University of California

Los Angeles

Bayesian Method for Support Union Recovery in Multivariate Multi-Response Linear Regression

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Wan-Ping Chen

2015

Abstract of the Dissertation

Bayesian Method for Support Union Recovery in Multivariate Multi-Response Linear Regression

by

Wan-Ping Chen

Doctor of Philosophy in Statistics

University of California, Los Angeles, 2015

Professor Yingnian Wu, Chair

Sparse modeling has become a particularly important and quickly developing topic in many applications of statistics, machine learning, and signal processing. The main objective of sparse modeling is discovering a small number of predictive patterns that would improve our understanding of the data. This paper extends the idea of sparse modeling to the variable selection problem in high dimensional linear regression, where there are multiple response vectors, and they share the same or similar subsets of predictor variables to be selected from a large set of candidate variables. In the literature, this problem is called multi-task learning, support union recovery or simultaneous sparse coding in different contexts.

We present a Bayesian method for solving this problem by introducing two nested sets of binary indicator variables. In the first set of indicator variables, each indicator is associated with a predictor variable or a regressor, indicating whether this variable is active for any of the response vectors. In the second set of indicator variables, each indicator is associated with both a predicator variable and a response vector, indicating whether this variable is active for the particular response vector. The problem of variable selection is solved by sampling from the posterior distributions of the two sets of indicator variables. We develop a Gibbs sampling algorithm for posterior sampling and use the generated samples to identify active support both in shared and individual level. Theoretical and simulation justification are performed in the paper.

The proposed algorithm is also demonstrated on the real image data sets. To learn the patterns of the object in images, we treat images as the different tasks. Through combining images with the object in the same category, we cannot only learn the shared patterns efficiently but also get individual sketch of each image.

The dissertation of Wan-Ping Chen is approved.

Frederic Paik Schoenberg

Hongquan Xu

Luminita Vese

Yingnian Wu, Committee Chair

University of California, Los Angeles

2015

To my parents, my husband and my children.

Table of Contents

List of Figures

List of Tables

Vita

1999	Bachelor of Science, Mathematics, National Taiwan University.
2001	Master of Science, Statistics, National Tsing Hua University.
2001–2005	Research Assistant, Institute of Statistical Science, Academia Sinica.
2006–2007	Teaching Assistant, Statistics, UCLA.
2007–2008	Teaching Associate, Statistics, UCLA.
2008	Advanced to Ph.D. Candidacy, Statistics, UCLA.
2008–2010	Teaching Fellow, Statistics, UCLA.
2009–2010	Research Assistant, Statistics, UCLA.

Publications

Nicole Chen, Y.-N. Wu, and R.-B. Chen. (2014) Bayesian Variable Selection for Multi-response Linear Regression. *Proceedings of the 2014 International conference on Technologies and Applications of Artificial Intelligence.*

CHAPTER 1

Introduction

Variable selection is a fundamental problem in linear regression, especially in modern applications where the number of predictor variables or regressors can exceed the number of observations. Under the sparsity assumption that the number of active variables is small, it is possible to select these active variables even if the number of candidate variables is very large.

During the past decade, the problem of variable selection in high dimensional linear regression has been intensely studied in statistics, machine learning and signal processing. Many variable selection methods have been developed, such as the Lasso by Tibshirani (1996) [Tib96], SCAD by Fan and Li (2001) [FL01], elastic net by Zou and Hastie (2005) [ZH03], and MCP by Zhang (2010) [Zha10]. In addition to these penalized least squares methods, Bayesian approaches have also been proposed, for example, stochastic search variable selection (SSVS) by George and McCulloch (1993) [GM93], Gibbs variable selection (GVS) by Dellaportas et al. (2000) [DFN00], and RVM by Tipping (2005) [Tip01].

Variable selection methods have also been proposed for group sparsity. For example, Yuan and Lin (2006) [YL06] proposed the group Lasso method under the group sparsity assumption. Simon et al. (2012) [ST12] generalized group Lasso to sparse group lasso. In the Bayesian framework, Farcomeni (2010) [Far10] proposed a Bayesian constrained variable selection approach that can also be used for group selection. Raman et al. (2009) [RFW09] proposed a Bayesian group Lasso method by extending the standard Bayesian Lasso. Chen et al. (2014) [CCCnt] introduced a Bayesian approach for the sparse group selection problem.

The linear regression problems treated by the above methods usually involve a single response vector. In some applications, there can be multiple response vectors, and these response vectors may be explained by the same or similar subsets of variables to be selected from a large set of candidate variables. Such shared sparsity pattern enables different response vectors to collaborate with or to borrow strength from each other to select the active variables. Such a problem has been studied by Tropp et al. (2006) [Tro06] under the name of simultaneous sparse coding, where each response vector is a signal, each predictor vector is a base signal or an atom, and the collection of all the base signals form a dictionary. The goal is to select a small number of base signals from the dictionary to represent the observed signals. The problem has been studied by Lounici et al. (2009) [LPT09] under the name of multi-task learning, where the regression of each response vector on the predictor variables is considered a single task. Obozinski et al. (2011) [OWJ11] studied this problem under the name of support union recovery, where the word "support" means the subset of variables selected for a response vector, and "support union" means the union of subsets of variables selected for all the response vectors. If the supports of different response vectors are similar, then the union of the supports will only be slightly bigger than the supports of individual response vectors.

In this paper, we propose a Bayesian method for solving the above support union recovery problem, by assuming two nested sets of binary indicator variables. In the first set of indicators, each indicator is associated with a variable, indicating whether this variable is active for any of the response vectors. The set of variables whose indicators are 1's then become the union of the supports. In the second set of indicators, each indicator is associated with both a variable and a response vector, indicating whether this variable is active for explaining the particular response vector. So the second set of indicators gives us the supports of individual response vectors. Variable selection can then be accomplished by sampling from the posterior distributions of the two sets of indicators. We develop the Gibbs sampling algorithm for posterior sampling and demonstrate the performances of the proposed method for both simulated and real data sets.

1.1 Variable selection

In machine learning and statistics, variable selection is a process of selecting a subset of relevant variables from all the candidate predictors. The desired task is that we can have good predictive ability on the new observations, or can explain the relationships in the data, through the promising model constructed by the selected features.

The algorithm of variable selection combines the search technique for the suggested new variable subsets, and a way of measuring used to evaluate the different variable subsets. The simplest algorithm is to test each possible subset of variables and find the best one which minimizes the error rate. However, when there are tens or hundreds of thousands of variables available in the dataset, it will be an exhaustive search of the predictor space and computationally intractable in most cases. Therefore, how to efficiently distinguish the relevant variables from other redundant variables has been the major issue. Different evaluation metrics and constrains are used in different algorithm. In this paper, we are going to focus on Lasso and Bayesian variable selection.

1.1.1 LASSO

Suppose we have a response vector $Y \in \mathbb{R}^n$ and a design matrix $\boldsymbol{X} = [X_1, \cdots, X_p] \in \mathbb{R}^{n \times p}$. We want of find a linear model $Y \approx \boldsymbol{X}\beta$ to describe the relationship between Y and $\boldsymbol{X}$, where $\beta \in \mathbb{R}^p$. If $n > p$, this is the classical linear regression problem. We can solve the problem by minimizing the ordinary least square,

$$\min_{\beta \in \mathbb{R}^n} ||Y - \boldsymbol{X}\beta||^2. \tag{1.1}$$

The solution is well-defined and can be found easily. However, high technology has made it possible to collect large amount data over the recent years, and the number of features often exceeds the number of examples, it means $p \gg n$. In this case, we believe many features in the data could be redundant and irrelevant. The idea is illustrated in Figure 1.1, where β is a sparse vector that has many zero components. Therefore, the goal is to dig out and identify nonzero coefficients and estimate their values. In order to incorporate the sparsity into the ordinary least square, a penalty term can be added in Eq (1.1). This goal becomes

$$\arg\min_{\beta \in \mathbb{R}^n} \left\{ ||Y - \boldsymbol{X}\beta||_2^2 + \lambda ||\beta||_0 \right\}, \tag{1.2}$$

where $||\beta||_0$, the 0-norm, denotes the number of nonzero components in β and $\lambda > 0$ is a regularization parameter. Unfortunately, this optimization problem is computationally intractable, because the 0-norm is non-convex.

Consider the convex relaxation, Tibshirani (1996) [Tib96] proposed an alternative version to replace the sparsity constraint by the l_1 norm $||\beta||_1$, which is the sum of the absolute coefficients. It is known as Lasso (least absolute shrinkage and selection

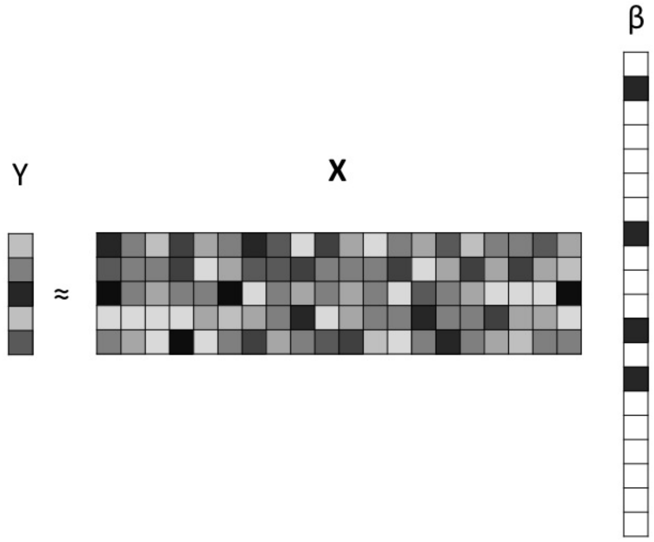


Figure 1.1: The linear regression model $Y \approx \boldsymbol{X}\beta$, with $p \gg n$ and a sparse X.

operator), and represented as

$$\arg\min_{\beta \in \mathbb{R}^n} \left\{ ||Y - \boldsymbol{X}\beta||_2^2 + \lambda ||\beta||_1 \right\}. \tag{1.3}$$

Through the l_1 norm, more and more coefficients will be driven to zero by increasing the value of λ. Thus, Lasso can automatically find a sparse model that includes more relevant features and discards the others.

Lasso with the l_1 norm regularization has achieved great success in many applications. However, in some cases, the explanatory factors used to predict the response variable are represented by a group of features but not just a single feature. For example, in microarray gene expression data analysis, these groups may be gene pathways. Then, the selection of relevant features is extended to the selection of groups of features. Yuan and Lin (2006) [YL06] introduced group Lasso for this problem. Suppose the p predictors can be divided into L different groups with p_l predictors in each group l, $\boldsymbol{X}_l$ is the sub matrix of $\boldsymbol{X}$ with columns corresponding to the predictors in group

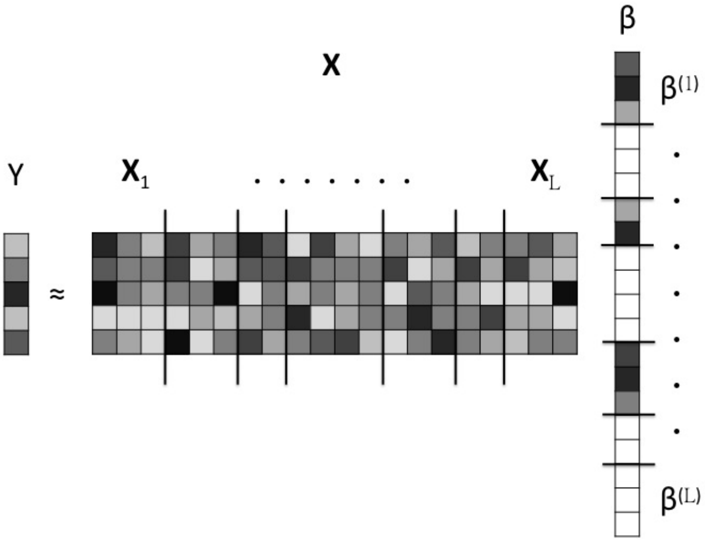


Figure 1.2: The group Lasso model with L groups for selection of relevant groups of features.

l, and $\beta^{(l)}$ is the coefficient vector of that group. It is illustrated in Figure 1.2. The introduced group Lasso criterion is

$$\arg\min_{\beta \in \mathbb{R}^n} \left\{ ||Y - \sum_{l=1}^{L} \boldsymbol{X_l} \beta^{(l)}||_2^2 + \lambda \sum_{l=1}^{L} \sqrt{p_l} ||\beta^{(l)}||_2 \right\}, \tag{1.4}$$

where the $\sqrt{p_l}$ terms accounts for the varying group size, and $|| \cdot ||_2$ is the Euclidean norm. In fact, If each group consists of just one variable, this reduces to the regular Lasso in Eq. (1.3).

The group Lasso gives a sparse set of groups with the tuning parameter λ controlling the sparsity level. Larger value of λ implies more regularization and entire predictors of some groups may be drop out of the model at the same time. In the contrary, if a group is included in the group Lasso model then all coefficients in the group will be nonzero. It means, the group Lasso considers just the group-wise sparsity but not the sparsity within each group. However, sometimes we would like

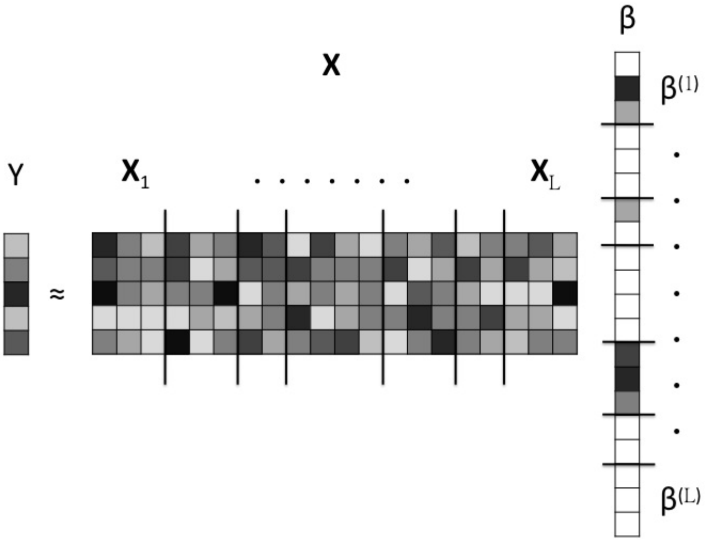


Figure 1.3: The sparse group Lasso model that considers both group-wise sparsity and sparsity within each group.

sparsity at both the group and individual predictor levels. For example, when we want to construct the land climate model using ocean climate variables in the climate research, not only the relevant location on the ocean, but also the particular important feature(s) at the location are the goals we are looking for. The revised model is illustrated in Figure 1.3, where the coefficient vectors corresponding to the active first and the third group have some zero terms.

To generalize group Lasso, Simon et al. (2012) [ST12] proposed the sparse group Lasso, and the estimator is given by

$$\arg\min_{\beta\in\mathbb{R}^n}\left\{||Y-\sum_{l=1}^{L}\boldsymbol{X_l}\beta^{(l)}||_2^2+\lambda_1\sum_{l=1}^{L}\sqrt{p_l}||\beta^{(l)}||_2+\lambda_2||\beta||_1\right\}, \tag{1.5}$$

where the regularization combines the Lasso and group Lasso penalties. If $\lambda_1 = 0$ its gives the Lasso criterion, while $\lambda_2 = 0$ gives the group Lasso criterion.

1.1.2 Bayesian Variable Selection

Consider the general linear regression

$$Y = \boldsymbol{X}\beta + \omega, \tag{1.6}$$

where Y corresponds to the $n \times 1$ response vector, $\boldsymbol{X} = [X_1, \cdots, X_p]$ corresponds to the $n \times p$ design matrix, β is the $p \times 1$ unknown coefficient vector, and ω is the $n \times 1$ random error. The random error ω is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$. The regression problem is trying to express the response variable with a number of the predictors. In sparse coding, the aim is to select a small and promising subset from the over-complete potential predictors, while controlling the trade-off between bias and variance.

To define a Bayesian approach for variable selection, the variable selection problem is considered as a model selection problem. In the 2^p possible models, each model S_δ is represented by a binary vector $\delta = (\delta_1, \cdots, \delta_p)$, where $\delta_j = 1$ indicates predictor X_j is included in the model. Let $p(\delta)$ denote the prior probability of model S_δ. Based on the Bayesian rule, the posterior probability of model S_δ is obtained through updating the prior probability with observation data Y and $\boldsymbol{X}$:

$$P(\delta|Y, \boldsymbol{X}) = \frac{P(\delta)L(Y|\delta, \boldsymbol{X})}{\sum_{\delta^\star} P(\delta^\star)L(Y|\delta^\star, \boldsymbol{X})}, \tag{1.7}$$

where $L(Y|\delta, \boldsymbol{X}) = \int L(Y|\beta_\delta, \boldsymbol{X})dP(\beta_\delta)$ is the marginal likelihood under model S_δ, and $L(Y|\beta_\delta, \boldsymbol{X},)$ is the likelihood of Y conditional on the coefficients β_δ in model S_δ. Eq. (1.7) describes the posterior probabilities for each of the candidate models, and these posterior probabilities also provide weights to be used in model selecting. In general, the goal is to find a single "best" model for further consideration.

George and McCulloch (1993) [GM93] proposed a stochastic search variable selection(SSVS) algorithm for normal linear regression. They used proposed Gibbs sampling to search the model with the highest posterior probability. In their approach, by using the latent variable $\delta_j = 1$ or 0, a normal mixture model with a low and a high variance centered at zero for each regression parameter β_j is represented:

$$\beta_j | \delta_j \sim (1 - \delta_j) N(0, \tau^2) + \delta_j N(0, \tau_j^2). \tag{1.8}$$

If $\delta_j = 0$, the coefficient β_j which corresponds to the normal distribution with very low variance τ^2, could be efficiently estimated by 0. In the other way, if $\delta_j = 1$, a non-zero estimate of β_j, which follows the normal distribution with high variance τ_j^2, should be included in the model.

In 1997, to improve the efficiency of computation, George and McCulloch modified SSVS by iteratively samples the predicator inclusion indicator for the jth predicator δ_j from its *Bernoulli* full conditional posterior distribution given the other predicators in the model, $\delta_{-j} = \{\delta_{:}, \neq j, = 1, \cdots, p\}$ for $= 1, \cdots, p$.

1.1.3 Gibbs sampling

Gibbs sampling is one of the Markov chain Monte Carlo (MCMC) algorithms for simulating a sequence of samples from the posterior distribution of a multivariate probability distribution, when the joint distribution is not know explicitly or is difficult to sample from directly. Suppose we have a joint distribution $p(\theta_1, \cdots, \theta_p)$ that we want to sample from. If we know the full conditional distribution for each variable, which is the distribution of the variable conditional on the known information and all the other variables: $p(\theta_j | \theta_{-j})$, we can use the Gibbs sampling to simulate samples from the joint distribution by sampling each variable in turn. The procedure is:

1. Begin with a vector of initial values $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_p^{(0)})$.

2. Repeat for $t = 1, 2, \cdots, T$.
   Generate $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)})$
   Generate $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)})$
   $\vdots$
   Generate $\theta_p^{(t)}$ from $p(\theta_p|\theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{p-1}^{(t)})$

3. Return samples $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(T)}\}$.

4. Discard samples in burn-in period.

In the algorithm, at each repeating procedure in step 2, each variable is sampled from the distribution conditional on the most recently generated parameter values in turn. Through exploiting the updating schemes, the sequence of the simulated samples will converge to a stationary distribution, which is the desired joint distribution. However, it may take a while for the stationary distribution to be reached. In order to discard samples that may not accurately represent the desired distribution, a burn-in period is commonly used to ignore samples from the beginning. Then just the left samples are considered.

1.2 Multi-task Learning

Multi-task learning [Car97] is a kind of machine learning that learns related tasks in parallel while using a shared representation. Based on the assumption that there are commonalities among related tasks, multi-task learning often leads to better performance than single-task learning. In fact, multi-task learning can be treated as an inductive mechanism. It improves generalization performance by using the information contained in the training signals of related tasks as inductive bias.

