eye towards big-data genomic applications. For the second project, some potential future applications of this methodology include health/social survey data with multiple groups defined by any demographic factors like ethnicity, age or gender and among others. Also, an extension that incorporates a Bayesian non-parametric component could be used to consider clustering of the groups.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52(3):317–332.

Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli*, 9(2):291–312.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4):1011–1018.

Bai, R. and Ghosh, M. (2018). High-dimensional multivariate posterior consistency under global–local shrinkage priors. *Journal of Multivariate Analysis*, 167:157–170.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, pages 1281–1311.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98(2):291–306.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490. PMID: 27019543.

Boik, R. J. (2002). Spectral models for covariance matrices. *Biometrika*, 89(1):159–182.

Brown, B., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach. *Biometrika*, 86(3):635–648.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3):627–641.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). High-dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456. PMID: 21218139.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Dunson, D. B., Xue, Y., and Carin, L. (2008). The matrix stick-breaking process: Flexible Bayes meta-analysis. *Journal of the American Statistical Association*, 103(481):317–327.

Flury, B. K. (1987). Two generalizations of the common principal component model. *Biometrika*, 74(1):59–69.

Flury, B. N. (1984). Common principal components in K groups. *Journal of the American Statistical Association*, 79(388):892–898.

Frey, P. W. and Slate, D. J. (1991). Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6(2):161–182.

Gaskins, J. (2019). Hyper markov laws for correlation matrices. *Statistica Sinica*, 29(1):165–184.

Gaskins, J. and Daniels, M. (2016). Covariance partition priors: A Bayesian approach to simultaneous covariance estimation for longitudinal data. *Journal of Computational and Graphical Statistics*, 25(1):167–186.

Gaskins, J. T. and Daniels, M. J. (2012). A nonparametric prior for simultaneous covariance estimation. *Biometrika*, 100(1):125–138.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(3):501–514.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Green, P. J. and Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3):1391–1403.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Hoff, P. D. (2009). A hierarchical eigenmodel for pooled covariance estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):971–992.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Manly, B. F. and Rayner, J. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, 74(4):841–847.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–117.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Polson, N. G. and Scott, J. (2010). Shrinkg globally, act locally: Sparse Bayesian regularization and prediction. In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., editors, *Bayesian Statistics 9*, pages 501–538.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M., Daniels, M. J., and Park, T. (2007). Simultaneous modelling of the cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, 98(3):568–587.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22(3):1195–1211.

APPENDIX

Appendix A

This section includes the posterior consistency of Bayesian variable selection for multi-outcome model Chapter 2.

A.1 Posterior Consistency of Bayesian Variable Selection for Multi-outcome Model

Here we provide details and the proof of the posterior consistency results from Section 2.2.5.

For our discussion we use the term multi-index to denote a model where the individual observations belong to a common multidimensional family $f(\cdot)$ but are indexed by possibly different parameters θ_{iB}. The second subscript denotes a global parameter B, which in our context is the (shared) matrix of regression coefficients. Thus, in our multivariate Gaussian regression we let $\theta_{iB} = X_{i\cdot}B$ be the K-vector representing the mean of the K responses.

Recall that the KL distance between two densities is defined as $E_{f_0}\left\{\log \frac{f_1(Y)}{f_0(Y)}\right\}$. For multi-index families, we extend the definition to have a notion KL distance for each i. To that end, the KL distance between the global parameter B and the true value B_0 for observation i can be written as

$$KL_i(B, B_0) = E_{B_0}\left\{\log\left(\frac{f(Y_i, \theta_{iB})}{f(Y_i; \theta_{iB_0})}\right)\right\},$$

where $\theta_{iB} = X_{i\cdot}B$ and $\theta_{iB_0} = X_{i\cdot}B_0$ are parameter vectors indexing the densities for

observation i under parameters B and B_0. We also define $V_i(B, B_0)$ as the variance analogue, i.e.,

$$V_i(B, B_0) = Var_{B_0}\left\{\log\left(\frac{f(Y_i, \theta_{iB})}{f(Y_i; \theta_{iB_0})}\right)\right\}.$$

We first state our Lemma 1 which establishes a uniformly exponentially consistent (UEC) sequence of tests that will be required in the proof of Theorem 1. Here, we include the dependence on n by letting Y_n and X_n denote the response and design matrices for a sample of size n.