Assume we have M learning tasks and all data for the tasks come from the same space $\{\boldsymbol{X}, Y\}$, where $\boldsymbol{X} \subset \mathbb{R}^p$, and $Y \subset \mathbb{R}$. For each task $m \in \{1, \cdots, M\}$, we have n samples

$$\{(X_{1m}, y_{1m}), (X_{2m}, y_{2m}), \cdots, (X_{nm}, y_{nm})\}$$

sampled from a distribution f_m on $\{\boldsymbol{X}, Y\}$. So the total data available is:

$$\{(X_{11}, y_{11}), \cdots, (X_{n1}, y_{n1})\}, \cdots, \{(X_{1M}, y_{1M}), \cdots, (X_{nM}, y_{nM})\}.$$

We assume that f_m is different for each task, but that the f_m are related. The goal of multi-task learning is to learn M functions $\hat{f}_1, \cdots \hat{f}_M$ such that $\hat{f}_m(X_{im}) \approx y_{im}$. If $M = 1$, it becomes the standard single-task learning problem.

The multi-task learning problem has been studied in the statistics literature and shown the benefits of such multi-task learning relative to individual task learning when tasks are related. Based on the minimization of regularization functions that have been successfully used in single-task learning, Evgeniou and Pontil (2004) [EP04] presented an multi-task learning approach by molding the relation between tasks in terms of a novel kernel function. Obozinski, Taskar, and Jordan (2006) [OTJ06] proposed a novel type of joint regularization of the model parameters in order to couple feature selection across tasks. Argyriou, Evgeniou, and Pontil (2008) [AEP08] presented a method for learning sparse representation, which is shared across multiple related tasks. This method built upon the 1-norm regularization problem using a new regularizer, which controls the number of learned features common for all tasks. These methods work on the assumption that all tasks are related. However, this assumption can be violated in many real-world problems and reduce the performance. Some methods assume that tasks can be grouped in clusters and parameters of tasks within the same cluster are shared (Bakker and Heskes (2003) [BH03]; Kumar and Daum III(2012) [I12]).

In this paper, we focus on learning related tasks. We adopt a simpler setting in which the same input data X_{im} are used for all the tasks. It means, for every $i \in \{1, \cdots, n\}$ the vector X_{im} is the same for all $m \in \{1, \cdots, M\}$. However, the output values y_{im} are different for each m. Therefore, the data we need is

$$\{y_{11}, \cdots, y_{n1}\}, \cdots, \{y_{1M}, \cdots, y_{nM}\}, \{X_1, \cdots, X_n\}.$$

We wish to learn a low-dimensional representation which could be shared across multiple related tasks.

1.2.1 Group Lasso in multi-task learning

In order to deal with the situation of coupling multiple related tasks, Obozinski, Wainwright, and Jordan (2011) [OWJ11] has extended the idea of group lasso penalty [YL06] and introduced a group lasso method to recover the union of the supports $S_S = \bigcup_{m=1}^{M} S_m$ in the multiple linear regression, where S_m is the support set that contains the variables with nonzero coefficient for the m-th singular regression model. The multiple linear regression is written as

$$\boldsymbol{Y} = \boldsymbol{X}B + W,$$

where $\boldsymbol{Y}$ is a $n \times M$ response matrix, $\boldsymbol{X}$ is a $n \times p$ design matrix, B is a $p \times M$ matrix of the unknown regression coefficients, and W is a $n \times M$ noise term. This method is know as the L_1/L_2-regularized multi-task regression. In this case, the criterion to estimate the coefficient matrix is

$$\arg\min_{B\in\mathbb{R}^{p\times M}}\left(\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}B\|_F^2+\lambda\|B\|_{l_1/l_2}\right), \tag{1.9}$$

where $\|\cdot\|_F$ is the Frobenius norm, and $\|B\|_{l_1/l_2}$ is the block l_1/l_2 norm

$$\|B\|_{l_1/l_2}=\sum_{j=1}^{p}\left(\sum_{m=1}^{M}\beta_{jm}^2\right)^{1/2}=\sum_{j=1}^{p}\|\beta^j\|_2. \tag{1.10}$$

The L_2 norm is applied to the regression coefficients for all responses for each predictor, β^j, and these L_2 norms for the p predictors are combined through the L_1 norm. Because the L_1 part of penalty prefer sparse solutions, there is only a sparse set of predictors to have nonzero regression coefficients. The L_2 part of penalty doesn't encourage sparsity. Once a predictor is selected in the model, all entries in the corresponding coefficient vector for all responses will be nonzero, although the values are allowed to vary across different responses. Otherwise, the predictor is not relevant to any of the responses, and is drop out of the model. Therefore, the structure assumption in the multivariate group lasso is that all responses in the multiple linear regression are relevant to the same set of predictors.

The multivariate group lasso just consider the shared sparsity, but ignore the individual sparsity for each response. It is not realistic in many natural situations. We assume some responses can be irrelevant to the predictors in the union support, and would like to identify them. A sparse group lasso, the more general penalty consider sparsities in both ways, is introduced in SLEP [LJY09]. The criterion to solve the coefficient matrix is:

$$\arg\min_{B\in\mathbb{R}^{p\times M}}\left(\frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}B\|_2^2+\lambda_1\|B\|_1+\lambda_2\|B\|_{l_1/l_2}\right). \tag{1.11}$$

Besides the block l_1/l_2 norm, the sparse group lasso apply the L_1 part of penalty to all entries in the coefficient matrix B. It encourages the sparsity within the row vector that the group lasso doesn't cover.

CHAPTER 2

Bayesian Variable Selection in Multi-response Linear Regression

2.1 Problem Set-up and the Model

Consider the following multiple linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}B + W, \tag{2.1}$$

where $\boldsymbol{Y} = [Y_1, \cdots, Y_M]$ is a $n \times M$ response matrix of observations, $\boldsymbol{X} = [X_1, \cdots, X_p]$ $\in \mathbb{R}^{n \times p}$ is the fixed $n \times p$ design matrix, $B = [\beta_1, \cdots, \beta_M]$ is a $p \times M$ matrix of the unknown regression coefficients, and $W = [\omega_1, \cdots, \omega_M] \in \mathbb{R}^{n \times M}$ is the corresponding noise matrix. Here the error term ω_m is an $n \times 1$ noise vector that follows a multivariate normal distribution with zero mean vector and covariance matrix $\sigma^2 I_n$, where I_n is the n-dimensional identify matrix. Thus a group of M response vectors are to be regressed on the same design matrix $\boldsymbol{X}$. The model can also be written as

$$Y_m = \boldsymbol{X}\beta_m + \omega_m \sim N_n(\boldsymbol{X}\beta_m, \sigma^2 I_n),\ m = 1, \cdots, M, \tag{2.2}$$

where $\beta_m = (\beta_{1,m}, \cdots, \beta_{p,m})'$ is the coefficient vector for the m-th response vector Y_m. The estimation of each column of B, β_m, is a single linear regression problem with response vector Y_m and design matrix $\boldsymbol{X}$, and can be solved individually. However, in this study, we solve the M individual regression problems together by exploiting the similarities among β_m, or by imposing constraint on the matrix B.

In particular, we are interested in finding the sparse model for the multi-response model (2.2). Suppose S_m is the support set for the m-th response vector, i.e.

$$S_m = \{j \in \{1, \cdots, p\} \mid \beta_{j,m} \neq 0\}. \tag{2.3}$$

In some applications, the multiple response vectors could be related by a set of shared sparsity variables. It means the support sets S_m may be the same or similar for different m. Thus each response vector depends on variables specific to itself in addition to the ones that are shared. In this case, finding the set of variables which are related to any of the multiple response vectors simultaneously is more benefit than identifying S_m separately. Therefore, in the assumption of structural, it is natural to believe certain variables are related to several responses, corresponding to rows of B have many non-zero entries, while certain variables are relevant to some but not all, corresponding to rows would be element-wise sparse, while certain variables are not relevant to any responses, corresponding to rows have all zero entries.

Obozinski et al. (2011) [OWJ11] focus on the problem of recovering the union of the supports,

$$S_S = \cup_{m=1}^{M} S_m = \{j \in \{1, \cdots, p\} \mid \sum_m \beta_{j,m} \neq 0\}, \tag{2.4}$$

which is denoted as S_S, support union of the shared model in this paper. This support union of the shared model corresponds to the subset of indices $j \in \{1, \cdots, p\}$ that are included in at least one support set S_m, $m = 1, \cdots, M$. In this paper, we go deeper to the second level. Consider not only the group sparsity but also the individual sparsity, we are interested in identifying each nonzero entries in the true coefficient matrix B. Thus another union of the support sets is introduced here, we call it the support union of individual models, which is defined as

$$S_I = \{(j, m) \mid \beta_{j,m} \neq 0, j \in 1, \cdots, p, m \in 1, \cdots, M\}. \tag{2.5}$$

This support union of individual models, S_I, is a subset of pair of indices, which indicate which particular model is the shared variable active for. To target the ”support union recovery” problem, a Bayesian approach is adopted and the corresponding Bayesian algorithms are proposed to recover the unknown support sets S_S and S_I.

2.2 Group-Wise Gibbs Sampler

In support union recover problem, Obozinski et al. (2011) [OWJ11] set the group structure fro each variable across multiple response vectors, and the group Lasso approach was adopted. Consider the corresponding Bayesian approach, it is straightforward to apply Bayesian group selection algorithm to replace the group Lasso approach. Thus one set of the indicators is defined to denote whether X_j is active or not. Similar to group Lasso, we want to select the ”best” subset of variables from $X_1, \cdots, X_p$ to explain the multiple responses $Y_1, \cdots, Y_M$ simultaneously.

First, following SSVS in George and McCulloch (1993) [GM93], a $p \times 1$ vector of indicator variables, $\delta = (\delta_1, \ldots, \delta_p)'$, is introduced to indicate which variables are selected. It is defined as:

$$\delta_j = \begin{cases} 1, & \text{if } X_j \text{ is selected or active} \\ 0, & \text{if } X_j \text{ is not selected or inactive} \end{cases} \quad j = 1, \cdots p. \tag{2.6}$$

Consider the prior assumption for (δ_j, β_j). The prior distribution of δ_j is assumed to follow the *Bernoulli* distribution with

$$P(\delta_j) = \begin{cases} \theta_j, & \text{if } \delta_j = 0 \\ 1-\theta_j, & \text{if } \delta_j = 1 \end{cases} \quad j = 1, \cdots p. \tag{2.7}$$

Then the prior distribution of the coefficient $\beta_{j,m}$ given the indicator δ_j is set as

$$\beta_{j,m}|\delta_j \sim (1-\delta_j)\gamma_0 + \delta_j N(0, \tau_{j,m}^2), m = 1, \cdots, M, \tag{2.8}$$

where γ_0 is a point mass at 0. That is if X_j is inactive, i.e. $\delta_j = 0$, then $\beta_{j,m} = 0$ for all $m = 1, \cdots, M$. Otherwise $N(0, \tau_{j,m}^2)$ is the prior distribution of $\beta_{j,m}$. We also assume that the prior distribution of $(\delta_j, \beta_{j,m})$ are independent for $m = 1, \cdots, M, j = 1, \ldots, p$, and they are independent of the prior distribution of the residual variance σ^2, which is assumed to follow an inverse Gamma distribution, $\sigma^2 \sim IG(a/2, b/2)$.

Based on the prior assumptions, the sampling scheme of component-wise Gibbs sampler in Chen et al. (2011) [CCL11] is modified. We sample $(\delta_j, \beta_{j,1}, \cdots, \beta_{j,M})$ one at a time by fixing the other components δ_{-j} and $\beta_{-j,m}$, $m = 1, \cdots, M$, where δ_{-j} denotes all the indicators except δ_j, and $\beta_{-j,m}$ denotes all the coefficients for Y_m except $\beta_{j,m}$. Therefore in the Gibbs sampler, we need to computer the posterior probability $P(\delta_j = 1|\boldsymbol{Y}, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma)$, where the calculation of the likelihood ratio

$$\tilde{Z}_j = \frac{P(\mathbf{Y}|\delta_j = 1, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma)}{P(\mathbf{Y}|\delta_j = 0, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma)}$$

is the key step. Due to the independent assumption of $Y_1, \cdots, Y_M$, it is easy to show that

$$\begin{aligned}
\tilde{Z}_j &= \prod_{m=1}^{M} \frac{P(Y_m|\delta_j = 1, \delta_{-j}, \beta_{-j,m}, \sigma)}{P(Y_m|\delta_j = 0, \delta_{-j}, \beta_{-j,m}, \sigma)} \\
&= \prod_{m=1}^{M} \frac{\int P(Y_m|\delta_j = 1, \delta_{-j}, \beta_{j,m}, \beta_{-j,m}, \sigma) P(\beta_{j,m}|\delta_j = 1) d\beta_{j,m}}{\int P(Y_m|\delta_j = 0, \delta_{-j}, \beta_{j,m}, \beta_{-j,m}, \sigma) P(\beta_{j,m}|\delta_j = 0) d\beta_{j,m}} \\
&= \prod_{m=1}^{M} \frac{\sigma}{\sqrt{X_j' X_j \tau_{j,m}^2 + \sigma^2}} \cdot \exp\left\{ \frac{(R_{j,m}' X_j)^2 \tau_{j,m}^2}{2\sigma^2(\sigma^2 + X_j' X_j \tau_{j,m}^2)} \right\} \qquad (2.9)
\end{aligned}$$

where $R_{j,m} = Y_m - \sum_{i \neq j} X_i \beta_{i,m}$. Define $r_{j,m} = \frac{R_{j,m}' X_j \tau_{j,m}^2}{\sigma^2 + X_j' X_j \tau_{j,m}^2}$, and $\sigma_{j,m}^{\star 2} = \frac{\sigma^2 \tau_{j,m}^2}{X_j' X_j \tau_{j,m}^2 + \sigma^2}$. We then can rewrite $\tilde{Z}_j$ as:

$$\tilde{Z}_j = \prod_{m=1}^{M} \sqrt{\sigma_{j,m}^{\star 2} / \tau_{j,m}^2} \exp\left\{ \frac{r_{j,m}^2}{2\sigma_{j,m}^{\star 2}} \right\}, \qquad (2.10)$$

The group-wise Gibbs sampler is described in Algorithm 1. In practice, we start from the null model and then iterate the steps in Algorithm 1 to generate the posterior samples of δ_j, β_j, for the posterior inference.

Algorithm 1: Group-Wise Gibbs Sampler for Support Recovery

1. Randomly select a variable X_j. Compute $R_{j,m} = Y_m - \sum_{i \neq j} X_i \beta_{i,m}$, for $m = 1, \cdots, M$.

2. Compute the likelihood ratio $\tilde{Z}_j$ according to Eq. (2.10), and then evaluate the posterior probability of δ_j

$$P(\delta_j = 1|\boldsymbol{Y}, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma) = \frac{(1-\theta_j)\tilde{Z}_j}{(1-\theta_j)\tilde{Z}_j + \theta_j}. \qquad (2.11)$$

3. Sample δ_j based on the posterior probability in (2.11). If $\delta_j = 0$, then set $\beta_{j,m} = 0$, $m = 1, \cdots, M$, otherwise, sample $\beta_{j,m} \sim N(r_{j,m}, \sigma_{j,m}^{\star 2})$.

4. After repeat above steps for all variables, compute the current residual matrix, $Res = \boldsymbol{Y} - \boldsymbol{X}B$. Then sample $\sigma^2 \sim IG(\frac{a+n\times M}{2}, \frac{\sum(diaq(Res'Res))+b}{2})$. Go to Step 1.

2.3 Two-Layer Structure and Two-Layer Gibbs sampler

In the group selection methods, once a variable, X_j, is selected, then X_j is active for all the responses, $Y_1, \cdots, Y_M$. However, we can further assume that the selected variable might not be active for all response vectors simultaneously. In other words, we are interested in finding the best union of supports sets, S_S, and we also assume that the variable in S_S might be inactive for some response vectors. Therefore, unlike the single indicator set-up in the group-wise Gibbs sampler, two nested sets of binary indicator variables are used. The first set of indicators $\delta = (\delta_1, \cdots, \delta_p)'$ is associated with variables, $X_1, \cdots, X_p$, respectively, and δ_j is defined to indicate if the variable, X_j, is active for any of the response vectors. Specifically if $\delta_j = 1$, then the variable X_j is selected, and $\delta_j = 0$ otherwise. In the second set of indicators, each indicator is associated with a variable an a response vector, indicating whether this variable is active for explaining the particular response vector. Thus for each variable X_j, we define the indicator vector $\eta^{(j)} = (\eta_{j,1}, \cdots, \eta_{j,M})$, and if $\eta_{j,m} = 1$, the variable X_j is active for the m-th response, Y_m, and $\eta_{j,m} = 0$ otherwise.

Similar the the group-wise Gibbs sampler, the prior distribution of δ_j is also assumed to follow the *Bernoulli* distribution with $P(\delta_j = 0) = \theta_j$ and $P(\delta_j = 1) = 1 - \theta_j$, i.e. $Ber(1 - \theta_j)$. Consider the prior assumption for the second set of indicators. Following Chen et al. (2014) [CCCnt], the prior distribution of the indicator in the second set, $\eta_{j,m}$, is chosen as a mixture distribution depending on the indicator in the first set: δ_j, and is represented as

$$\eta_{j,m}|\delta_j \sim (1-\delta_j)\gamma_0 + \delta_j Ber(1-\rho_{j,m}), \tag{2.12}$$

where $P(\eta_{j,m}=0)=\rho_{j,m}$. Based on Eq. (2.12), if the j-th variable, X_j, is not selected in S_S, i.e. $\delta_j = 0$, then $\eta_{j,m}=0$ for all $m=1,\cdots,M$, however, when $\delta_j=1$, $\eta_{j,m}$ still could be 0 or 1 due to the *Bernoulli* prior distribution. Then for the coefficient, $\beta_{j,m}$, given the indicators δ_j and $\eta_{j,m}$, the prior distribution of $\beta_{j,m}$ can be defined as

$$\beta_{j,m}|\delta_j,\eta_{j,m} \sim (1-\delta_j\eta_{j,m})\gamma_0 + \delta_j\eta_{j,m}\mathcal{N}(0,\tau_{j,m}^2), \tag{2.13}$$

where γ_0 is a point mass at 0. That is the prior of $\beta_{j,m}$ is $N(0,\tau_{j,m}^2)$ only when $\delta_j=1$ and $\eta_{j,m}=1$, i.e. X_j is in S_S and is active for the m-th response Y_m. Otherwise $\beta_{j,m}$ is set to be zero. In fact, this coefficient prior has also been used in Chen et al. (2014) [CCCnt]. For the prior assumption on the noise variance σ^2, as usual, we choose the inverse gamma conjugate prior $\sigma^2 \sim IG(a/2,b/2)$. Finally in the prior distribution, $(\delta_j,\eta_{j,m},\beta_{j,m}),\ j=1,\cdots,p$ are assumed to be independent and given $\delta_j=1$, $(\eta_{j,m},\beta_{j,m}),\ m=1,\cdots,M$ are assumed to be independent of each others, too.

Based on the prior set-up, we can use Gibbs sampler to draw posterior samples of the indicators and the coefficients. Similar to group-wise Gibbs sampler in Algorithm 1, the key step is to compute the likelihood ratios of the indicators in the first and second sets respectively, and then the posterior probabilities for $\delta_j=1$ and $\eta_{j,m}=1$ can be computed accordingly. Thus we can sample these indicators from the corresponding posterior *Bernoulli* distributions. First, consider the multi-response model in Eq. (2.1). Based on the assumption of independence between $Y_1,\cdots,Y_M$, the likelihood ration Z_j of the variable X_j is represented as

$$\begin{aligned} Z_j &= \frac{P(\mathbf{Y}|\delta_j = 1, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma)}{P(\mathbf{Y}|\delta_j = 0, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma)} \\ &= \prod_{m=1}^{M} \frac{P(Y_m|\delta_j = 1, \delta_{-j}, \beta_{-j,m}, \sigma)}{P(Y_m|\delta_j = 0, \delta_{-j}, \beta_{-j,m}, \sigma)} \end{aligned} \tag{2.14}$$

Let $k = \{(k_1, \cdots, k_M) : \ k_m = 0 \text{ or } 1, m = 1, \cdots, M\}$ denote the set of all possible combinations of $(\eta_{j,1}, \cdots, \eta_{j,M})$. It is easy to show that

$$Z_j = \sum_{k=(k_1,\cdots,k_M)} (\prod_{m=1}^{M} b_{j,k_m}),$$

where

$$b_{j,k_m} = \frac{\int P(Y_m|\beta_{j,m}, \eta_{j,m} = k_m, \delta_j = 1, \delta_{-j}, \beta_{-j,m}, \sigma) P(\beta_{j,m}, \eta_{j,m} = k_m|\delta_j = 1) d\beta_{j,m}}{P(Y_m|\delta_j = 0, \delta_{-j}, \beta_{-j,m}, \sigma)}.$$

If $k_m = 0$, then we can simply obtain $b_{j,k_m=0} = \rho_{j,m}$. When $k_m = 1$, then

$$\begin{aligned} b_{j,k_m=1} &= \frac{\frac{(1-\rho_{j,m})}{\sqrt{2\pi\tau_{j,m}^2}} \int \exp\left\{-\frac{1}{2\sigma^2}(R_{j,m} - \beta_{j,m}X_j)'(R_{j,m} - \beta_{j,m}X_j) - \frac{\beta_{j,m}^2}{2\tau_{j,m}^2}\right\} d\beta_{j,m}}{\exp\left(-\frac{1}{2\sigma^2}R_{j,m}'R_{j,m}\right)} \\ &= \frac{(1-\rho_{j,m})}{\sqrt{2\pi\tau_{j,m}^2}} \cdot \int \exp\left\{(-\frac{1}{2\tau_{j,m}^2} - \frac{1}{2\sigma^2}X_j'X_j)\beta_{j,m}^2 + \frac{1}{\sigma^2}R_{j,m}'X_j\beta_{j,m}\right\} d\beta_{j,m} \\ &= \frac{(1-\rho_{j,m})}{\sqrt{2\pi\tau_{j,m}^2}} \cdot \exp\left\{\frac{(\frac{1}{\sigma^2}R_{j,m}'X_j)^2}{2(\frac{1}{\tau_{j,m}^2} + \frac{1}{\sigma^2}X_j'X_j)}\right\} \cdot \sqrt{\frac{2\pi}{\frac{X_j'X_j}{\sigma^2} + \frac{1}{\tau_{j,m}^2}}} \\ &= \frac{(1-\rho_{j,m})\sigma}{\sqrt{X_j'X_j\tau_{j,m}^2 + \sigma^2}} \cdot \exp\left\{\frac{(R_{j,m}'X_j)^2\tau_{j,m}^2}{2\sigma^2(\sigma^2 + X_j'X_j\tau_{j,m}^2)}\right\} \\ &= (1-\rho_{j,m}) \times \sqrt{\sigma_{j,m}^{\star 2}/\tau_{j,m}^2} \exp\left\{\frac{r_{j,m}^2}{2\sigma_{j,m}^{\star 2}}\right\}. \end{aligned} \tag{2.15}$$

Thus the likelihood ration Z_j of the indicator δ_j can be represented as

$$Z_j = \sum_{k=(k_1,\cdots,k_M)} \left\{\prod_{m=1}^{M} \left[\left((1-\rho_{j,m}) \times \sqrt{\sigma_{j,m}^{\star 2}/\tau_{j,m}^2} \exp\left\{\frac{r_{j,m}^2}{2\sigma_{j,m}^{\star 2}}\right\}\right)^{k_m} \cdot \rho_{j,m}^{(1-k_m)}\right]\right\}. \tag{2.16}$$

Once X_j is not selected, then we can simply set $\eta_{j,m} = 0$ and $\beta_{j,m} = 0$ for all $m = 1, \cdots, M$. Otherwise, if the variable X_j is included in S_S, i.e. $\delta_j = 1$, then we need to check if X_j is active or not for each individual response Y_m separately. Following the component-wise Gibbs sampler in Chen et al. (2011) [CCL11], the likelihood ratio $Q_{j,m}$ of the variable X_j with respect to the m-th model, Y_m, is computed and can be show as

$$\begin{aligned}
Q_{j,m} &= \frac{P(Y_m|\eta_{j,m} = 1, \eta_{-j,m}, \beta_{-j,m}, \sigma, \delta_j = 1)}{P(Y_m|\eta_{j,m} = 0, \eta_{-j,m}, \beta_{-j,m}, \sigma\delta_j = 1)} \\
&= \frac{\int P(Y_m|\beta_{j,m}, \eta_{j,m} = 1, \eta_{-j,m}, \beta_{-j,m}, \sigma\delta_j = 1)P(\beta_{j,m}|\eta_{j,m} = 1\delta_j = 1)d\beta_{j,m}}{P(Y_m|\eta_{j,m} = 0, \eta_{-j,m}, \beta_{-j,m}, \sigma\delta_j = 1)} \\
&= \frac{\sigma}{\sqrt{X_j'X_j\tau_{j,m}^2 + \sigma^2}} \cdot \exp\left\{\frac{(R_{j,m}'X_j)^2\tau_{j,m}^2}{2\sigma^2(\sigma^2 + X_j'X_j\tau_{j,m}^2)}\right\} \\
&= \sqrt{\sigma_{j,m}^{\star 2}/\tau_{j,m}^2}\exp\left\{\frac{r_{j,m}^2}{2\sigma_{j,m}^{\star 2}}\right\}. \qquad (2.17)
\end{aligned}$$

Based on both likelihood ratio functions, Eq. (2.16) and Eq. (2.17), the corresponding posterior probabilities of $\delta_j = 1$ and $\eta_{j,m} = 1$ can be derived. The proposed Gibbs sampling algorithm is summarized in Algorithm 2. Note that we would start from the null model by setting $\delta_j = 0; \eta_{j,m} = 0$ and $\beta_{j,m} = 0$ for all j and m. Based on our experiences, this initial model works well.

Algorithm 2: The Two-Layer Gibbs Sampler for Support Recovery

1. Randomly select a variable X_j. Compute $R_{j,m} = Y_m - \sum_{i \neq j} X_i\beta_{i,m}$, for $m = 1, \cdots, M$.

2. Compute the likelihood ratio Z_j according to Eq. (2.16), and then evaluate the posterior probability of δ_j

$$P(\delta_j = 1|\boldsymbol{Y}, \delta_{-j}, \{\beta_{-j,m}, m = 1, \cdots, M\}, \sigma) = \frac{(1-\theta_j)Z_j}{(1-\theta_j)Z_j + \theta_j}. \qquad (2.18)$$

3. Sample δ_j based on the posterior probability in (2.18). If $\delta_j = 0$, then set $\eta_{j,m} = 0$, and $\beta_{j,m} = 0$, for all $m = 1, \cdots, M$. Otherwise, for each $m = 1, \cdots, M$, compute the likelihood ratio $Q_{j,m}$ according to Eq. (2.17), and sample $\eta_{j,m}$ based on the posterior probability

$$P(\eta_{j,m} = 1 | Y_m, \eta_{-j,m}, \beta_{-j,m}, \sigma, \delta_j = 1) = \frac{(1-\rho_{j,m})Q_{j,m}}{(1-\rho_{j,m})Q_{j,m} + \rho_{j,m}}. \qquad (2.19)$$

   If $\eta_{j,m} = 0$, set $\beta_{j,m} = 0$; otherwise, sample $\beta_{j,m} \sim N(r_{j,m}, \sigma^{\star 2}_{j,m})$.

4. After repeat above steps for all variables, compute the current residual matrix, $Res = \boldsymbol{Y} - \boldsymbol{X}B$. Then sample $\sigma^2 \sim IG(\frac{a+n\times M}{2}, \frac{\sum(diaq(Res'Res))+b}{2})$. Go to Step 1.

2.4 Sample Version of Two-Layer Gibbs Sampler

In Algorithm 2, the computation of Z_j in Eq. (2.16) involves 2^M cases and can be computational expensive, especially when the number of the responses, M, is large. To save computational cost, instead of deciding whether the j-th variable, X_j, is selected or not based on the posterior probability in Eq. (2.18) directly, we adopt another method as below. If the current variable is not selected in the support union of the shared model, i.e., $\delta_j = 0$, we propose to active this variable first by setting $\delta_j = 1$, and sample the individual indicators $\eta_{j,m}$ and coefficients $\beta_{j,m}$ from the corresponding conditional distributions via the component-wise Gibbs sampling approach in Chen et al. (2011) [CCL11], i.e. the Step 3 in Algorithm 2. We then decide whether to keep the sampled indicators and coefficients via the Metropolis-Hasting acceptance-rejection rule. Conversely, if the variable is selected in S_S, i.e. $\delta_j = 1$, we then propose to turn down this indicator by switching δ_j to 0, and setting all the corresponding indicators $\eta_{j,m}$ and coefficients $\beta_{j,m}$ to be zero. Therefore, we determine whether to accept this proposal or not via the Metropolis-Hastings acceptance-rejection rule, too. Thus this proposal method can be treated as the sample version of the two-ayer Gibbs sampler. The details of these stages are shown in following.

Let $\Theta_j = (\delta_j, \eta^{(j)}, \beta^{(j)})$ be the parameter set of the j-th variable, X_j, where $\eta^{(j)} = (\eta_{j,1}, \cdots, \eta_{j,M})$, and $\beta^{(j)} = (\beta_{j,1}, \cdots, \beta_{j,M})$ are the corresponding second set indicators and coefficients. The proposed transition of Θ_j can be defined as

$$T(\Theta_j^0 \to \Theta_j^1) \qquad = P(\hat{\beta}^{(j)}, \hat{\eta}^{(j)} | R^{(j)}, \delta_j = 1, \sigma) \tag{2.20}$$

$$T(\Theta_j^1 \to \Theta_j^0) \qquad = 1, \tag{2.21}$$

where $\Theta_j^0 = (\delta_j = 0, \eta^{(j)} = \mathbf{0}, \beta^{(j)} = \mathbf{0})$, $\Theta_j^1 = (\delta_j = 1, \hat{\eta}^{(j)}, \hat{\beta}^{(j)})$, $R^{(j)} = (R_{j,1}, \cdots, R_{j,M})$, and $\{\beta^{(j)}, \hat{\eta}^{(j)}\}$ are sampled from the joint posterior distribution. Here $T(\Theta_j^0 \to \Theta_j^1)$ is the proposal distribution for changing δ_j from 0 to 1, and $T(\Theta_j^1 \to \Theta_j^0)$ is the proposal distribution to switch δ_j to 0. Suppose the variable X_j is not included in S_S currently, i.e. $\delta_j = 0$. Then after sampling $\hat{\eta}_{j,m}$ and $\hat{\beta}_{j,m}$ by setting $\delta_j = 1$, we calculate the acceptance probability $\hat{A}_j$ as:

$$\begin{aligned}
& \hat{A}_j(\Theta_j^0 \to \Theta_j^1) && (2.22)\\
= \; & \frac{P(\Theta_j^1)}{P(\Theta_j^0)} \cdot \frac{T(\Theta_j^1 \to \Theta_j^0)}{T(\Theta_j^0 \to \Theta_j^1)} \\
= \; & \frac{P(\delta_j = 1, \hat{\eta}^{(j)}, \hat{\beta}^{(j)} | \boldsymbol{Y}, \delta_{-j}, \eta^{(-j)}, \beta^{(-j)}, \sigma)}{P(\delta_j = 0, \eta^{(j)} = \mathbf{0}, \beta^{(j)} = \mathbf{0} | \boldsymbol{Y}, \delta_{-j}, \eta^{(-j)}, \beta^{(-j)}, \sigma)} \cdot \frac{1}{P(\hat{\beta}^{(j)}, \hat{\eta}^{(j)} | R^{(j)}, \delta_j = 1, \sigma)} \\
= \; & \left(\prod_{m=1}^{M} \frac{P(Y_m | \hat{\beta}_{j,m}, \hat{\eta}_{j,m}, \delta_j = 1, \delta_{-j}, \beta_{-j,m}, \sigma) P(\hat{\beta}_{j,m}, \hat{\eta}_{j,m} | \delta_j = 1)}{P(Y_m | \delta_j = 0, \delta_{-j}, \beta_{-j,m}, \sigma)} \right) \times \frac{1-\theta_j}{\theta_j} \\
& \times \frac{1}{\prod_{m=1}^{M} P(\hat{\beta}_{j,m} | \hat{\eta}_{j,m}, R_{j,m}, \delta_j = 1, \sigma) P(\hat{\eta}_{j,m} | R_{j,m}, \delta_j = 1, \sigma)} \\
= \; & \prod_{m=1}^{M} \left[\left(\frac{1-\rho_{j,m}}{\sqrt{2\pi\tau_{j,m}^2}} \cdot \exp\left(-\frac{\sigma^2 + \tau_{j,m}^2 X_j' X_j}{2\tau_{j,m}^2 \sigma^2} \hat{\beta}_{j,m}^2 + \frac{R_{j,m}' X_j}{\sigma^2} \hat{\beta}_{j,m} \right) \right)^{\hat{\eta}_{j,m}} \cdot \rho_{j,m}^{(1-\hat{\eta}_{j,m})} \right] \\
& \times \prod_{m=1}^{M} \left[\left(\frac{1}{p_{j,m}} \sqrt{2\pi\sigma_{j,m}^{*2}} \cdot \exp(\frac{(\hat{\beta}_{j,m} - r_{j,m})^2}{2\sigma_{j,m}^{*2}}) \right)^{\hat{\eta}_{j,m}} \left(\frac{1}{1-p_{j,m}} \right)^{(1-\hat{\eta}_{j,m})} \right] \\
& \times \frac{1-\theta_j}{\theta_j} \\
= \; & \prod_{m=1}^{M} \left[\left(\frac{1-\rho_{j,m}}{p_{j,m}} \cdot \frac{\sigma_{j,m}^*}{\tau_{j,m}} \cdot \exp\left(-\frac{\sigma^2 + \tau_{j,m}^2 X_j' X_j}{2\tau_{j,m}^2 \sigma^2} \hat{\beta}_{j,m}^2 + \frac{R_{j,m}' X_j}{\sigma^2} \hat{\beta}_{j,m} + \frac{(\hat{\beta}_{j,m} - r_{j,m})^2}{2\sigma_{j,m}^{*2}} \right) \right)^{\hat{\eta}_{j,m}} \right] \\
& \times \prod_{m=1}^{M} \left[\left(\frac{\rho_{j,m}}{1-p_{j,m}} \right)^{(1-\hat{\eta}_{j,m})} \right] \cdot \frac{1-\theta_j}{\theta_j} && (2.23)
\end{aligned}$$

where $p_{j,m} = P(\eta_{j,m} = 1 | R_{j,m}, \delta_j = 1, \sigma) = \frac{(1-\rho_{j,m})Q_{j,m}}{(1-\rho_{j,m})Q_{j,m} + \rho_{j,m}}$, $\eta^{(-j)}$ denotes all the second set indicator vectors except $\eta^{(j)}$, and $\beta^{(-j)}$ denotes all the coefficient vectors except $\beta^{(j)}$. So based on Metropolis-Hastings acceptance-rejection rule, we accept the proposed samples, $\delta_j = 1$ and $(\beta^{(j)}, \eta^{(j)}) = (\hat{\beta}^{(j)}, \hat{\eta}^{(j)})$, with probability $P_{add} = \min\{1, \hat{A}_j\}$. Otherwise if the variable X_j is active already, that is $\delta_j = 1$, then based on the current $\beta^{\star(j)} = (\beta^\star_{j,1}, \cdots, \beta^\star_{j,M})$ and $\eta^{\star(j)} = (\eta^\star_{j,1}, \cdots, \eta^\star_{j,M})$, we have

$$\begin{aligned}
& \hat{D}_j(\Theta^1_j \to \Theta^0_j) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(2.24)} \\
= \; & \frac{P(\Theta^0_j)}{P(\Theta^1_j)} \cdot \frac{T(\Theta^0_j \to \Theta^1_j)}{T(\Theta^1_j \to \Theta^0_j)} \\
= \; & \frac{P(\delta_j = 0, \eta^{(j)} = \mathbf{0}, \beta^{(j)} = \mathbf{0} | \boldsymbol{Y}, \delta_{-j}, \eta^{(-j)}, \beta^{(-j)}, \sigma)}{P(\delta_j = 1, \eta^{\star(j)}, \beta^{\star(j)} | \boldsymbol{Y}, \delta_{-j}, \eta^{(-j)}, \beta^{(-j)}, \sigma)} \cdot P(\beta^{\star(j)}, \eta^{\star(j)} | R^{(j)}, \delta_j = 1, \sigma) \\
= \; & \left(\prod_{m=1}^{M} \frac{P(Y_m | \delta_j = 0, \delta_{-j}, \beta_{-j,m}, \sigma)}{P(Y_m | \beta^*_{j,m}, \eta^*_{j,m}, \delta_j = 1, \delta_{-j}, \beta_{-j,m}, \sigma) P(\beta^*_{j,m}, \eta^*_{j,m} | \delta_j = 1)} \right) \times \frac{\theta_j}{1 - \theta_j} \\
& \times \prod_{m=1}^{M} P(\beta^*_{j,m} | \eta^*_{j,m}, R_{j,m}, \delta_j = 1, \sigma) P(\eta^*_{j,m} | R_{j,m}, \delta_j = 1, \sigma) \\
= \; & \prod_{m=1}^{M} \left[\left(\frac{p_{j,m}}{1 - \rho_{j,m}} \cdot \frac{\tau_{j,m}}{\sigma^*_{j,m}} \cdot \exp\left(\frac{\sigma^2 + \tau^2_{j,m} X'_j X_j}{2\tau^2_{j,m}\sigma^2} \beta^{\star 2}_{j,m} - \frac{R'_{j,m} X_j}{\sigma^2} \beta^\star_{j,m} - \frac{(\beta^\star_{j,m} - r_{j,m})^2}{2\sigma^{*2}_{j,m}} \right) \right)^{\eta^\star_{j,m}} \right] \\
& \times \prod_{m=1}^{M} \left[\left(\frac{1 - p_{j,m}}{\rho_{j,m}} \right)^{(1 - \eta^\star_{j,m})} \right] \cdot \frac{\theta_j}{1 - \theta_j}. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(2.25)}
\end{aligned}$$

Thus the probability of accepting the proposed to remove the variable X_j from S_S is $P_{del} = \min\{1, \hat{D}_j\}$.

The modified algorithm is shown in Algorithm 3. As mentioned before, in this algorithm, component-wise Gibbs sampler is used to generate the proposal samples of $\eta_{j,m}$ and $\beta_{j,m}$ for the corresponding response Y_m individually.

Algorithm 3: Sample Version of Two-Layer Gibbs Sampler for Support Union Recovery

1. Randomly select a variable X_j. Compute $R_{j,m} = Y_m - \sum_{i \neq j} X_i \beta_{i,m}$, for $m = 1, \cdots, M$.

2. If $\delta_j = 0$, sample $\{(\hat{\eta}_{j,m}, \hat{\beta}_{j,m}), m = 1, \cdots, M\}$ based on the component-wise Gibbs sample (step 3 in Algorithm 2) through $R_{j,m}$. Compute $\hat{A}_j$ in Eq. (2.23).

Switch δ_j from 0 to 1 with probability $P_{add} = \min\{1, \hat{A}_j\}$. If $\delta_j = 1$, set $\eta_{j,m} = \hat{\eta}_{j,m}$, $\beta_{j,m} = \hat{\beta}_{j,m}$, $m = 1, \cdots, M$.

3. If $\delta_j = 1$, suppose the current coefficients and indicators in the second set are $\beta_{j,m} = \beta^{\star}_{j,m}$ and $\eta_{j,m} = \eta^{\star}_{j,m}$, $m = 1, \cdots, M$. Compute $\hat{D}_j$ in Eq. (2.25). Change δ_j from 1 to 0 with the probability $P_{del} = \min\{1, \hat{D}_j\}$. If the proposal is rejected, it means the variable X_j is kept in the shared model. Then we can re-sample $\eta_{j,m}$ and $\beta_{j,m}$, $m = 1, \cdots, M$ for each individual regression model according to the component-wise Gibbs sampler by $R_{j,m}$.