Lemma 1. *For any $\epsilon > 0$, define $B_\epsilon = \{B : ||B - B_0|| > \epsilon\}$. Let $\Phi_n = I\left(Y_n \in C_n\right)$ be the test statistic based on the critical region $C_n = \left\{Y_n : ||\hat{B} - B_0|| > \frac{\epsilon}{2}\right\}$ and $\hat{B} = \left(X_n^T X_n\right)^{-1} X_n^T Y_n$. Further, assume condition (A3), and let d be the largest eigenvalue of Ψ. Then, for the likelihood* (2.1)*, we have the following:*

1. $E_{B_0}\left(\Phi_n\right) \leq \exp\left(-n\frac{\epsilon^2 c}{16d}\right)$,
2. $\sup_{B \in B_\epsilon} E_B\left(1 - \Phi_n\right) \leq \exp\left(-n\frac{\epsilon^2 c}{16d}\right)$.

Proof: Proof of this lemma follows as in Lemma 1 of Bai and Ghosh (2018).

Next, we state and prove Lemma 2 which establishes the prior positivity condition.

Lemma 2. *Assume a fixed Ψ, the likelihood* (2.1)*, (A1), and (A2). Then, for all $\epsilon > 0$, there exists a set C_ϵ with $\pi(B \in C_\epsilon) > 0$, such that for all $B \in C_\epsilon$*

$$\begin{aligned} KL_i(B, B_0) &< \epsilon \quad \text{for all } i, \\ \sum_{i=1}^{\infty} \frac{1}{i^2} V_i(B, B_0) &< \infty. \end{aligned}$$

Proof: A little algebra shows that

$$KL_i(B, B_0) = (X_i B - X_i B_0)\Psi^{-1}(X_i B - X_i B_0)'.$$

Let $\tilde{X}_i = I_K \otimes X_{i\cdot}$, $\beta = \text{vec}(B)$, and $\beta_0 = \text{vec}(B_0)$. Then, it follows that

$$\begin{aligned} KL_i(B, B_0) &= (X_iB - X_iB_0)\Psi^{-1}(X_iB - X_iB_0)' = (\tilde{X}_i\beta - \tilde{X}_i\beta_0)'\Psi^{-1}(\tilde{X}_i\beta - \tilde{X}_i\beta_0) \\ &= (\beta - \beta_0)\tilde{X}_i\Psi^{-1}\tilde{X}_i'(\beta - \beta_0)' = ||M_i(\beta - \beta_0)||^2, \end{aligned}$$

where $M_i = \Psi^{-\frac{1}{2}}\tilde{X}_{i\cdot}$. From the sub-multiplicativity of the Frobenius norm, $||M_i||$ is bounded by $||\Psi^{-\frac{1}{2}}||\,||\tilde{X}_i|| = K^{1/2}||\Psi^{-\frac{1}{2}}||\,||X_{i\cdot}||$, which is bounded by $GK^{1/2}||\Psi^{-\frac{1}{2}}||$ using (A2). Clearly, $||\beta - \beta_0|| = ||B - B_0||$. Thus, a set $C_\epsilon = \left\{ B : ||B - B_0|| < \frac{\epsilon}{GK^{1/2}||\Psi^{-\frac{1}{2}}||} \right\}$ will clearly satisfy $KL_i(B, B_0) < \epsilon$ for all i. By (A1) the continuous prior $\pi(B)$ assigns positive probability to any such open neighborhood C_ϵ. Similar steps show that for all B in C_ϵ, the V_is are bounded uniformly by a constant across all n, proving convergence of $\sum_{i=1}^{\infty} \frac{1}{i^2} V_i(B, B_0)$.

We first introduce and sketch the proof of a more general theorem that establishes posterior consistency for a wide range of multi-index models.

Theorem 2. *Consider a multi-index model with global parameter B and independent observations Y_i, $i = 1, \ldots, n, \ldots$ with $Y_i \sim f(\cdot, \theta_{iB_0})$ under the true global parameter value B_0. Further assume the following two conditions:*

1. *There exist tests Φ_n such that $E_{B_0}(\Phi_n) < \exp(-nC_1)$ and that for all $B \neq B_0$, $E_B(1 - \Phi_n) < \exp(-nC_2)$. Here, C_1 and C_2 are constants not depending on the parameter of interest.*

2. *There exists a set C_ϵ with $\pi(B \in C_\epsilon) > 0$, such that for all $B \in C_\epsilon$,*

$$\begin{aligned} KL_i(B, B_0) &< \epsilon \quad \textit{for all } i, \\ \sum_{i=1}^{\infty} \frac{1}{i^2} V_i(B, B_0) &< \infty. \end{aligned}$$

Then, the posterior distribution for B is consistent. That is, for any $\epsilon > 0$,

$$P_{B_0}\{\,||B - B_0|| > \epsilon \mid Y_1, \ldots, Y_n\} \to 0, \quad \textit{a.s. as } n \to \infty.$$

Proof: The proof of this theorem is a combination of arguments in Schwartz (1965), Amewou-Atisso et al. (2003) and Choi and Schervish (2007), and we omit the technical details. Briefly the argument is as follows. The posterior probability of interest, denoted by L_n^ϵ, can be written as a ratio of integrals of two likelihood ratios in the following way

$$L_n^\epsilon = P_{B_0}\{\,||B - B_0|| > \epsilon \mid Y_1, \ldots, Y_n\} = \frac{\int_{U_\epsilon} \frac{\prod_i f(Y_i, \theta_{iB})}{\prod_i f(Y_i, \theta_{iB_0})} dB}{\int_U \frac{\prod_i f(Y_i, \theta_{iB})}{\prod_i (f(Y_i, \theta_{iB_0})} dB},$$

where $U_\epsilon = \{B : ||B - B_0|| > \epsilon\}$ is the ϵ-ball around B_0 and U is the entire parameter space. The aim is to show L_n^ϵ converges to 0 a.s. under P_{B_0} for all $\epsilon > 0$.

As shown in Schwartz (1965), we may bound L_n^ϵ using the test statistic Φ_n as

$$L_n^\epsilon \leq \Phi_n + \frac{J_{1n}}{J_{2n}},$$

where $J_{1n} = \int_{U_\epsilon} \frac{(1-\Phi_n)\prod_i f(Y_i, \theta_{B_i})}{\prod_i f(Y_i, \theta_{B_0})} dB$ and $J_{2n} = \int_U \frac{\prod_i f(Y_i, \theta_{B_i})}{\prod_i f(Y_i, \theta_{B_0})} dB$. Following the arguments from Schwartz (1965) (also used in Bai and Ghosh (2018) and Armagan et al. (2013)), the first condition in Theorem 2 can be shown to imply $\Phi_n \to 0$ a.s. Further, $e^{nC} J_{1n} \to 0$ a.s., for a constant $C > 0$ that may depend on auxiliary parameters (such as Ψ and the eigenvalues of the design matrix) but not on B_0. Similarly, the second condition of the Theorem 2 can be shown to imply that for any constant $c > 0$, $e^{nc} J_{2n} \to \infty$ a.s. In combination, these imply that L_n^ϵ converges almost surely to zero under the true parameter B_0, guaranteeing posterior consistency.

We note that the proof of this theorem has a general flavor in that it only requires a UEC sequence of tests and prior positivity. The first condition can be

satisfied for several settings involving multivariate Gaussian likelihoods. The second condition is applicable to a variety of model specifications and holds simply when observations are independent but not identically distributed. Of note, condition 2 was proved in Schwartz (1965) for single-index families and later adapted to multi-index families (Choi and Schervish, 2007). For a proof of this, we refer the reader to the proof of part A.5 in Theorem 1 from Choi and Schervish (2007).