4. After repeat above steps for all variables, compute the current residual matrix, $Res = \boldsymbol{Y} - \boldsymbol{X}B$. Then sample $\sigma^2 \sim IG(\frac{a+n\times M}{2}, \frac{\sum(diaq(Res'Res))+b}{2})$. Go to Step 1.

CHAPTER 3

Simulation

In this chapter, through simulation examples, we illustrate the performance of the proposed two-layer Gibbs sampler for support union recovery in multi-response linear regression.

3.1 Group-Wise v.s. Two-Layer Gibbs Sampler

Example 3.1 Consider a multi-response linear regression example with $M = 3$ response vectors. In this example, there are $p = 50$ predictor variables of length $n = 80$. The predictor variables are defined by

$$X_j = G_j + kG, \tag{3.1}$$

where k is a pre-specified constant, and G_j's and G are independently generated from multivariate normal distribution with zero mean vector and identical covariance matrix I_{80}. Here we set $k = 1$, then the correlation between any two variables is 0.5. The true active variables in the shared model are $X_7, X_8, X_9, X_{11}, X_{12}$, i.e. $S_S = \{7, 8, 9, 11, 12\}$, and the corresponding coefficients of 3 single regression models are shown in Table 3.1. Some variables in S_S are not active in all individual regression models. The other coefficients are all set to be zero. Then each response vector Y_m

is generated according to the linear model $Y_m = \boldsymbol{X}\beta_m + \omega_m$ where $\omega_m \sim N_{80}(\boldsymbol{0}, I_{80})$.

Table 3.1: Example 3.1: The true coefficients of the support union in the shared model

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$
X_7	1.5	1.7	0
X_8	1.5	1.7	2.2
X_9	1.5	0	2.2
X_{11}	3.2	2.5	4.1
X_{12}	3.2	0	4.1

The group-wise Gibbs sampler, Algorithm 1, and two-layer Gibbs sampler, Algorithm 2, are used in this example. In the group-wise Gibbs sampler, the first set of indicator variables, $\delta_j, j = 1, \cdots, p$, are only adopted in the model, while in the two-layer Gibbs sampler, in addition to δ_j, the second set of indicator variables, $\eta_{j,m}, j = 1, \cdots, p,\ m = 1, \cdots, M$ are also added into the model. The prior parameters are set as $\theta_j = P(\delta_j = 0) = 0.5,\ \rho_{j,m} = P(\eta_{j,m} = 0|\delta_j = 1) = 0.5,\ \tau^2_{j,m} = 20$ for all $j \in \{1, \cdots, 50\},\ m \in \{1, 2, 3\}$, and $a = b = 0.001$ as the non-informative parameter for inverse *Gamma* prior for σ^2. The initial model is set as null model, i.e. $\delta_j = 0$; $\eta_{j,m} = 0$ and $\beta_{j,m} = 0$ for all j and m. Totally we run 500 sweeps. After discarding the first 300 sweeps, samples collected from the last 200 sweeps are used for the inference about support union recovery. First the posterior probabilities $P(\delta_j = 1|\boldsymbol{Y})$ for Algorithm 1 and Algorithm 2, and $P(\eta_{j,} = 1|\delta_j = 1, \boldsymbol{Y})$ for Algorithm 2, are estimated based on the posterior samples. Then for the posterior inference, the median probability criterion is used according to Barbieri and Berger (2004)[BB04]. Thus the threshold probabilities for including predictor in the shared and individual model are both set to 0.5, i.e. $\hat{P}(\delta_j = 1|\boldsymbol{Y}) \geqslant 0.5$ and $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y}) \geqslant 0.5$.

The posterior probabilities of $\hat{P}(\delta_j = 1|\boldsymbol{Y})$ for Algorithm 1 and Algorithm 2 are shown in Figure 3.1. In fact, in both algorithms, the estimated posterior probabilities of $X_7, X_8, X_9, X_{11}, X_{12}$ are all higher than 0.5, and even equal to 1. Therefore, for the support union in the shared model S_S, the selection results of the support union recovery in group-wise Gibbs sampler and two-layer Gibbs sampler both agree with the true model. There is one phenomenon that the posterior probabilities, $\hat{P}(\delta_j = 1|\boldsymbol{Y})$, of the inactive variables in the two-layer Gibbs sampler, i.e. the bottom figure in Figure 3.1, are generally higher than those in the group-wise Gibbs sampler, i.e. the top figure in Figure 3.1. The reason is the likelihood ratio, the key step in calculating the the posterior probability $\hat{P}(\delta_j = 1|\boldsymbol{Y})$. The likelihood ratio of indicators in the first set in the two-layer Gibbs sampler, i.e. Z_j in Eq. 2.16, consider all possible combination of indicators in the second set. However, because only indicators in the first set are adopted in group-wise Gibbs sampler, the likelihood ratio, $\tilde{Z}_j$ in Eq. 2.10, just consider the situation that the variable X_j is active or inactive for all regression models simultaneously. Therefore, Z_j is higher than $\tilde{Z}_j$, and then the posterior probabilities in two-layer Gibbs sampler are higher than those in group-wise Gibbs sampler.

Then, for the two-layer Gibbs sampler, i.e. Algorithm 2, $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ for $j \in S_S$ are shown in Figure 3.2. For those nonzero coefficients in Table 3.1, all corresponding indicators in the second set have posterior probability higher than 0.5. Therefore, based on the median probability criterion, these are treated as active.

Thus, both the group-wise Gibbs sampler and two-layer Gibbs sampler can successfully recover the support union in the shared model, S_S, correctly. In addition, by using the second set of indicators $\eta_{j,m}$, the two-layer Gibbs sampler even indicate

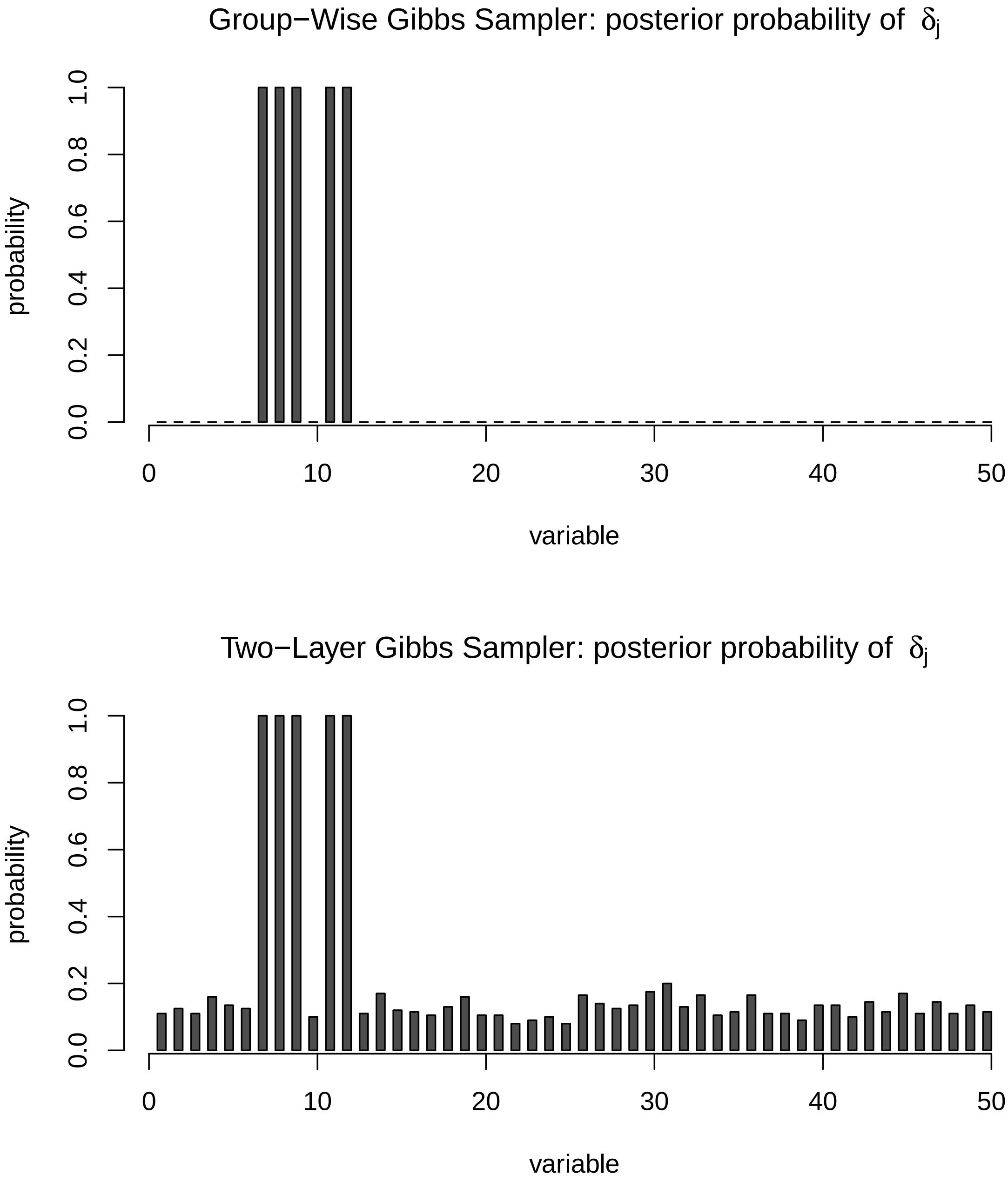


Figure 3.1: Example 3.1: The estimated posterior probabilities of δ_j : $\hat{P}(\delta_j = 1|\boldsymbol{Y})$. Top: Group-wise Gibbs sampler. Bottom: Two-layer Gibbs sampler.

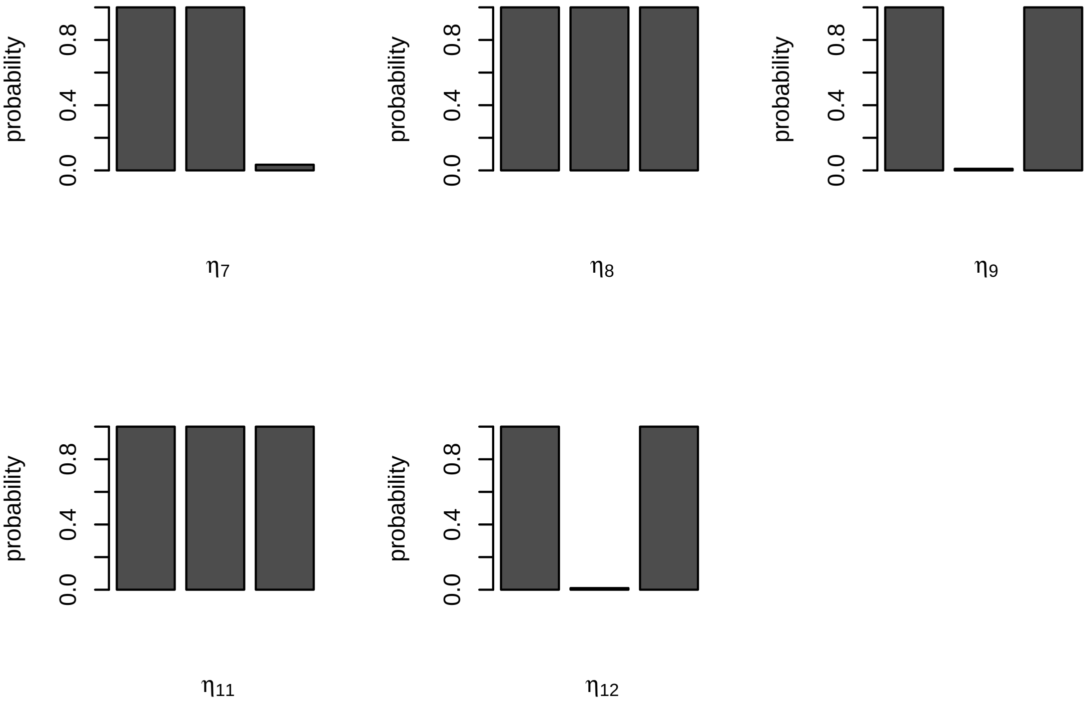


Figure 3.2: Example 3.1: The estimated posterior probabilities of $\eta_{j,m} : \hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ obtained by the two-layer Gibbs sampler

the particular regression model these variables in S_S are active for. The posterior means of the coefficients for the selected variables in both algorithms are shown in Table 3.2.

Table 3.2: Example 3.1: The estimated coefficients. (a) Group-wise Gibbs sampler (Algorithm 1).(b) Two-layer Gibbs sampler (Algorithm 2).

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$
X_7	1.55	1.56	0.13
X_8	1.37	1.84	2.38
X_9	1.49	0.02	2.12
X_{11}	3.10	2.57	4.15
X_{12}	3.21	-0.02	3.94

(a)

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$
X_7	1.57	1.55	0
X_8	1.37	1.84	2.40
X_9	1.49	0	2.16
X_{11}	3.09	2.57	4.16
X_{12}	3.18	0	3.97

(b)

3.2 Two-Layer Gibbs Sampler

Example 3.2 Performance of sample version of two-layer Gibbs sampler

In this example, we set $M = 5$ and there are $p = 200$ predictor variables of length $n = 80$. The variables are generated by Eq. 3.1 and $k = 2$ is chosen here, then the correlation between any two variables is 0.8. The true active variables are $\{X_7, X_8, X_9, X_{11}, X_{12}, X_{19}, X_{20}, X_{21}\}$, i.e. $S_S = \{7, 8, 9, 11, 12, 19, 20, 21\}$, and the corresponding coefficients of 5 individual regression model are shown in Table 3.3. The other coefficients are all set to be zero. Then each response vector is generated according to the linear model $Y_m = \boldsymbol{X}\beta_m + \omega_m$, where $\omega_w \sim N_{80}(\mathbf{0}, I_{80})$.

Table 3.3: Example 3.2: The true coefficients of the support union in the shared model

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$
X_7	0.9	1.7	0	1.2	1.5
X_8	0.9	1.7	2.2	1.2	0
X_9	0.9	1.7	0	0	0
X_{11}	0	2.5	0	0	1.3
X_{12}	3.2	0	4.1	2.3	0
X_{19}	0	0.6	0	0.4	0
X_{20}	0	0	0	0	0.7
X_{21}	1.5	0	0	0	0

To demonstrate the efficiency of the sample version of the two-layer Gibbs sampler, both the two-layer Gibbs sampler, Algorithm 2, and sample version of two-layer Gibbs sampler, Algorithm 3, are used in this example. In both methods, the prior parameters set-up are chosen as $\theta_j = 0.5$, $\rho_{j,m} = 0.5$, $\tau^2_{j,m} = 20$ for all $j \in \{1, \cdots, 50\}$, $m \in \{1, \cdots, 5\}$, and $a = b = 0.001$ as the non-informative parameter for inverse *Gamma* prior for σ^2. The last 200 draws are kept from the total 500 sweeps as the posterior samples, and the median probability criterion is also adopt for the posterior inference.

The estimated posterior probabilities of $\hat{P}(\delta_j = 1|\boldsymbol{Y})$ are shown in Figure 3.3. The estimated posterior probabilities of $X_7, X_8, X_9, X_{11}, X_{12}, X_{19}, X_{20}$ and X_{21} in both figures are all singinificantly higher than 0.5. Compare the probabilities of those inactive variables in the two figures, due to the Metropolis-Hasting acceptance-rejection rule, more inactive variables in sample version of two-layer Gibbs sampler have probabilities higher than 0.1. Then $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ for $j \in S_S$ are shown in Figure 3.4 and Figure 3.5. The results in two methods are very similar to each other.

The elapsed time is 90.13 seconds for the original two-layer Gibbs sampler, and 29.70 seconds for the sample version of two-layer Gibbs sampler. It means the sample version do speed up the calculation and save more than half of time to get the correct results. The posterior means of the coefficients for the selected variables are shown in Table 3.4.

Table 3.4: Example 3.2: The estimated coefficients. (a) Two-layer Gibbs sampler (b) Sample version of the two-layer Gibbs sampler

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$
X_7	1.04	1.58	0	1.04	1.53
X_8	0.74	1.71	2.26	1.16	0
X_9	0.82	1.73	0	0	0
X_{11}	0	2.61	0	0	1.32
X_{12}	3.14	0	4.07	2.36	0
X_{19}	0	0.68	0	0.56	0
X_{20}	0	0	0	0	0.64
X_{21}	1.69	0	0	0	0

(a)

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$
X_7	1.01	1.60	0	0.99	1.54
X_8	0.74	1.71	2.25	1.14	0
X_9	0.86	1.73	0	0	0
X_{11}	0	2.60	0	0	1.32
X_{12}	3.18	0	4.07	2.40	0
X_{19}	0	0.71	0	0.59	0
X_{20}	0	0	0	0	0.65
X_{21}	1.69	0	0	0	0

(b)

Example 3.3 In this example, three different values of residual variance, 1, 5, and 10, are used for generating three different sets of response vectors. By applying the proposed sample version of two-layer Gibbs sampler on the three sets of response

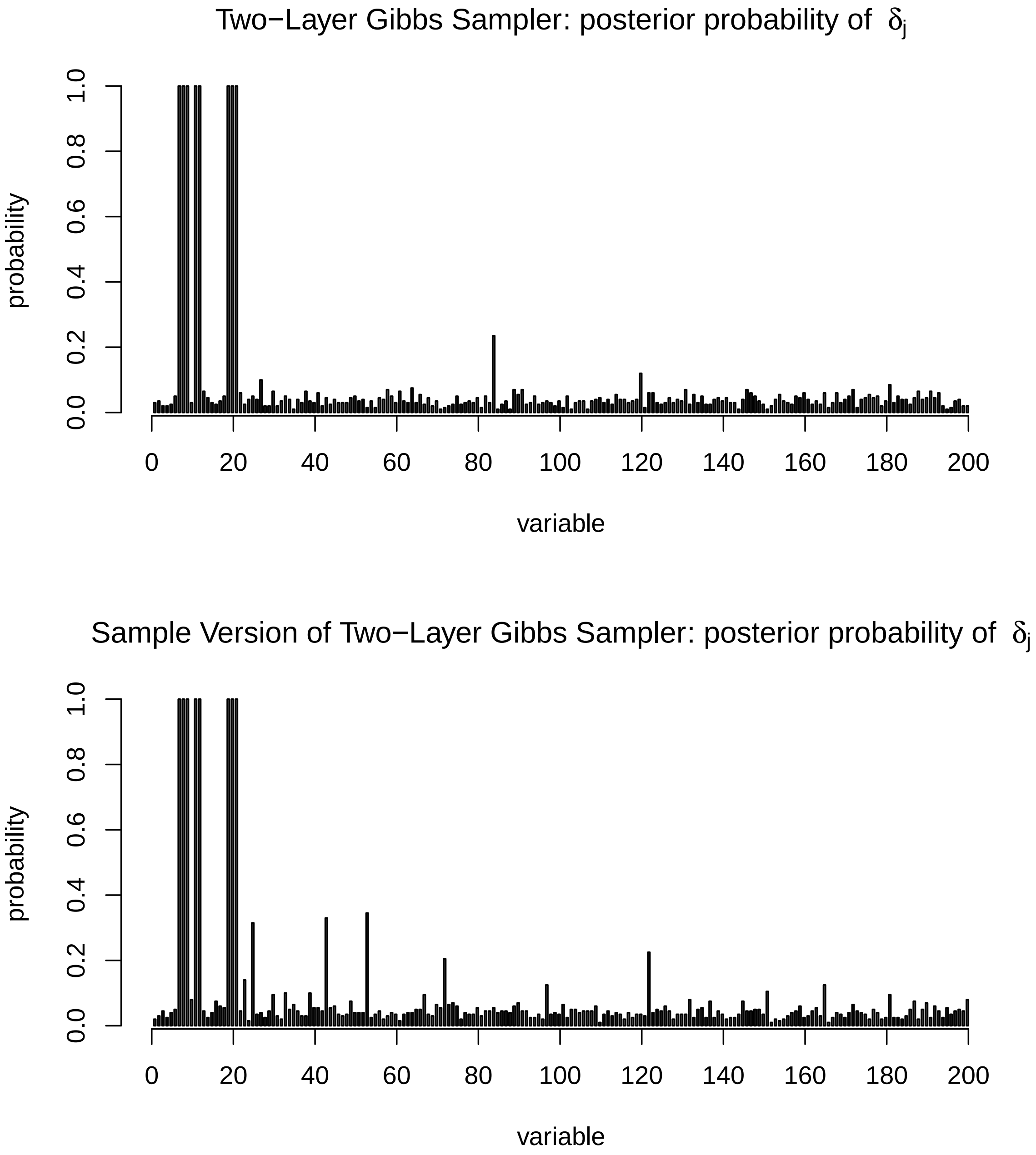


Figure 3.3: Example 3.2: The estimated posterior probabilities of δ_j : $\hat{P}(\delta_j = 1|\boldsymbol{Y})$. (a) two-layer Gibbs sampler (b) sample version of two-layer Gibbs sampler.

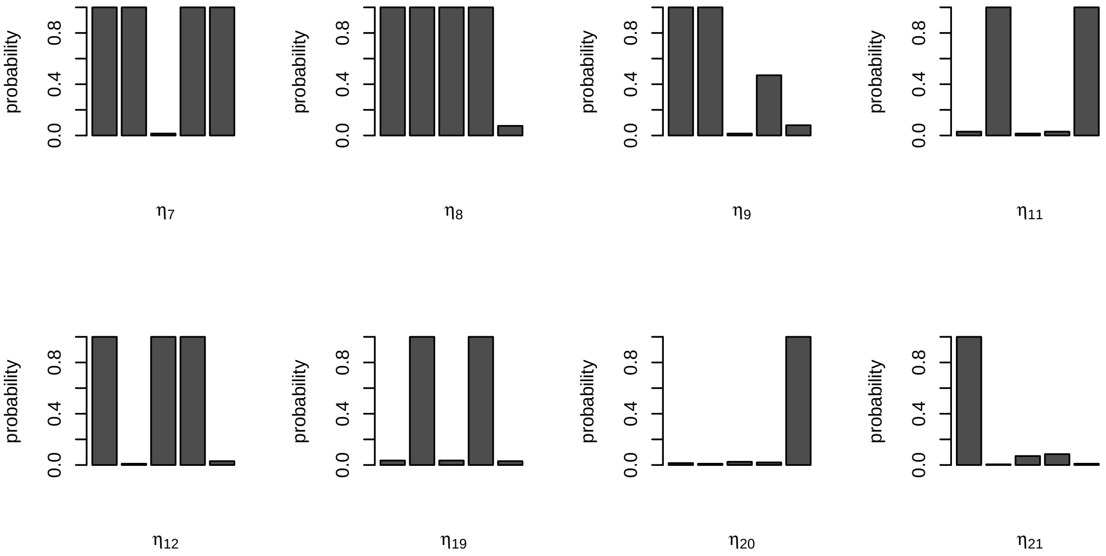


Figure 3.4: Example 3.2: The estimated posterior probabilities of $\eta_{j,m} : \hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ obtained by the two-layer Gibbs sampler

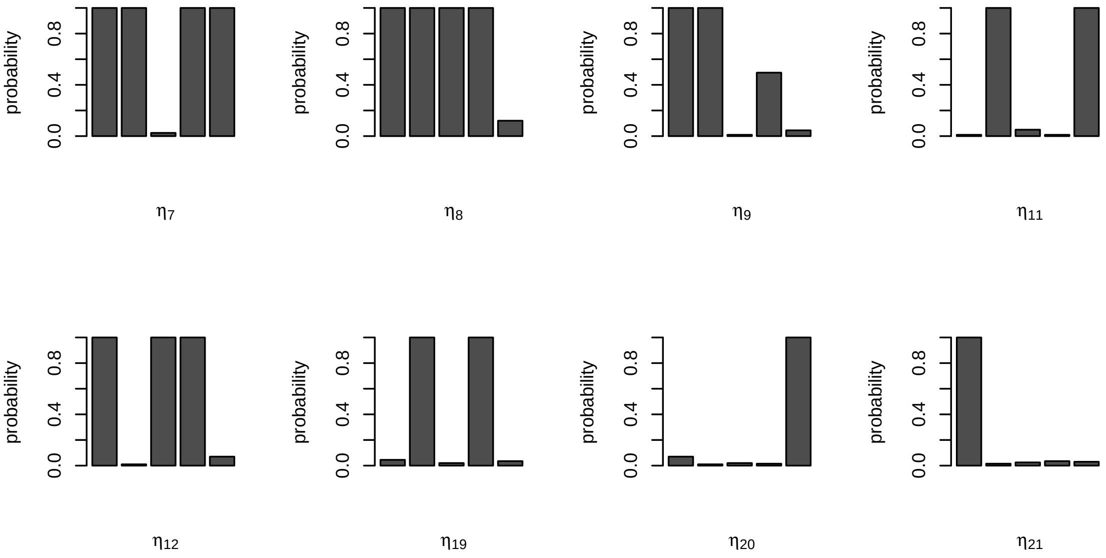


Figure 3.5: Example 3.2: The estimated posterior probabilities of $\eta_{j,m}$: $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ obtained by the sample version of two-layer Gibbs sampler

vectors, we want to know the influence of residual variance on the recovery results. Using the same date generating scheme and parameters setting in Example 3.2, the selection results with 50 replications are summarized in Table 3.5.

We measure the true positive rate (TPR), false positive rate (FPR) and accuracy for both first and second set of indicator variables, δ_j and $\eta_{j,m}$ respectively. True positive rate is the probability that the estimated set contains the true active variables. It is a measure of correct recovery. False positive rate is the rate of mis-containing the inactive variables. It measures the prediction errors. Accuracy is the rate of correct prediction, which is the sum of true positive and true negative over total variables. For TPR and accuracy, the higher the better, and for FPR, the lower the better. The values in Table 3.5 are the mean values over 50 replications.

Table 3.5: Example 3.3: The average rates of δ and η after 50 replications.

	δ			η		
	TPR	FPR	Accuracy	TPR	FPR	Accuracy
$\sigma^2 = 1$	0.9950	0.0055	0.9945	0.9884	0.0011	0.9987
$\sigma^2 = 5$	0.8475	0.0229	0.9719	0.8505	0.0042	0.9930
$\sigma^2 = 10$	0.7900	0.0436	0.9497	0.7495	0.0082	0.9872

From the results in Table 3.5, we can see large variance do affect the recovery results. Both TPR and accuracy decrease, and FPR increases when the variance increases, no matter for indicators in the first or second sets. However, the results may not be unacceptable. With the extreme high variance 10, the TPR for $\delta$is 0. 79 , it means it miss less than two active variables in average. The FPR for δ is 0.0436. it means about 7 from 192 inactive variables are incorrectly included in the model. As for the accuracy, both values for two sets of indicators are higher than

94%. Therefore, the proposed two-layer Gibbs sampler still did a good job in support union recovery.

Example 3.4 Consider another multi-response linear regression model. There are $M = 10$ regression models and $p = 400$ predicator variables of length $n = 100$. The variables are generated by Eq. 3.1 with $k = 1$. Thus the correlation between any two variables is 0.5. The true active variables are $\{X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}\}$, i.e. $S_S = \{7, 8, 9, 11, 12, 13, 14, 15\}$, and the corresponding coefficients of the 10 individual regression models are shown in Table 3.6. The other coefficients are all set to zero.

Table 3.6: Example 3.4: The true coefficients of the support union in the shared model.

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$	$\beta_{j,6}$	$\beta_{j,7}$	$\beta_{j,8}$	$\beta_{j,9}$	$\beta_{j,10}$
X_7	0.9	1.7	0	1.2	0.5	0	2.1	0.7	0	0.8
X_8	0.9	1.7	2.2	1.2	0	0.4	2.1	0.7	0	0.8
X_9	0.9	1.7	0	0	0.5	0.4	2.1	0	0.5	0.8
X_{11}	0	1.3	0	0.9	0	0	1.2	1.3	0	0
X_{12}	0	0	0	0.9	0	0.7	1.2	0	0.8	0
X_{13}	0	0	0	0	1.3	0	0	0	0	0
X_{14}	0	0	0	0	0	0	0	0	0.7	0
X_{15}	0	0.6	0	0	0	0.5	0	0	0.7	0

The sample version of two-layer Gibbs sampler, Algorithm 3, is applied in this example. The tuning parameter τ^2 is chosen from the pre-specified candidate set $\{1, 20, 40, 100\}$ by 5-fold cross validation. The other prior parameters set-up are chosen the same as those in Example 3.2, and the median probability criterion is also adopted for the posterior inference.

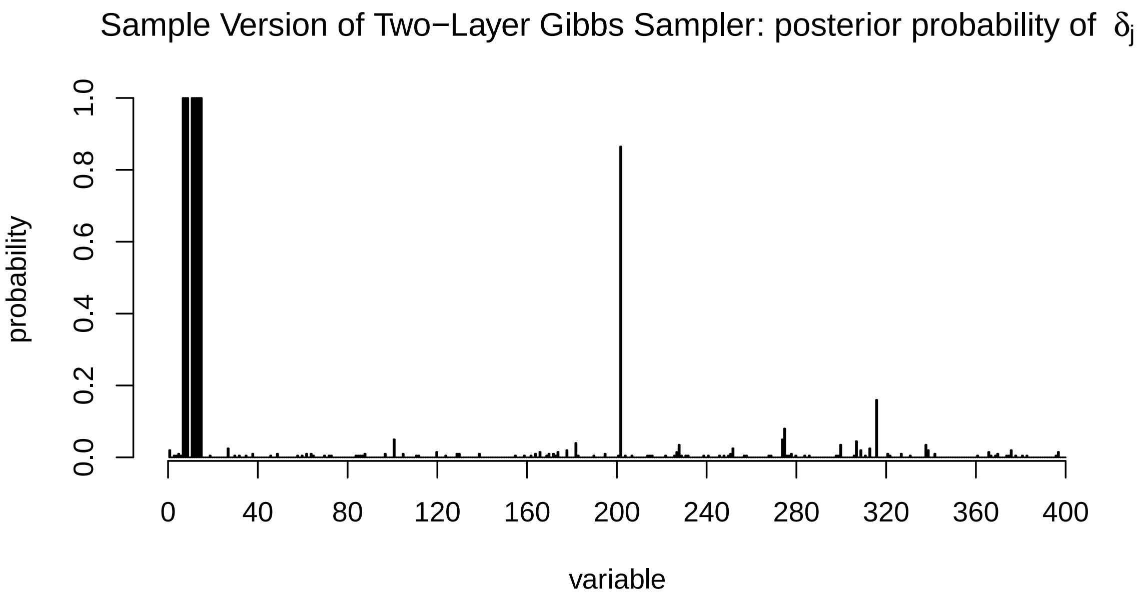


Figure 3.6: Example 3.4: The estimated posterior probabilities of δ_j : $\hat{P}(\delta_j = 1|\boldsymbol{Y})$.

By cross validation, τ^2 is chosen as 20. The estimated posterior probabilities of $\hat{P}(\delta_j = 1|\boldsymbol{Y})$ are show in Figure 3.6. Probabilities of all active predictor variables, and one additional variable X_{202}, are higher than 0.5. The estimated posterior probabilities of $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ for $j \in \{S_S, 202\}$, are shown in Figure 3.7. The posterior means of the coefficients for the selected variables and X_{202} are shown in Table 3.7.

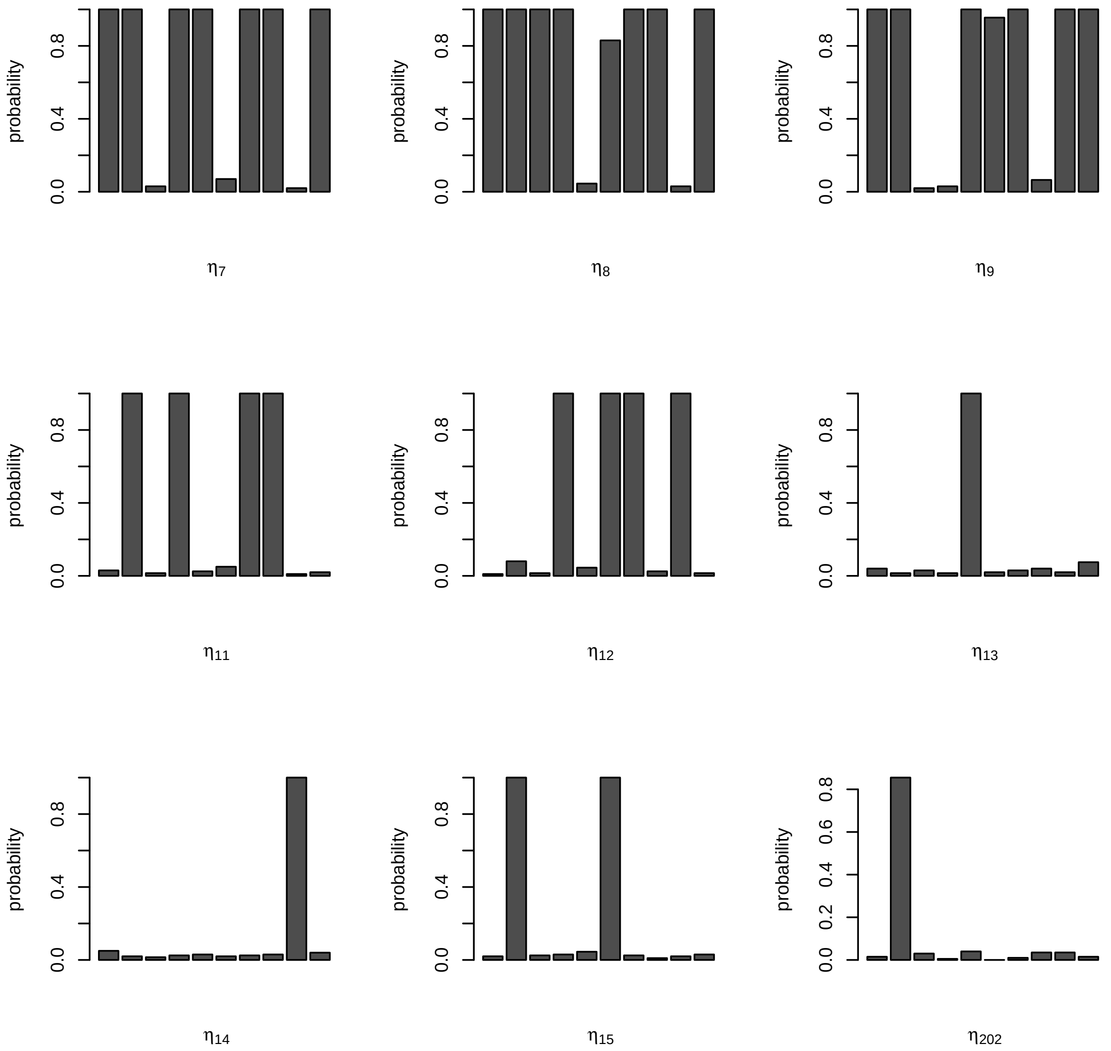


Figure 3.7: Example 3.4: The estimated posterior probabilities of $\eta_{j,m}$: $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$.

Table 3.7: Example 3.4: The estimated coefficients of the support union in the shared model.

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$	$\beta_{j,6}$	$\beta_{j,7}$	$\beta_{j,8}$	$\beta_{j,9}$	$\beta_{j,10}$
X_7	0.82	1.56	0	1.08	0.47	0	2.16	0.73	0	0.81
X_8	0.85	1.70	2.18	1.33	0	0.29	2.00	0.68	0	0.81
X_9	0.82	1.91	0	0	0.40	0.34	2.22	0	0.52	0.85
X_{11}	0	1.40	0	0.88	0	0	1.23	1.23	0	0
X_{12}	0	0	0	1.00	0	0.63	1.15	0	0.79	0
X_{13}	0	0	0	0	1.23	0	0	0	0	0
X_{14}	0	0	0	0	0	0	0	0	0.67	0
X_{15}	0	0.69	0	0	0	0.62	0	0	0.7	0
X_{202}	0	-0.33	0	0	0	0.62	0	0	0.7	0

3.3 Comparison between Gibbs and Lasso on Different Response setting

In this section, we compare the performance between Gibbs and Lasso on different response setting. First, we can use the two sets of indicators, $\{\delta_j, j = 1, \cdots, p\}$ and $\{\eta_{j,m}, j = 1, \cdots, p, m = 1, \cdots, M\}$, to describe the linear regression for each response vector, Y_m, in the three different settings separately. They are shown as:

$$\text{Component-Wise} \qquad Y_m = \eta_{1,m}(\beta_{1,m}X_1) + \cdots + \eta_{p,m}(\beta_{p,m}X_p) + \omega_m \tag{3.2}$$

$$\text{Group-Wise} \qquad Y_m = \delta_1(\beta_{1,m}X_1) + \cdots + \delta_p(\beta_{p,m}X_p) + \omega_m \tag{3.3}$$

$$\text{Two-layer} \qquad Y_m = \delta_1\eta_{1,m}(\beta_{1,m}X_1) + \cdots + \delta_p\eta_{p,m}(\beta_{p,m}X_p) + \omega_m \tag{3.4}$$

In Eq. (3.2), each component, $\beta_{j,m}X_j$, is multiplied by the indicator in the second set, $\eta_{j,m}$. That is, in the component-wise setting, whether the variable X_j is active for the m-th response vector, Y_m, depends individually on the specific indicator $\eta_{j,m}$. Instead, in Eq. (3.3), the component $\beta_{j,m}X_j$ in the group-wise setting is multiplied by the indicator in the first set, δ_j. That is whether the variable X_j is active for all response vectors, $Y_m, m = 1, \cdots, M$, depends on the same indicator δ_j simultaneously. Consider the proposed two-layer setting. By Combining two sets of indicators together, each component in Eq. (3.4) is multiplied by $\delta_j\eta_{j,m}$. Thus whether the variable X_j is active in the m-th response vector, Y_m, is decided by both δ_j and $\eta_{j,m}$. Only when $\delta_j = 1$ and $\eta_{j,m} = 1$, the variable X_j is active for the response vector Y_m.

Example 3.5. Comparision

In this example, we extend the dimension of the response vectors to $M = 15$. There are $p = 200$ predictor variables of length $n = 80$. As defined before, the variables are generated by $X_j = G_j + kG$, where $k = 2$ is a pre-specified constant. In the setting, the correlation between any two variables is 0.8. The true active variable set is $\{X_7, X_8, X_9, X_{11}, X_{12}, X_{13}\}$, i.e. $S_S = \{7, 8, 9, 11, 12, 13\}$, and the corresponding coefficients of 15 single regression models are shown in Table 3.8. The cell with gray color means the variable is not active for this singular regression model. The other coefficients are all set to be zero. Then each response vector is generated according to the linear model $Y_m = \boldsymbol{X}\beta_m + \omega_m$, where $\omega_m \sim N_{80}(\mathbf{0}, I_{80})$.

Table 3.8: Example 3.5: The true coefficients of the support union in the shared model.