Proof of Theorem 1: Results from Lemmas 1 and 2 are immediately obtained from assumptions (A1)–(A3), and these lemmas establish the two conditions required for Theorem 2. Hence, Theorem 1 is proved.

Appendix B

This section includes the additional computation in Chapter 3.

A.2 Calculation of the posterior probability p^* for Simultaneous Covariance Estimation

In section 3.2.2 we have discussed the blockwise sampling algorithm for Z_{jk}, ω_{jk} and λ_{mjk}. From the equation (3.1) we have $Y_{mij} \sim N\left(\sum_{k=1}^{K} \lambda_{mjk}\eta_{mik}, \sigma_{mj}^2\right)$, which implies, $e_{mij} = Y_{mij} - \sum_{\substack{l=1 \\ l\neq k}}^{K} \lambda_{mjk}\eta_{mik} \sim N\left(\lambda_{mjk}\eta_{mik}, \sigma_{mj}^2\right)$.

From our hierarchical model (3.2), the posterior distribution ω_{jk} will be

$$\omega_{jk}|e_{mjk} \sim p^* I\left(\omega_{jk} = 0\right) + (1 - p^*)N(\mu_w^*, {\sigma_w^*}^2),$$

with $p^* = \frac{B}{A+B}$, where

$$A = \pi \prod_{m=1}^{M} \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk} = 0\right),$$

$$B = (1-\pi) \prod_{m=1}^{M} \left[\int \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk}\right) f\left(\lambda_{mjk}|\omega_{jk}\right) d\lambda_{mjk}\right] f\left(\omega_{jk}|Z_{jk} = 1\right) d\omega_{jk}.$$

Note that these conditional distributions rely on many other parameters in the conditioning statement. To simplify notation we only include those that are involved in the calculations. Now we derive A and B separately,

$$\begin{aligned} A &= \pi \prod_{m=1}^{M} \prod_{i=1}^{n_m} f(Y_{mij}|\lambda_{mjk} = 0) \\ &= \pi \prod_{m=1}^{M} \prod_{i=1}^{n_m} (2\pi\sigma_{mj}^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} e_{mij}^2\right] \\ &= \pi \left(\prod_{m=1}^{M} (2\pi\sigma_{mj}^2)^{-\frac{n_m}{2}}\right) \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\sum_{i=1}^{n_m} \frac{e_{mij}^2}{\sigma_{mj}^2}\right]. \end{aligned}$$

Taking the log transformation over A, we have,

$$a = \log A = \log \pi - \sum_{m=1}^{M} \frac{n_m}{2} \log(2\pi\sigma_{mj}^2) - \frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \frac{e_{mij}^2}{\sigma_{mj}^2}. \tag{3.6}$$

Now,

$$B = (1-\pi) \prod_{m=1}^{M} \left[\int \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk}\right) f\left(\lambda_{mjk}|\omega_{jk}\right) d\lambda_{mjk} \right] f\left(\omega_{jk}|Z_{jk}=1\right) d\omega_{jk}.$$

First we simplify the $\left[\int \prod_{i=1}^{n_m} f\left(Y_{mij}|\lambda_{mjk}\right) f\left(\lambda_{mjk}|\omega_{jk}\right) d\lambda_{mjk}\right]$ term in the following.