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$	$\beta_{j,6}$	$\beta_{j,7}$	$\beta_{j,8}$	$\beta_{j,9}$	$\beta_{j,10}$	$\beta_{j,11}$	$\beta_{j,12}$	$\beta_{j,13}$	$\beta_{j,14}$	$\beta_{j,15}$
X_7	0.9	1.7	0	1.2	0.5	0	2.1	0.7	0	0.8	0.8	2.5	0	0	0.9
X_8	0.9	1.7	2.2	1.2	0	0.4	2.1	0.7	0	0.8	0.8	2.5	1.3	0	0
X_9	0.9	1.7	0	0	0.5	0.4	2.1	0	0.5	0.8	0.8	2.5	0	0.5	0
X_{11}	0	0	0	0	1.3	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0
X_{13}	0	0.6	0	0	0	0.5	0	0	0	0	0	0	0	0	0

In sample version of two-layer Gibbs sampler, i.e. Algorithm 3, two sets of indicators, δ_j and $\eta_{j,m}$, are adopted, and the prior parameters are set as $\theta_j = 0.5$, $\rho_{j,m} = 0.5$ for all $j \in \{1, \cdots, 200\}$, $m \in \{1, \cdots, 15\}$. In Group-wise Gibbs sampler, i.e. Algorithm 1, only the first set of indicators, δ_j, is adopted, and the parameter θ_j, $j = 1, \cdots, p$ are all set to be 0.5. As for the component-wise Gibbs sampler, we treat each response as a special case of multi-response linear regression with $M = 1$, and apply Algorithm 3 on each response vector separately. Then the union of the selected predictors from each singular regression model are chosen as the support union in the shared model. The other prior parameters for the three algorithms are set as $\tau^2_{j,m} = 20$ for all $j \in \{1, \cdots, 200\}$, $m \in \{1, \cdots, 15\}$, and $a = b = 0.001$ as the non-informative parameter for inverse gamma prior of σ^2. The initial model is set as the null model, i.e. $\delta_j = 0$; $\eta_{j,m} = 0$ and $\beta_{j,m} = 0$ for all j and m, and the median probability criterion is adopt for the posterior inference. Totally we run 500 sweeps. After discarding the first 300 sweeps, samples collected from the last 200 sweeps are used for the inference.

The chosen sets of the support union in the shared model, S_S, by the three different algorithms are shown in Table 3.12. The selection results of the two-layer Gibbs sampler agree the true model, but the selection results of Group-wise Gibbs sampler miss identify three active predictor variables, X_{11}, X_{12}, and X_{13}, which are active for just one or two singular regression models. Thus due to the weak group signal for the variables X_{11}, X_{12}, and X_{13}, this group-wise Gibbs sampler has the under-selection problem. As for the component-wise Gibbs sampler, the selection result reveal the problem of over-selection. 29 predictor variables are chosen as active. Focus on the true active variable set, S_S. The posterior means of the coefficients are shown in Table 3.9, Table 3.10, and Table 3.11.

Table 3.9: Example 3.5: The estimated coefficients by the sample version of the two-layer Gibbs sampler

Var	$\hat{\beta}_{j,1}$	$\hat{\beta}_{j,2}$	$\hat{\beta}_{j,3}$	$\hat{\beta}_{j,4}$	$\hat{\beta}_{j,5}$	$\hat{\beta}_{j,6}$	$\hat{\beta}_{j,7}$	$\hat{\beta}_{j,8}$	$\hat{\beta}_{j,9}$	$\hat{\beta}_{j,10}$	$\hat{\beta}_{j,11}$	$\hat{\beta}_{j,12}$	$\hat{\beta}_{j,13}$	$\hat{\beta}_{j,14}$	$\hat{\beta}_{1,15}$
X_7	0.90	1.46	0	1.15	0.57	0	2.07	0.60	0	0.83	0.82	2.58	0	0	0.85
X_8	0.91	1.86	2.24	1.18	0	0	2.03	0.78	0	0.67	0.84	2.35	1.22	0	0
X_9	0.90	1.68	0	0	0.45	0.36	2.20	0	0.51	0.78	0.77	2.47	0	0.54	0
X_{11}	0	0	0	0	1.23	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0.68	0	0	0	0	0	0
X_{13}	0	0.73	0	0	0	0.69	0	0	0	0	0	0	0	0	0

Table 3.10: Example 3.5: The estimated coefficients by the Group-wise Gibbs sampler

Var	$\hat{\beta}_{j,1}$	$\hat{\beta}_{j,2}$	$\hat{\beta}_{j,3}$	$\hat{\beta}_{j,4}$	$\hat{\beta}_{j,5}$	$\hat{\beta}_{j,6}$	$\hat{\beta}_{j,7}$	$\hat{\beta}_{j,8}$	$\hat{\beta}_{j,9}$	$\hat{\beta}_{j,10}$	$\hat{\beta}_{j,11}$	$\hat{\beta}_{j,12}$	$\hat{\beta}_{j,13}$	$\hat{\beta}_{j,14}$	$\hat{\beta}_{j,15}$
X_7	0.84	1.71	0.15	1.19	0.93	0.39	2.07	0.54	0.17	0.87	0.82	2.58	0.07	0.01	1.01
X_8	0.87	2.15	2.19	1.10	0.51	0.29	2.07	0.72	0.24	0.67	0.86	2.34	1.14	0.01	-0.14
X_9	0.89	1.84	-0.09	0.05	0.71	0.51	2.17	0.11	0.74	0.73	0.75	2.49	0.03	0.52	-0.07
X_{11}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{13}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 3.11: Example 3.5: The estimated coefficients by the Component-wise Gibbs sampler

Var	$\hat{\beta}_{j,1}$	$\hat{\beta}_{j,2}$	$\hat{\beta}_{j,3}$	$\hat{\beta}_{j,4}$	$\hat{\beta}_{j,5}$	$\hat{\beta}_{j,6}$	$\hat{\beta}_{j,7}$	$\hat{\beta}_{j,8}$	$\hat{\beta}_{j,9}$	$\hat{\beta}_{j,10}$	$\hat{\beta}_{j,11}$	$\hat{\beta}_{j,12}$	$\hat{\beta}_{j,13}$	$\hat{\beta}_{j,14}$	$\hat{\beta}_{j,15}$
X_7	0.90	1.63	0	1.14	0.46	0	2.03	0.53	0	1.02	0.78	2.58	0	0	0.95
X_8	0.93	1.91	2.28	1.09	0	0	2.06	0.70	0	0.95	0.75	2.23	1.19	0	0
X_9	0.93	1.68	0	0	0.37	0	2.16	0	0.42	0.86	0.67	2.42	0	0.55	0
X_{11}	0	0	0	0	1.11	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0.54	0	0	0	0	0	0
X_{13}	0	0.96	0	0	0	0.56	0	0	0	0	0	0	0	0	0

Table 3.12: Example 3.5: The selection results of support union in the shared model.

Algorithm	S_S
Two-layer	7 8 9 11 12 13
Group-wise	7 8 9
Component-wise	7 8 9 11 12 13 5 10 16 40 45 47 49 50 51 54 60 69 83 91 94 106 130 134 155 179 183 184 199
Sparse group Lasso	7 8 9 13 10 27 46 55 72 105 149 154 156 184
Group Lasso	7 8 9 13 10 27 46 55 62 105 131 154 184

In addition to Bayesian approaches, we also compare the simulation results with the Lasso type methods. Here the group Lasso function, *mcLeastR.m*, and sparse group Lasso function, $mc - sgLeastR.m$, in SLEP MATLAB toolbox are used for this simulation study. According to the response vector setting in the multi-response linear regression described in Eq. 3.3 and Eq. 3.4, the group Lasso corresponds to the group-wise Gibbs sampler, and sparse group Lasso corresponds to the two-layer setting.

Consider the selection results of support union in the shared model, S_S, shown in the last two rows for sparse group Lasso and group Lasso in Table 3.12. The results of both Lasso methods are similar. 4 out of 6 true active variables, X_7, X_8, X_9, X_{13}, are detected as active in both methods. Additional 10 and 9 variables, i.e. numbers in the gray color, are false detected as active by the two Lasso methods respectively. Therefore, both Lasso methods have over-selection problem for the support union in the shared model. Focus on the true active variable set, S_S. The coefficient estimations, $\beta_{j,m}$ for $j \in S_S$ and $m = 1, \cdots, 15$, are shown in Table 3.13 and Table 3.14.

Table 3.13: Example 3.5: The estimated coefficients by the sparse group Lasso

Var	$\hat{\beta}_{j,1}$	$\hat{\beta}_{j,2}$	$\hat{\beta}_{j,3}$	$\hat{\beta}_{j,4}$	$\hat{\beta}_{j,5}$	$\hat{\beta}_{j,6}$	$\hat{\beta}_{j,7}$	$\hat{\beta}_{j,8}$	$\hat{\beta}_{j,9}$	$\hat{\beta}_{j,10}$	$\hat{\beta}_{j,11}$	$\hat{\beta}_{j,12}$	$\hat{\beta}_{j,13}$	$\hat{\beta}_{j,14}$	$\hat{\beta}_{j,15}$
X_7	0.47	1.12	0	1.165	0.20	0	1.52	0.13	0	0.43	0.37	1.91	0	0	0.08
X_8	0.09	1.04	1.05	0.25	0	0	1.03	0	0	0	0.07	1.22	0.24	0	0
X_9	0.87	1.71	0	0.06	0.43	0.22	2.08	0.09	0.40	0.73	0.67	2.33	0.02	0	0
X_{11}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{13}	0	0.29	0	0.03	0	0.08	0.03	0	0	0	0	0.03	0	0	0

Table 3.14: Example 3.5: The estimated coefficients by the group Lasso

Var	$\hat{\beta}_{j,1}$	$\hat{\beta}_{j,2}$	$\hat{\beta}_{j,3}$	$\hat{\beta}_{j,4}$	$\hat{\beta}_{j,5}$	$\hat{\beta}_{j,6}$	$\hat{\beta}_{j,7}$	$\hat{\beta}_{j,8}$	$\hat{\beta}_{j,9}$	$\hat{\beta}_{j,10}$	$\hat{\beta}_{j,11}$	$\hat{\beta}_{j,12}$	$\hat{\beta}_{j,13}$	$\hat{\beta}_{j,14}$	$\hat{\beta}_{j,15}$
X_7	0.70	1.45	0.46	0.75	0.63	0.29	1.67	0.40	0.22	0.63	0.65	2.01	0.24	0.08	0.44
X_8	0.59	1.37	0.88	0.63	0.43	0.23	1.43	0.39	0.21	0.51	0.58	1.66	0.47	0.08	0.08
X_9	0.81	1.73	0.36	0.42	0.63	0.40	1.98	0.27	0.49	0.70	0.72	2.28	0.22	0.31	0.11
X_{11}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{12}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
X_{13}	0.02	0.05	0.02	0.02	0.02	0.02	0.04	0.01	0.01	0.02	0.02	0.05	0.01	0	0.01

Based on the same tuning parameter set-up and the number of iterations, the selection results with 100 replications are summarized in Table 3.15. We measure the true positive rate (TPR), false positive rate (FPR) and accuracy for both first and second set of indicator variables, δ_j and $\eta_{j,m}$ respectively. For TPR and accuracy, the higher the better, and for FPR, the lower the better. The values in Table 3.15 are the mean values over 100 replications.

Consider the first set of indicator variables, δ_j. Although the component-wise Gibbs sampler produces the perfect identified rates for the true active variables, i.e. TPR $= 1$, it has the highest FPR in three Gibbs sampler algorithms. It means when learning related tasks independently, it is easy to have over-selection problem. However, the problem can be solved by multi-task learning through Bayesian approaches. Both group-wise and two-layer Gibbs sampler have very low FPR, 0.0005, and 0.0006 respectively. It means in the two Bayesian algorithms, about 10 variables are false selected as active in total 100 replications. In contrast, the over-selection problem still exists in two Lasso algorithms. Both group Lasso and sparse group Lasso have high FPR, 0.0373, 0.0424 respectively. In other words, on average 8 variables are false selected as active in each replication. The advantage of two-layer setting is in TPR, where the value rise significant from 0.5282 in group-wise Gibbs sampler to 0.9833 in two-layer Gibbs sampler, and from 0.5833 in group Lasso to 0.7783 in sparse group Lasso. The same as accuracy, the value of accuracy goes up when the second set of indicator variables, $\eta_{j,m}$, are added in the multi-task regression model. As for the second set of indicators, $\eta_{j,m}$. Two-layer Gibbs sampler produces over 99% TPR and accuracy, and has the lowest FPR: 0.0002. It means on average 0.5 indicator from total 2965 inactive indicators of the second set are false selected as active. Thus, by adopting the two sets of indicator variables together in the multi-task learning, the proposed two-layer Gibbs sampler cannot only dig out the shared variables but also indicate which particular response vector the variable is active for.

Table 3.15: Example 3.5: The average rates of δ and η after 100 replications.

	δ			η		
	TPR	FPR	Accuracy	TPR	FPR	Accuracy
Component-wise	1	0.0414	0.9223	0.9669	0.0056	0.9944
Group-wise	0.5283	0.0005	0.9853			
Two-layer	0.9833	0.0006	0.9989	0.9909	0.0002	0.9997
Group Lasso	0.5833	0.0373	0.9496			
Sparse group Lasso	0.7783	0.0424	0.9583	0.8660	0.0249	0.9753

CHAPTER 4

Application in Image Studies

In this section, the proposed Bayesian algorithm is applied in image analysis. Based on the sparse coding theory of Olshausen and Field (1996) [OF96], an image $\boldsymbol{I}$ can be represented as a linear composition of Gabor wavelet elements

$$\boldsymbol{I} = \sum_{j=1}^{p} c_j G_j + U, \tag{4.1}$$

where $(G_j, j = 1, \ldots, p)$ is a dictionary of Gabor basis functions defined on the same domain as $\boldsymbol{I}$, $\{c_j, j = 1, \ldots p\}$ are the coefficients, and U is the unexplained residual image. In this situation, the basis functions are treated as representational features and assumed over-complete.

Example 4.1 Given a domain $\chi = \{(x_1, x_2) | \, x_1 \in \{1, 2, \cdots, 10\}, \; x_2 \in \{1, 2, \cdots, 10\}\}$, the image can be represented as $(10 \times 10) \times 1$ image vector. We define the Gabor basis dictionary as

$$\begin{aligned} G(u, v) &= \exp\left[-\frac{1}{2}(\frac{u^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2})\right] \cos\left[\frac{2\pi u}{\lambda}\right], && (4.2) \\ u &= u_0 + x_1 \cos\theta + x_2 \sin\theta, && (4.3) \\ v &= v_0 - x_1 \sin\theta + x_2 \cos\theta, && (4.4) \end{aligned}$$

where $(u_0, v_0) \in \chi$ has the same domain as image, $\sigma_u = 1, \sigma_v = \sqrt{2}, \lambda = \sqrt{2\pi}$ and $\theta = \{0, \pi/3, 2\pi/3\}$ is the angle between the x_1-axis of the image and the u-axis of the Gabor functions. Thus, we have 300 Gabor basis functions in total whose norms are all equal to 1 on χ. For simplicity, we use $X_1, \cdots, X_{300}$ to index all the basis functions, and X_i is a 100×1 vector, $i = 1, 2, \cdots, 300$. The true active variables in the shared model are $\{X_{17}, X_{71}, X_{161}, X_{180}\}$, i.e. $S_S = \{17, 71, 161, 180\}$, and the corresponding coefficients of 5 single regression models are shown in Table 4.1. Then each response vector Y_m is generated according to the linear model $Y_m = \boldsymbol{X}\beta_m + \omega_m$ where $\omega_m \sim N_{100}(\mathbf{0}, I_{100})$.

Table 4.1: Example 4.1: The true coefficients of the support union in the shared model.

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$
X_{17}	7	-9	8	-7	-8
X_{71}	-7	9	0	0	8
X_{161}	7	9	-8	7	-8
X_{180}	-7	-9	0	0	8

The sample version of two-layer Gibbs sampler, i.e. Algorithm 3, is applied in the example, and the prior parameters are set as $\theta_j = 0.5, \rho_{j,m} = 0.5, \tau_{j,m} = 40$ for all $j \in \{1, \cdots, 300\}$, $m \in \{1, \cdots, 5\}$, and $a = b = 0.001$ as the non-informative parameter for inverse gamma prior of σ^2. Totally we run 1000 sweeps. After discarding the first 500 sweeps, examples collected from the last 500 sweeps are used for the inference about support union recovery. The estimated posterior probabilities of indicators in the first set, $\hat{P}(\delta_j = 1|\boldsymbol{Y}),\ j = 1, \cdots, p$, are shown in Figure 4.1. It is clear that the posterior probabilities of X_{17}, X_{71}, X_{161}, and X_{180} are all higher than 0.5. Therefore, based on the median probability criterion, the selection result agrees with the true model. The $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ for $j \in S_s$ are shown in Figure 4.2. For those nonzero

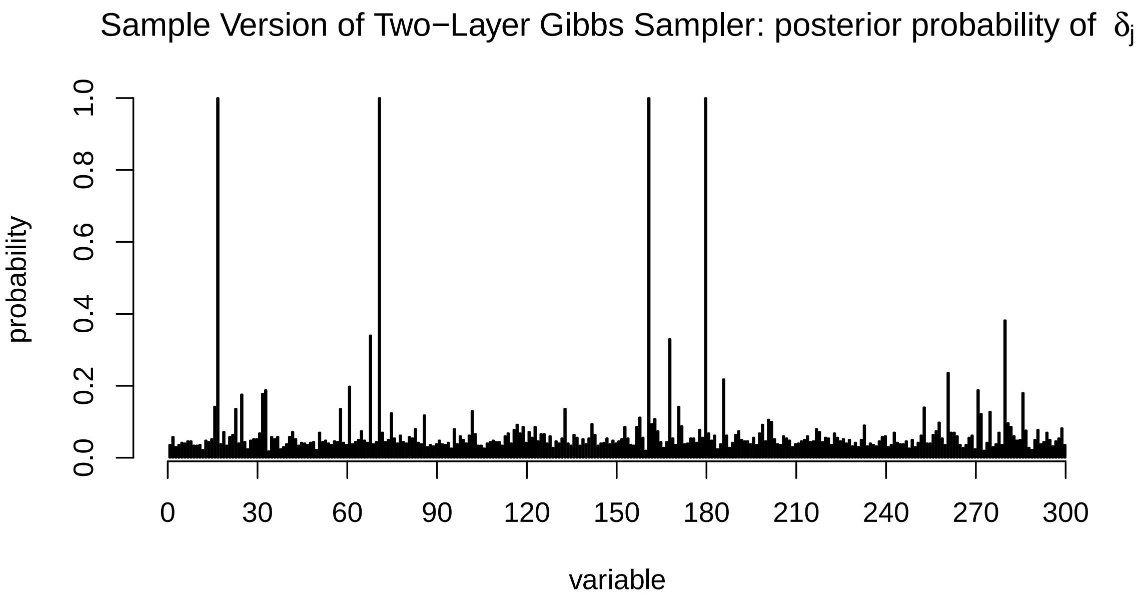


Figure 4.1: Example 4.1: The estimated posterior probabilities of δ_j : $\hat{P}(\delta_j = 1|\boldsymbol{Y})$.

coefficients, the corresponding indicators in the second set have posterior probabilities larger than 0.5. Therefore, these are treated as active. The corresponding means of the coefficients for the selected variable are also shown in Table 4.2.

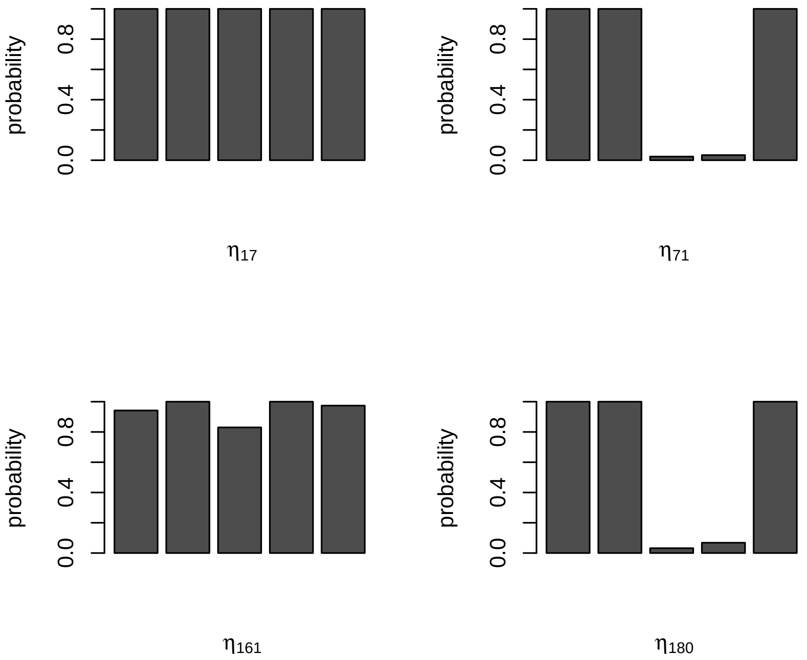


Figure 4.2: Example 4.1: The estimated posterior probabilities of $\eta_{j,m}$: $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})$ obtained by the sample version of the two-layer Gibbs sampler.