Now,

$$
\begin{aligned}
& \int \prod_{i=1}^{n_m} f(e_{mij}|\lambda_{mjk})\, f(\lambda_{mjk}|w_{jk})\, d\lambda_{mjk} \\
= & \int \prod_{i=1}^{n_m} \left\{ \left(2\pi\sigma_{mj}^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} \left(e_{mij} - \lambda_{mjk}\eta_{mik}\right)^2 \right] \right\} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\nu_m^2} \left(\lambda_{mjk} - \omega_{jk}\right)^2 \right] d\lambda_{mjk} \\
= & \int \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} \sum_{i=1}^{n_m} \left(e_{mij}^2 - 2e_{mij}\lambda_{mjk}\eta_{mik} + \lambda_{mjk}^2\eta_{mik}^2\right) \right] \\
& \times \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\nu_m^2} \left(\lambda_{mjk}^2 - 2\lambda_{mjk}\omega_{jk} + \omega_{jk}^2\right) \right] d\lambda_{mjk} \\
= & \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} \sum_{i=1}^{n_m} e_{mij}^2 \right] \exp\left[-\frac{1}{2\nu_m^2}\omega_{jk}^2 \right] \\
& \times \int \exp\left[-\frac{1}{2} \left\{ \lambda_{mjk}^2 \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right) - 2\lambda_{mjk} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2} \right) \right\} \right] d\lambda_{mjk} \\
= & \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} \sum_{i=1}^{n_m} e_{mij}^2 \right] \exp\left[-\frac{1}{2\nu_m^2}\omega_{jk}^2 \right] \\
& \times \int \exp\left[-\frac{1}{2} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right) \left\{ \lambda_{mjk}^2 - 2\lambda_{mjk} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \right.\right. \\
& \left.\left. \times \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2} \right) + \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-2} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2} \right)^2 \right\} \right] d\lambda_{mjk} \\
& \times \exp\left[\frac{1}{2} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2} \right)^2 \right] \\
= & \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2} \sum_{i=1}^{n_m} e_{mij}^2 \right] \exp\left[-\frac{1}{2\nu_m^2}\omega_{jk}^2 \right] \\
& \times (2\pi)^{\frac{1}{2}} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-\frac{1}{2}} \exp\left[\frac{1}{2} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2} \right)^2 \right].
\end{aligned}
$$

Hence,

$$
\begin{aligned}
B = \quad & (1-\pi)\int \prod_{m=1}^{M} \left\{ \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2}\sum_{i=1}^{n_m} e_{mij}^2\right] \exp\left[-\frac{1}{\nu_m^2}\omega_{jk}^2\right] \right. \\
& \left. \times \left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-\frac{1}{2}} \exp\left[\frac{1}{2}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} + \frac{\omega_{jk}}{\nu_m^2}\right)^2\right] \right\} \\
& \times f\left(\omega_{jk}|Z_{jk}=1\right) d\omega_{jk} \\
= \quad & (1-\pi)\int \prod_{m=1}^{M} \left\{ \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \left(2\pi\nu_m^2\right)^{-\frac{1}{2}} (2\pi)^{\frac{1}{2}} \left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma_{mj}^2}\sum_{i=1}^{n_m} e_{mij}^2\right] \right. \\
& \left. \times \exp\left[\frac{1}{2}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left\{\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2}\right)^2 + 2\omega_{jk}\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2}\right) + \frac{\omega_{jk}^2}{\nu_m^4}\right\}\right] \right\} \\
& \times \exp\left[-\frac{1}{\nu_m^2}\omega_{jk}^2\right] \left(2\pi\tau^2\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\tau^2}\omega_{jk}^2\right] d\omega_{jk} \\
= \quad & (1-\pi)(2\pi\tau^2)^{-\frac{1}{2}} \left(\prod_{m=1}^{M}\left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}}\right) \exp\left[-\frac{1}{2}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\frac{e_{mij}^2}{\sigma_{mj}^2}\right] \\
& \times \exp\left[\frac{1}{2}\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2}\right)^2\right] \\
& \times \left(\prod_{m=1}^{M}\left(\nu_m^2\right)^{-\frac{1}{2}}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-\frac{1}{2}}\right) \\
& \times \int \exp\left[-\frac{1}{2}\left\{\omega_{jk}^2\left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\left\{\frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\frac{1}{\nu_m^4}\right\}\right)\right.\right. \\
& \left.\left. -2\omega_{jk}\left(\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2}\right)\right\}\right] d\omega_{jk}
\end{aligned}
$$

$$
\begin{aligned}
B = \quad & (1-\pi)(2\pi\tau^2)^{-\frac{1}{2}} \left(\prod_{m=1}^{M} \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \right) \exp\left[-\frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \frac{e_{mij}^2}{\sigma_{mj}^2} \right] \\
& \times \exp\left[\frac{1}{2} \sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} \right)^2 \right] \\
& \times \left(\prod_{m=1}^{M} \left(\nu_m^2\right)^{-\frac{1}{2}} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-\frac{1}{2}} \right) \\
& \times \int \exp\left[-\frac{1}{2} \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \left\{ \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right\} \right) \right. \\
& \times \left\{ \omega_{jk}^2 - 2\omega_{jk} \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \left\{ \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right\} \right)^{-1} \right. \\
& \times \left[\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2} \right) \right] \\
& \left. + \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \left\{ \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right\} \right)^{-2} \right. \\
& \left. \left. \times \left(\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2} \right)^2 \right\} \right] d\omega_{jk} \\
& \times \exp\left[\frac{1}{2} \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \left\{ \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right\} \right)^{-1} \right. \\
& \left. \left(\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2} \right)^2 \right]
\end{aligned}
$$

$$
\begin{aligned}
B = \quad & (1-\pi)(2\pi\tau^2)^{-\frac{1}{2}} \left(\prod_{m=1}^{M} \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \right) \exp\left[-\frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \frac{e_{mij}^2}{\sigma_{mj}^2} \right] \\
& \times \exp\left[\frac{1}{2} \sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} \right)^2 \right] \\
& \times \left(\prod_{m=1}^{M} \left(\nu_m^2\right)^{-\frac{1}{2}} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-\frac{1}{2}} \right) \\
& \times (2\pi)^{\frac{1}{2}} \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \left\{ \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right\} \right)^{-\frac{1}{2}} \\
& \exp\left[\frac{1}{2} \left\{ \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right)^{-1} \right. \right. \\
& \left. \left. \left(\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2 \nu_m^2} \right)^2 \right\} \right] \\
= \quad & (1-\pi)(2\pi\tau^2)^{-\frac{1}{2}} \left(\prod_{m=1}^{M} \left(2\pi\sigma_{mj}^2\right)^{-\frac{n_m}{2}} \right) \exp\left[-\frac{1}{2} \sum_{m=1}^{M} \sum_{i=1}^{n_m} \frac{e_{mij}^2}{\sigma_{mj}^2} \right] \\
\times \quad & \exp\left[\frac{1}{2} \sum_{m=1}^{M} \left\{ \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \left(\frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2} \right)^2 \right\} \right] \\
\times \quad & \exp\left[\frac{1}{2} \left\{ \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right)^{-1} \right. \right. \\
& \left. \left. \left(\sum_{m=1}^{M} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{\sum_{i=1}^{n_m} e_{mij}\eta_{mik}}{\sigma_{mj}^2 \nu_m^2} \right)^2 \right\} \right] \\
\times \quad & \left(\prod_{m=1}^{M} \left(\nu_m^2\right)^{-\frac{1}{2}} \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-\frac{1}{2}} \right) \\
\times \quad & (2\pi)^{\frac{1}{2}} \left(\frac{1}{\tau^2} + \sum_{m=1}^{M} \frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m} \eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2} \right)^{-1} \frac{1}{\nu_m^4} \right)^{-\frac{1}{2}} .
\end{aligned}
$$

By simplifying, $\frac{1}{\nu_m^2} - \left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\frac{1}{\nu_m^4} = \frac{1}{\nu_m^2+\sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}$ and taking the log of B we have,

$$
\begin{aligned}
b &= \log B \\
&= \log(1-p) - \frac{1}{2}\log(2\pi\tau^2) - \sum_{m=1}^{M}\frac{n_m}{2}\log(2\pi\sigma_{mj}^2) - \frac{1}{2}\sum_{m=1}^{M}\sum_{i=1}^{n_m}\frac{e_{mij}^2}{\sigma_{mj}^2} \\
&\quad + \frac{1}{2}\sum_{m=1}^{M}\left\{\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m}e_{mij}\eta_{mik}}{\sigma_{mj}^2}\right)^2\right\} \\
&\quad + \frac{1}{2}\left\{\left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\frac{1}{\nu_m^2+\sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right)^{-1}\right. \\
&\quad \left.\left(\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\frac{\sum_{i=1}^{n_m}e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2}\right)^2\right\} \\
&\quad - \frac{1}{2}\sum_{m=1}^{M}\left\{\log\nu_m^2 + \log\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right) - \log(2\pi)\right\} \\
&\quad - \frac{1}{2}\left\{\log\left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\frac{1}{\nu_m^2+\sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right)\right\}. \qquad (3.7)
\end{aligned}
$$