Table 4.2: Example 4.1: The estimated coefficients.

X_j	$\beta_{j,1}$	$\beta_{j,2}$	$\beta_{j,3}$	$\beta_{j,4}$	$\beta_{j,5}$
X_{17}	6.14	-7.10	8.75	-7.74	-7.74
X_{71}	-7.04	9.29	0	0	8.60
X_{161}	6.27	8.53	-6.01	7.41	-8.93
X_{180}	-9.26	-10.09	0	0	8.52

4.1 Real Images

This section presents the performance of the proposed two-year Gibbs sampler on the real images. To fix notation, a *Gabor* basis defined in 4.2 is of the form $G(u,v) = \exp\{-[(u^2/\sigma_x^2) + (v^2/\sigma_y^2)]/2\}\cos(2\pi u/\lambda)$, where $\sigma_u < \sigma_v$. We can translate, rotate, and dilate $G(u,v)$ to obtain a general form of *Gabor* basis: $B_{x,y,s,\alpha}(u',v') = g(\tilde{u}/s, \tilde{v}/s)/s^2$ where $\tilde{u} = (u'-x)\cos\alpha + (u'-y)\sin\alpha$, $\tilde{y} = -(u'-x)\sin\alpha + (v'-y)\cos\alpha$. (x,y) is the central position, s is the scale parameter, and α is the orientation. Let D be the domain of image lattice. The dictionary of *Gabor* basis is Dictionary $= \{B_{x,y,s,\alpha}, \forall(x,y,s,\alpha)\}$, where $(x,y) \in D$, and $\alpha \in \{a\pi/A, a = 0, \cdots, A-1\}$ (e.g., A $= 5$).

In real image learning, we assume that the images are defined on the same image lattice which is the bounding box of the objects in these images. The Gabor basis are generated on the same domain of the image lattice. The following are the parameter values we use in all image examples in this paper (unless otherwise stated). Length of Gabor wavelets = 7. The orientation α takes $A = 5$ equally spaced angles in $[0, \pi]$. $\theta_j = 0.5$, $\rho_{j,m} = 0.5$ and $\tau_{j,m} = 20$ for all $j \in \{1, \cdots, p\}$. $m \in \{1, \cdots, M\}$. σ^2 is given as 0.001.

Example 4.2. In Example 4.2.1, we apply the sample version of the two-layer Gibbs sampler to a set of $M = 5$ cup images. The cup images are resized to 50×50 (height × length), so each cup image is represented as $(50 \times 50) \times 1$ image vector. Totally $50 \times 50 \times 5 = 12500$ Gabor basis functions are chosen in this example. There are 4852 out of 12500 Gabor basis are chosen as active in the shared model. Figure 4.3 displays the results. The image on the top displays the shared model learned by the sample version of two-layer Gibbs sampler, where active Gabor basis are multiplied

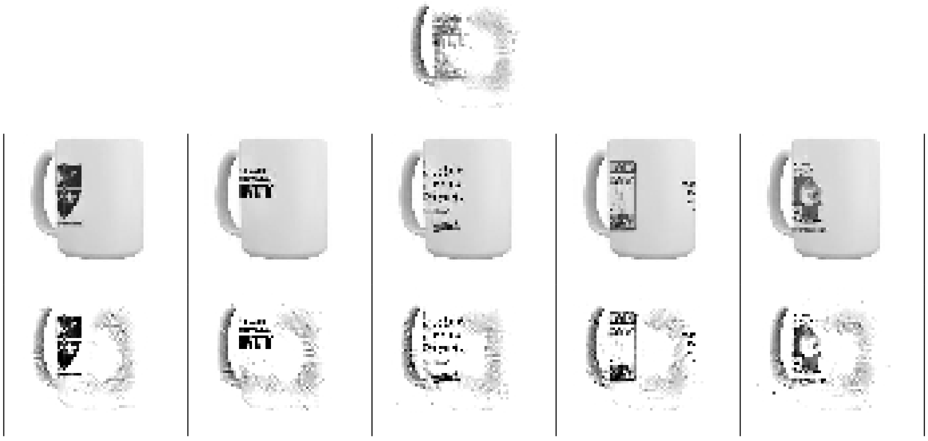


Figure 4.3: Example 4.2.1. The 5 cup images are 50 × 50 (height × length). The top image shows the image recovered by the shared model. Each block displays the observed image and the corresponding recovered images by sample version of the two-layer Gibbs sampler. Samples collected from the last 500 sweeps are used for inference after discarding the first 500 sweeps.

by the average of the estimated coefficients. For the remaining 5 pairs of plots, the top plot shows the original image I_m, and the bottom plot shows the recovered image by the individual support S_m. Figures 4.4 - 4.7 display more examples, where the results are obtained by the same algorithm.

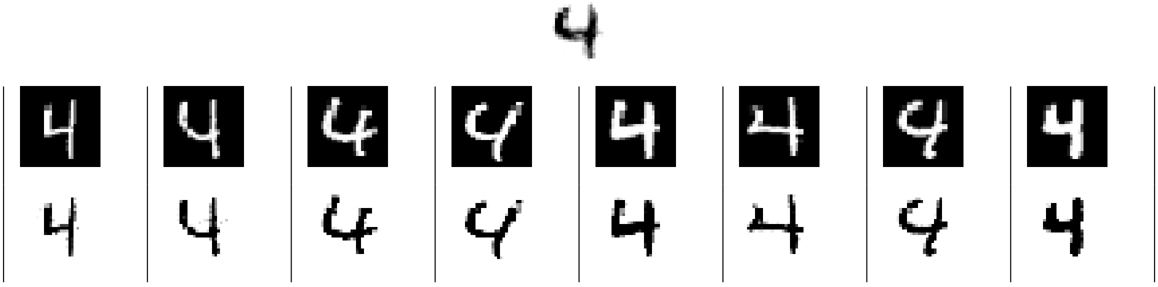


Figure 4.4: Example 4.2.2. The 8 handwritten digits of number 4 images are 30 × 30. Samples collected from the last 200 sweeps are used for inference after discarding the first 300 sweeps.

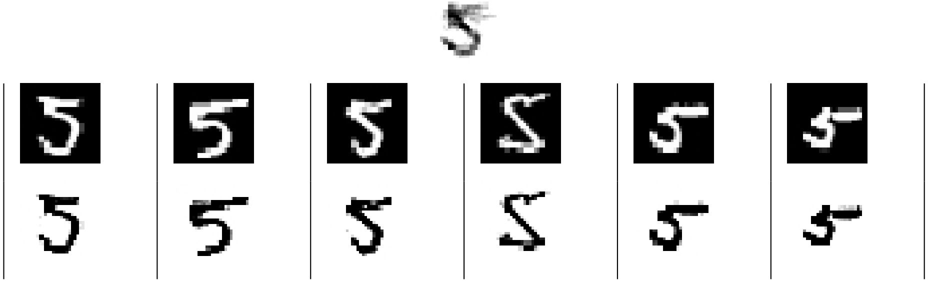


Figure 4.5: Example 4.2.3. The 6 handwritten digits of number 5 images are 30 × 30. Samples collected from the last 200 sweeps are used for inference after discarding the first 300 sweeps.

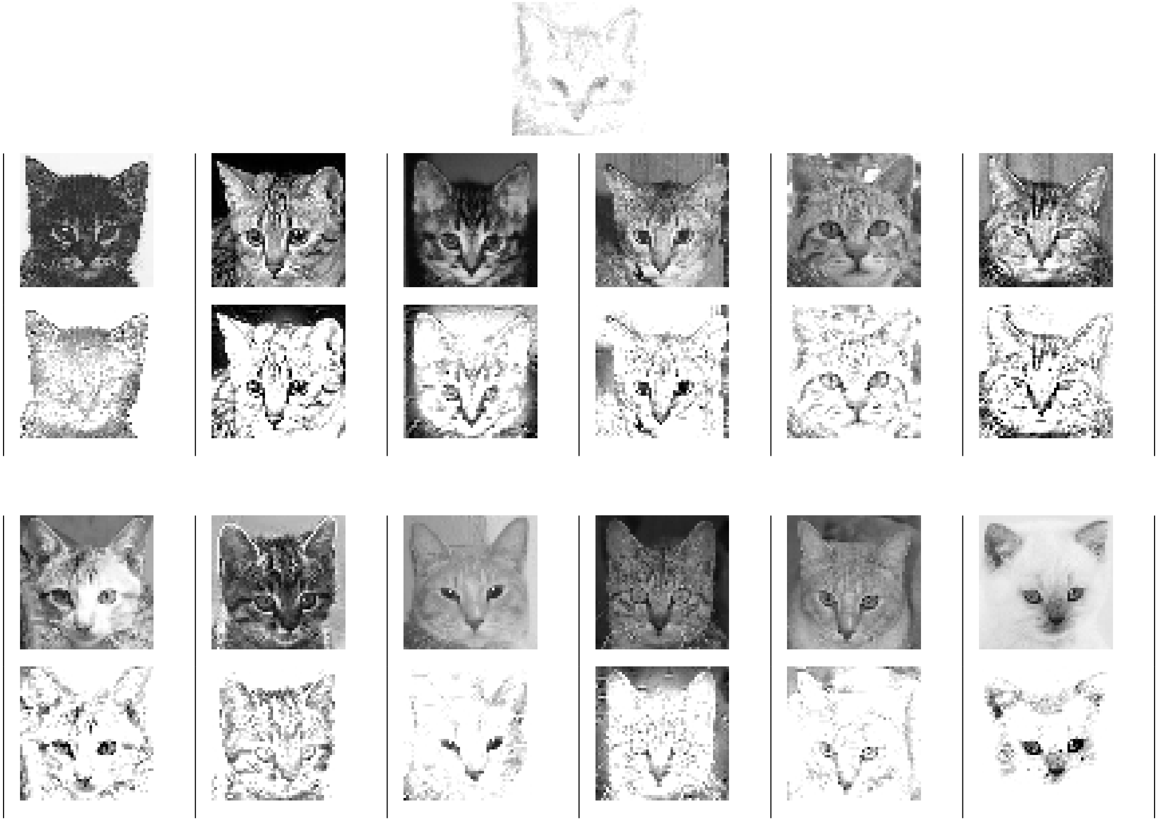


Figure 4.6: Example 4.2.4. The 12 cat images are 50 × 50. Samples collected from the last 500 sweeps are used for inference after discarding the first 500 sweeps.

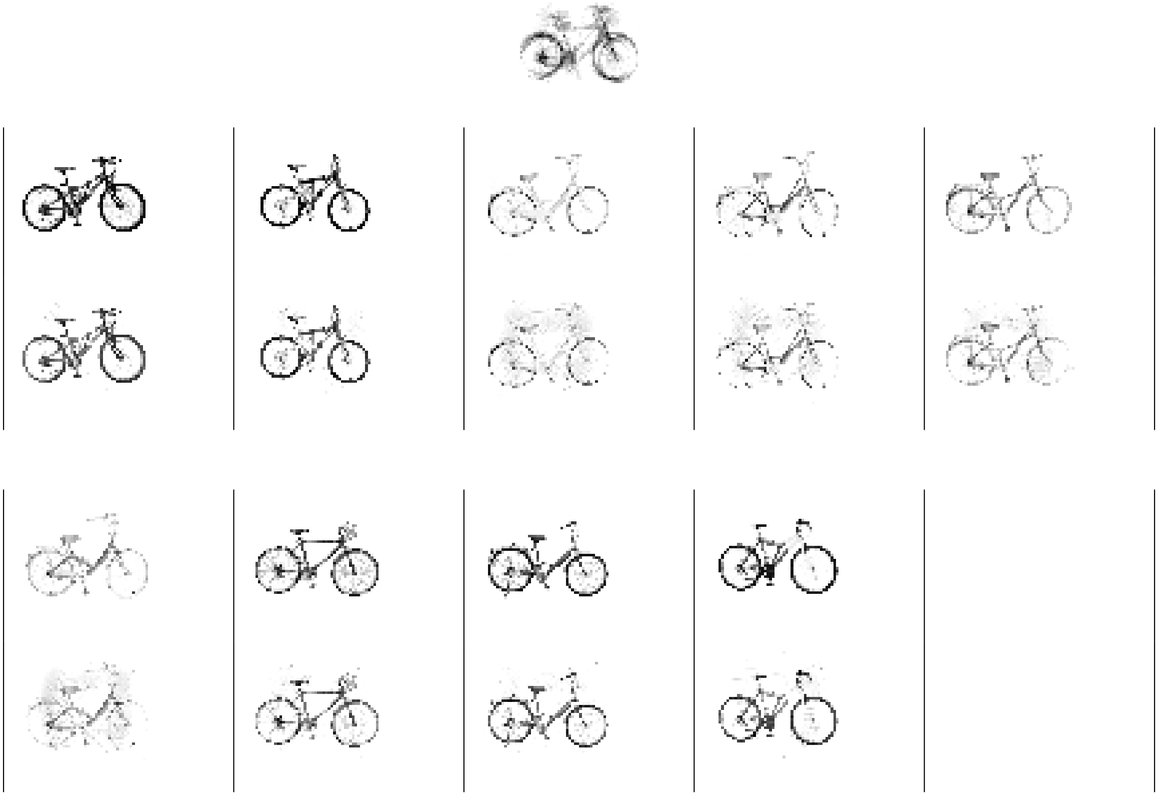


Figure 4.7: Example 4.2.5. The 9 bicycle images are 50 × 50. Samples collected from the last 500 sweeps are used for inference after discarding the first 500 sweeps.

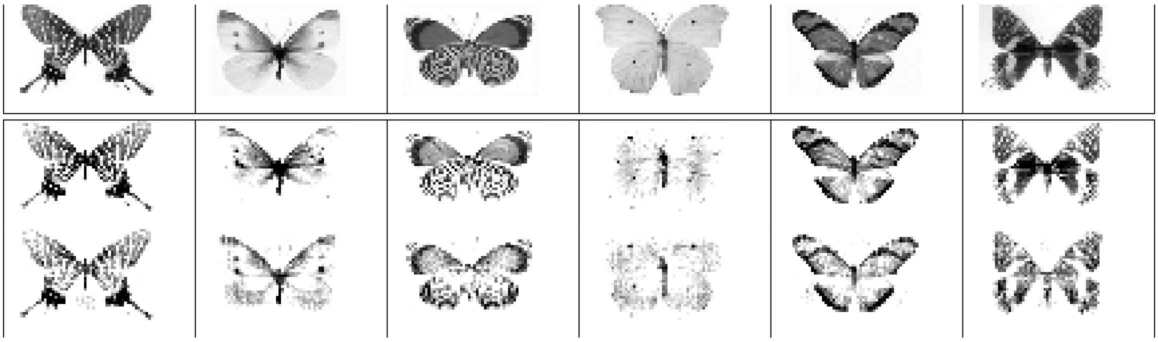
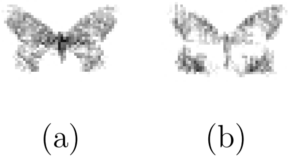


Figure 4.8: Example 4.3. The 6 butterfly images are 34 × 50. First row: original images. Second row: recovered images by Gabor basis with length = 5. Third row: recovered images by Gabor basis with length = 15.

Example 4.3. In Example 4.3, we apply the same algorithm to a set of $M = 6$ butterfly images. The butterfly images are resized to 34×50 (height × length), then 6 butterfly images are represented as 1700×6 image matrix. The parameters set-up are the same with those in Example 4.2. However, Gabor basis of two different scales are adopted for comparison. The lengths of the Gabor basis at these two scales are 5 and 15 respectively. Figure 4.8 and Figure 4.9 display the results. In each block of Figure 4.8, the top plot shows the original image, the middle plot shows the recovered image by Gabor basis with length = 5, and the bottom plot shows the recovered image by Gabor basis with length = 15. The two plots in Figure 4.9 display the two shared models learned by the two different Gabor basis sets, where active Gabor basis are multiplied by the average of the estimated coefficients. From the recovered results on both individual images and shared models, we can see the results by Gabor basis with length = 15 are foggier and cannot detect the details.

Example 4.4. In Example 4.4, two-layer Gibbs sampler is applied to a set of $M = 4$ deer images with different backgrounds. In order to get more clear images,



(a) (b)

Figure 4.9: Example 4.3. The shared models. (a): Gabor basis with length = 5. (b): Gabor basis with length = 15.

the number of sweeps is increased to 2000, and only the last 500 sweeps are used for inference after discarding the first 1500 sweeps. 4971 Gabor bases are chosen as active from the total 12500 bases. The sum of active bases for each image, i.e. the size of S_I, is 8273.

The recovered results are shown in Figure 4.10. Due to the contrast between the deer in the image and the background, the recovered results are different. We can see there are a lot of sketches for the background in the first and the last recovered images, because compare to the deers, the two images have a higher greyscale background. Therefore, the deers are left white in the recovered images. As for the third image, because it has a more consistent and low greyscale background, the recovered image clearly sketch the deer and let the background white.

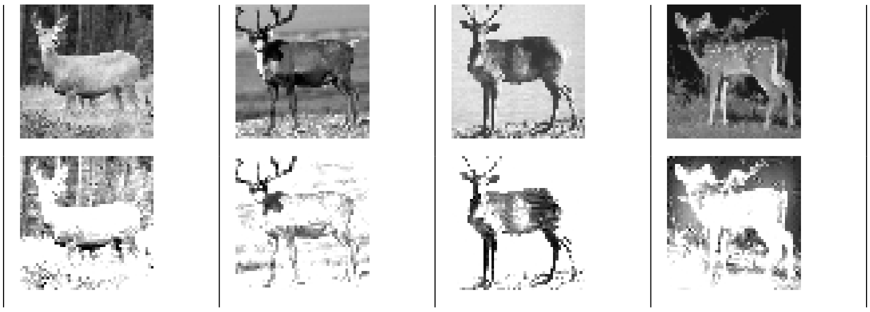


Figure 4.10: Example 4.4. The 4 deer images are 50 × 50. Samples collected from the last 500 sweeps are used for inference after discarding the first 1500 sweeps. First row: original images. Second row: recovered images.

CHAPTER 5

Full Bayesian approach

In the Bayesian framework, the prior distribution is used to express one's uncertainty about the parameter. The parameters of the prior distributions are called hyper parameters. One could assign the given value for a hyper parameter. For example, the failure probabilities, θ_j and $\rho_{j,m}$ in Eq. (2.7) and Eq. (2.12), of the *Bernoulli* prior for the first and second sets of indicators, δ_j and $\eta_{j,m}$, and the variance, $\tau^2_{j,m}$ in Eq. 2.13, of the normal prior for the coefficients, are all pre-specified hyper parameters in the proposed two-layer Gibbs sampler. One could also assign a probability distribution, which is called a hyper-prior, on the hyper parameter itself, so the value of the hyper parameter could be updated by iteration. For example, we assume the hyper-prior distribution of residual variance, σ^2 in Eq. (2.2), follow an inverse *Gamma* distribution, $\sigma^2 \sim IG(a/2, b/2)$. Then it is kept updating by the inverse *Gamma* distribution $IG((a+n\times M)/2, (\sum(diag(Res'Res))+b)/2)$, i.e. the Step 4 in Algorithm 3. In this chapter, we consider to extend the proposed two-layer Gibbs sampler, i.e. algorithm 3, to a full Bayesian framework by adding hyper-priors for the hyper parameters θ_j, $\rho_{j,m}$, and $\tau^2_{j,m}$.