Finally from (3.6) and (3.7) we have,

$$
\begin{aligned}
&c = b - a \\
&= \log(1-\pi) - \frac{1}{2}\log(2\pi\tau^2) - \log(\pi) - \frac{1}{2}\left\{\log\left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\frac{1}{\nu_m^2+\sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right)\right\} \\
&\quad + \frac{1}{2}\sum_{m=1}^{M}\left\{\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\left(\frac{\sum_{i=1}^{n_m}e_{mij}\eta_{mik}}{\sigma_{mj}^2}\right)^2\right\} + \\
&\frac{1}{2}\left\{\left(\frac{1}{\tau^2} + \sum_{m=1}^{M}\frac{1}{\nu_m^2+\sigma_{mj}^2/\sum_{i=1}^{n_m}\eta_{mik}^2}\right)^{-1}\left(\sum_{m=1}^{M}\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right)^{-1}\frac{\sum_{i=1}^{n_m}e_{mij}\eta_{mik}}{\sigma_{mj}^2\nu_m^2}\right)^2\right\} \\
&\quad - \frac{1}{2}\sum_{m=1}^{M}\left\{2\log\nu_m + \log\left(\frac{\sum_{i=1}^{n_m}\eta_{mik}^2}{\sigma_{mj}^2} + \frac{1}{\nu_m^2}\right) - \log(2\pi)\right\}.
\end{aligned}
$$

CURRICULUM VITA

NAME: Debamita Kundu

ADDRESS: Department of Biostatistics and Bioinformatics
University of Louisville
Louisville, KY 40202

EDUCATION: Bachelor of Science in Statistics,
University of Calcutta, India, 2012
Master of Science in Statistics,
Presidency University, India, 2014

PUBLICATIONS: Kundu, D., Mitra, R., Gaskins, J.T. (2018)
Bayesian Variable Selection for Multi-Outcome
Models Through Shared Shrinkage.
Scandinavian Journal of statistics (Submitted).

PRESENTATIONS: Southern Regional Council on Statistics Conference. June 3, 2019.
A Bayesian Hierarchical Sparse Factor Model for
Simultaneous Covariance Estimation.

Kentucky ASA Chapter Meeting. April 5, 2019.
A Bayesian Hierarchical Sparse Factor Model for

Simultaneous Covariance Estimation.

Joint Statistical Meetings (JSM). August 1, 2018.
Bayesian Variable Selection for Multi-Outcome Models
Through Shared Shrinkage.

Southern Regional Council on Statistics Conference. June 5, 2018.
Bayesian Variable Selection for Multi-Outcome Models
Through Shared Shrinkage.

Department of Bioinformatics & Biostatistics Seminar Series.
April 20, 2018.
Bayesian Variable Selection for Multi-Outcome Models
Through Shared Shrinkage.

Eastern North American Region Spring Meeting. March 28, 2018.
Bayesian Variable Selection for Multi-Outcome Models
Through Shared Shrinkage.

Kentucky ASA Chapter Meeting. March 2, 2018.
Bayesian Variable Selection for Multi-Outcome Models
Through Shared Shrinkage.

HONORS AND AWARDS

NSF funded Harshbarger Travel award for the SRCOS Summer Research Conference, Carollton, KY, 2019.

NSF funded Harshbarger Travel award for the SRCOS Summer Research Conference, Jekyll Island, GA, 2017.

School of Public Health Travel Scholarship (University of Louisville) for ENAR January 2018.

School of Interdisciplinary and Graduate Studies Travel Scholarship (University of Louisville) for JSM August 2018.