5.1 *Beta* Hyper-Priors for θ_j and $\rho_{j,m}$

First, consider the hyper parameters, θ_j and $\rho_{j,m}$, the failure probabilities of the *Bernoulli* prior in the first and second sets of indicators, δ_j and $\eta_{j,m}$. In the proposed two-layer Gibbs sampler, we assume δ_j and $\eta_{j,m}$ follow the *Bernoulli* distributions respectively as below:

$$\begin{aligned} \delta_j &\sim Ber(1-\theta_j), j = 1, \cdots, p \\ \eta_{j,m}|\delta_j &\sim (1-\delta_j)\gamma_0 + \delta_j Ber(1-\rho_{j,m}), j = 1, \cdots, p,\ m = 1, \cdots, M. \end{aligned}$$

Without any prior information, we simply set the failure rates $\theta_j = \rho_{j,m} = 0.5$, $j = 1, \cdots, p,\ m = 1, \cdots, M$. However, Scott and Berger (2010) [SB10] mentioned that to set the probability at 0.5 in the *Bernoulli* prior might not have multiplicity control, if p is large and the model is sparse. Taking advantage of the characteristic that *Bernoulli* and *Beta* are conjugate distributions, they suggested setting a *Beta* hyper-prior distribution instead of the given the probability $= 0.5$. Thus, the *Beta* hyper-prior assumption is adopted to modify the proposed method here.

In order to reduce the number of hyper parameters and simplify the calculation, we set $\theta_j = \theta$ and $\rho_{j,m} = \rho$ for all $j = 1, \cdots, p$ and $m = 1, \cdots, M$. We assume the hyper-priors of θ and ρ follow $Beta(r, s)$ and $Beta(t, u)$, respectively, and other prior setting of δ_j, $\eta_{j,m}$, $\beta_{j,m}$, and σ^2 are the same as those in the two-layer structure. Then we can add one more step in Algorithm 3 to draw θ and ρ from the posterior distributions listed below:

$$\theta \sim Beta\left(r + p - \sum_j \delta_j, s + \sum_j \delta_j\right), \tag{5.1}$$

$$\rho \sim Beta\left(t + \sum_j \sum_m (1-\eta_{j,m})\delta_j, u + \sum_j \sum_m \eta_{j,m}\delta_j\right), \tag{5.2}$$

where $\sum_j \delta_j$ is the sum of the indicators in the first set, i.e. the number of variables in the support union of the shared model S_S, and $\sum_j \sum_m (1-\eta_{j,m})\delta_j$ and $\sum_j \sum_m \eta_{j,m}\delta_j$ are the sum of two different products of the indicators in first and second sets, which correspond to the total number of zero and nonzero corresponding coefficients of the active variables in all regression models, respectively.

To demonstrate the performance of the modified algorithm that include *Beta* hyper-priors for θ and ρ, we use the data generated by the setting in Example 3.2. Thus there were $M = 5$ response vectors and $p = 200$ predictor variables of length $n = 80$. The true active variable set was $\{X_7, X_8, X_9, X_{11}, X_{12}, X_{19}, X_{20}, X_{21}\}$, i.e. $S_S = \{7, 8, 9, 11, 12, 19, 20, 21\}$, and the corresponding coefficients of 5 single regression models are shown in Table 3.3.

Consider the parameter setting in the *Beta* hyper-priors. According to Castillo and van der Vaar (2012) [CV12], to have small variance in the *Beta* distribution and to meet the sparsity assumption, s and u should be small, and r and t can be decided through the variable size and the number of the response vectors, respectively. We set $s = u = 1$, $r = 32$, and $t = M + 1 = 6$ as the parameters of *Beta* hyper-priors for θ and η in this example, while the hyper parameter τ^2 is set as 20, and σ^2 is assigned by inverse *Gamma* with parameters $a = b = 0.001$. We use the last 200 sweeps from total 500 sweeps for the inference in each replication. Based on the median probability criterion, the selection results of 100 replications of the simulation are shown in Table 5.1. In order to compare the results with those in the sample version of the two-layer Gibbs sampler, Algorithm 3 is applied on the same 100 simulation data, and the same parameter setting is adopted, except θ and ρ are set as the given value 0.5. The results are shown in Table 5.2.

The first column in Table 5.1 and Table 5.2 shows the frequencies of $\hat{P}(\delta_j =$

Table 5.1: The selection frequency of the modified algorithm with *Beta* hyper-priors for θ and ρ

X_j	δ_j	$\eta_{j,1}$	$\eta_{j,2}$	$\eta_{j,3}$	$\eta_{j,4}$	$\eta_{j,5}$
X_7	100	100	100	1	100	100
X_8	100	100	100	100	100	0
X_9	100	100	100	0	0	2
X_{11}	100	1	100	1	0	100
X_{12}	100	100	0	100	100	0
X_{19}	97	1	97	0	78	1
X_{20}	100	0	0	0	3	100
X_{21}	100	100	0	0	0	1

Table 5.2: The selection frequency of the sample version of the two-layer Gibbs sampler

X_j	δ_j	$\eta_{j,1}$	$\eta_{j,2}$	$\eta_{j,3}$	$\eta_{j,4}$	$\eta_{j,5}$
X_7	100	100	100	1	100	100
X_8	100	100	100	100	100	1
X_9	100	100	100	0	0	0
X_{11}	100	0	100	1	2	100
X_{12}	100	100	0	100	100	1
X_{19}	96	0	96	0	75	1
X_{20}	99	0	0	0	1	99
X_{21}	100	100	0	0	0	0

$1|\boldsymbol{Y}) > 0.5$ in 100 replications of the simulation, i.e. the frequencies of including the true active predictors in the shared model. In fact, both two methods work well in recovery of the support union in the shared model S_S. 7 and 6 out of 8 true predictors are chosen as active with 100 percentage in the modified algorithm with *beta* hyperpriors and in sample version of the two-layer Gibbs sampler, respectively, and the 3 others are all equal or higher than 97 percentage.

Consider the indicators in the second set: $\eta_{j,m}$. The left 5 columns in Table 5.1 and Table 5.2 show the frequencies of $\hat{P}(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y}) > 0.5$ for $j \in S_S$ in 100 replications of the simulation, i.e. the frequencies of including the active predictor in each individual model. The cell with grey color means the corresponding coefficient is zero, i.e. the predicator is not active in this particular regression model, and therefore the lower the frequency is better. In Table 5.1 and Table 5.2, the values in grey cells are all equal or smaller than 3. As for the frequencies in white cells, it means the corresponding coefficients are nonzero, only $\eta_{19,2}$: 97, $\eta_{19,4}$: 78 in Table 5.1, and $\eta_{19,2}$: 96, $\eta_{19,4}$: 75, $\eta_{20,5}$: 99 in Table 5.2 are not equal to 100 percentage. In particular, these frequencies correspond to $\beta_{19,2} = 0.6, \beta_{19,4} = 0.4$ and $\beta_{20,5} = 0.7$, which are relatively smaller values.

The average true positive rate, the average probability of correct recovery, and the average false positive rate, the average probability of incorrect recovery, are listed in Table 5.3. The results coming form the two algorithms both give good performance in identify the true 8 predictors in the shared model. The average true positive rates for the indicators in the first set δ_j are 0.9963 and 0.9938 respectively. It means in total 100 replication of simulation, only about 5 and 3 predictors are not correctly included in the shared mode respectively. However, compare the average false positive rates, the modified algorithm has about five times lower average rate, 0.0012, than

0.0059 in sample version of two-layer Gibbs sampler, in false including any of the 192 inactive predictors in the shared model. As for recovery of the support union of the individual models S_I, i.e. finding the particular regression model the predicator is active for, there is no significant difference between the two algorithms. 0.9868 and 0.9842 are the average true positive rates of the 19 active indicators in the second set $\eta_{j,m}$ respectively. The average false positive rate decrease from 0.0052 to 0.0038 by adding *Beta* hyper-priors, the change is not very much neither.

Based on the results, updating the hyper parameters θ and ρ by *Beta* hyper-priors can decrease the false positive rate, prediction error, in both sets of indicators. However, the influence is not significant. Therefore, using the value 0.5 as the given hyper parameters for θ_j and $\rho_{j,m}$ is acceptable.

Table 5.3: The average rates of δ and η.

	δ		η	
	TPR	FPR	TPR	FPR
Sample version of two-layer Gibbs sampling	0.9963	0.0059	0.9868	0.0052
With *Beta* hyper-priors for θ and ρ	0.9938	0.0012	0.9842	0.0038
With inverse *Gamma* hyper-prior for τ^2	0.9463	0.0001	0.9305	0

5.2 Inverse *Gamma* hyper-prior for τ^2

In the two-layer structure, we assume the coefficients follow the distribution as below,

$$\beta_{j,m}|\delta_j, \eta_{j,m} \sim (1-\delta_j\eta_{j,m})\gamma_0 + \delta_j\eta_{j,m}\mathcal{N}(0, \tau_{j,m}^2),$$

i.e. if the variable X_j is active for the m-th regression model, the corresponding coefficient $\beta_{j,m}$ follow the $Normal$ distribution with zero mean and variance $\tau^2_{j,m}$. In order to simplify the calculation, we set $\tau^2_{j,m} = \tau^2$ for all $j = 1, \cdots, p$ and $m = 1, \cdots, M$.

Like what we did in Example 3.4, we can adopt cross validation to tune the proper τ^2 value based on the current data set. However, due to the high sparsity in the coefficient matrix, the value of τ^2 used in the algorithm usually has to be set much higher than the assumption. To modify the current approach, we can assign a hyper-prior distribution for the hyper parameter τ^2. Here we choose inverse *Gamma* with parameters c and d as the hyper-prior for τ^2, i.e. $\tau^2 \sim IG(c/2, d/2)$. To incorporate the hyper-prior in the sample version of the two-layer Gibbs sampler, we can add a step to draw τ^2 from the posterior distribution

$$\tau^2 \sim IG\left(\frac{c}{2}, \frac{d + \sum_m \sum_j \beta^2_{j,m} / \sum_m \sum_j \delta_j \eta_{j,m}}{2}\right), \tag{5.3}$$

where $\sum_m \sum_j \beta^2_{j,m} / \sum_m \sum_j \delta_j \eta_{j,m}$ takes the average of squared coefficients over the support union of the individual models S_I.

The simulation data in Section 5.1 are used to demonstrate the performance. Here we set $c = d = 0.1$ as the parameters in the inverse *gamma* hyper-prior of τ^2. The other hyper parameters are chosen as $\theta_j = \eta_{j,m} = 0.5$, for $j = 1, \cdots, p$, $m = 1, \cdots, M$, and $a = b = 0.001$ for the inverse *Gamma* hyper-prior of σ^2. We run 500 iterations and the last 200 posterior samples are collected. With 100 replications, the frequencies of including the predictors in the shared and in the individual models are shown in Table 5.4.

The under-detection problem happens more frequently both in the shared model S_S, and in the individual models S_I. The frequencies are 65 in δ_{19}, 92 in δ_{20}, 63 in

$\eta_{19,2}$, 15 in $\eta_{19,4}$, and 92 in $\eta_{20,5}$. The poor performance happens especially in the predictor X_{19}, whose two nonzero corresponding coefficients are relatively smaller: $\beta_{19,2} = 0.6$, and $\beta_{19,4} = 0.4$. However, it has very good performance in not including the false predictors in each of the individual models. The values in the grey cells in Table 5.4 are all equal to zero. It means in the 100 replication of simulation, the modified algorithm with inverse *Gamma* hyper-prior for τ^2 didn't have any chance to include the false predictors in all individual regression models.

The TPR and FPR shown in the third row of Table 5.3 also reveal this phenomenon. The TPR of including the active predictors in the shared and in the individual models are 0.9463 and 0.9305 respectively. Both values are lower than the rates in the sample version of two-layer Gibbs sampler and in the modified algorithm with *Beta* hyper-priors. However, the FPR of incorrectly including the inactive in the shared and in the individual models are extremely low as 0.0001 and 0, which definitely have the dominant position.

Table 5.4: The selection frequency of the modified algorithm with inverse *Gamma* hyper prior for τ^2

X_j	δ_j	$\eta_{j,1}$	$\eta_{j,2}$	$\eta_{j,3}$	$\eta_{j,4}$	$\eta_{j,5}$
X_7	100	100	100	0	100	100
X_8	100	100	100	100	100	0
X_9	100	100	100	0	0	0
X_{11}	100	0	100	0	0	100
X_{12}	100	100	0	100	100	0
X_{19}	65	0	63	0	15	0
X_{20}	92	0	0	0	0	92
X_{21}	100	100	0	0	0	0

CHAPTER 6

Model Selection Consistency

In this chapter, we investigate the theoretical properties of our proposed procedure in terms of model selection consistency in sparse Bayesian variable selection. Let $S_{m,0}$ denote the true individual model for $m = 1, \cdots, M$, and S_0 denote the true multi-response model which consider all M singular models together. We use $\beta^*_{j,m}$ as the true parameter of variable X_j in m-response vector, and $B*$ for the true parameter matrix. In fact, consistency of an estimate in linear regression includes the model selection consistency and parameter estimation consistency. However, the two issues are independent. In this article, we focus on the consistency of model selection. Once the consistency of model selection is proven, the corresponding parameters can be estimated through specified technique.

Theorem 1. *Let $\boldsymbol{Y} = [Y_1, \cdots, Y_M]$ be the multiple response matrix and $\boldsymbol{X}$ be the $n \times p$ design matrix, where p is fixed. Assume $(\boldsymbol{X}'\boldsymbol{X})/n \to C$ when $n \to \infty$, where C is positive definite and the set that is the collection of possible models, S, contains the true model S_0, and median model S_{me}, which is the model defined with median criterion. Fix $\theta_j > 0, \rho_{j,m} > 0, \tau_{j,m} > 0$, and assume*

$$\sqrt{2\pi}\tau_{(j,1)} > \min\{\frac{1-\theta_j}{\theta_j}, 1\} \text{ and } \sqrt{2\pi}\tau_{j,m} > \frac{1-\rho_{j,m}}{\rho_{j,m}}, \tag{6.1}$$

for all $j = 1, \cdots, p$ and $m = 1, \cdots, M$, where $\tau_{(j,1)} = min\{\tau_{j,1}, \cdots, \tau_{j,M}\}$. Let $\tau_{(1)}$ be *the minimum of* $\tau_{(j,1)}, j = 1, \cdots, p$. *Then we have*

$$\lim_{n\to\infty} P(S_{me} = S_0|\boldsymbol{Y}) = 1. \tag{6.2}$$

With the minor restrictions shown in Eq. 6.1, the theorem shows the proposed selection results based on the median probability criterion converge to the true model. It means if the number of observations is large enough, we can find the true model.

The consistency of model selection means that the true model will be eventually selected if there is enough data. According to Zhao and Yu (2006) [ZY06], the requirement in the multiple linear regression is: for each m $\in \{1, \cdots, M\}$:

$$P(\{j : \hat{\beta}_{j,m} \neq 0\} = \{j : \beta^*_{j,m} \neq 0\}) \to 1, \text{as } n \to \infty. \tag{6.3}$$

Based on the posterior median criterion and the two-layer model we proposed, we could rewrite the requirement in Eq. (6.3) in two statements:

$$\begin{aligned}
&P(\{j : P(\delta_j = 1|\boldsymbol{Y}) > \frac{1}{2}\} \\
&= \{j : \exists\ m \in \{1, \cdots, M\}, \beta^*_{j,m} \neq 0\}) \to 1, \text{as } n \to \infty, \text{and} \qquad (6.4)\\
&P(\{j : P(\eta_{j,m} = 1|Y_m) > \frac{1}{2}\} \\
&= \{j : \beta^*_{j,m} \neq 0\}) \to 1, \text{as } n \to \infty, \text{for each m} \in \{1, \cdots, M\}. \qquad (6.5)
\end{aligned}$$

The first statement is for the shared model, and the second statement is for the singular models. It means the posterior median model will eventually coincide with the true model.

Before the proof of Theorem 1, we need the following two Lemmas.

Lemma 1. *Let* Y_m *be the* m*-th response vector and* $\boldsymbol{X}$ *be the* $n \times p$ *shared design matrix, where* p *is fixed.* $S_m = \{j : \beta_{j,m} \neq 0\}$ *is the support set for the* m*-th* *response vector. Let $\boldsymbol{X}_{S_m}$ denote the sub matrix of $\boldsymbol{X}$ with columns corresponding to the set S_m, and $\beta^\star_{S_m}$ denote the vector of true parameters for model S_m. Note that with β_{S_m} that we refer to the subset of parameters included in model S_m. Assume $(\boldsymbol{X}'\boldsymbol{X})/n \to C$ when $n \to \infty$, where C is positive define. By conditioning on S_m that we mean model S_m is deemed to be the true model. There exist a product measure $P^\infty_{S_m}$ on $(\mathcal{R}^\infty, \mathcal{B}(\mathcal{R}\infty))$ such that there exist $\Omega \in \mathcal{B}(\mathcal{R}\infty)$, of $P^\infty_{S_m}$-probability 1, such that*

$$\beta_{S_m}|S_m, Y_m \xrightarrow{P} \beta^\star_{S_m}, \tag{6.6}$$

where $\xrightarrow{P}$ denotes convergence in probability. Further,

$$\sqrt{n}(E[\beta_{S_m}|S_m, Y_m] - \beta^\star_{S_m}) \xrightarrow{d} N(0, \sigma^2 C^{-1}_{S_m}), \tag{6.7}$$

where $\xrightarrow{d}$ denotes convergence in distribution and note that $E[\beta_{S_m}|S_m, Y_m]$ is a random variable as a function of Y_m.

Proof. For the condition in the following proof, we need

$$\frac{\boldsymbol{X}'_{S_m}\boldsymbol{X}_{S_m}}{n} \xrightarrow[n\to\infty]{} D_m,$$

for some D_m that is positive definite matrix. We take a matrix E_{S_m} whose columns are elementary vectors which corresponding to S_m. From the assumption, we have $\lim_{n\to\infty}(\boldsymbol{X}'\boldsymbol{X})/n = C$, where C is positive define. Applying the Theorem 5.2 in Friedberg [FWR02], we obtain that

$$\lim_{n\to\infty} \frac{\boldsymbol{X}'_{S_m}\boldsymbol{X}_{S_m}}{n} = \lim_{n\to\infty} \frac{E'_{S_m}\boldsymbol{X}'\boldsymbol{X}E_{S_m}}{n} = E'_{S_m}CE_{S_m} = C_{S_m}, \tag{6.8}$$

where C_{S_m} that corresponds to S_m is the principal sub matrix of C, and C_{S_m} is positive definite, since every principal sub matrix of a positive definite matrix is positive definite.

If S_m is assumed as the true model, we have $\beta_{S_m}|S_m \sim N(\mathbf{0}, \sigma^2 I_{d_m})$, where $d_m = \#S_m$ is chosen from Θ which is an open set of R^{d_m} and the prior density of $\beta_{S_m}|S_m$ is continuous and positive on Θ. Using the condition in Eq. (6.8), we have two asymptotic theorems from Ferguson [Fer96] in the the following.

From Chapter 21 in Ferguson [Fer96], the asymptotic of posterior distribution is

$$\sqrt{n}(B_{S_m}|S_m, Y_m - \hat{B}_{S_m,n}) \xrightarrow{d} N(0, J^{-1}(\beta^\star_{S_m})), \tag{6.9}$$

where $\hat{\beta}_{S_m,n}$ is the MLE of β_{S_m} and $J(\beta^\star_{S_m}) = (\boldsymbol{X}'_{S_m}\boldsymbol{X}_{S_m})/\sigma^2_{S_m}$ is the Fisher information. It implies that

$$\beta_{S_m}|S_m, Y_m - \hat{\beta}_{S_m,n} \xrightarrow{P} 0. \tag{6.10}$$

According to Theorem 18 in Ferguson [Fer96], it leads that

$$\sqrt{n}(\hat{\beta}_{S_m,n} - \beta^\star_{S_m}) \xrightarrow{d} N(\mathbf{0}, J^{-1}(\beta^\star_{S_m})), \tag{6.11}$$

and it also implies

$$\hat{\beta}_{S_m,n} - \beta^\star_{S_m} \xrightarrow{P} \mathbf{0}. \tag{6.12}$$

Therefore, we have

$$\begin{aligned}
&\lim_{n\to\infty} P(|\beta_{S_m}|S_m, Y_m - \beta^\star_{S_m}| > \epsilon) \\
&\leqslant \lim_{n\to\infty} P(|\beta_{S_m}|S_m, Y_m - \hat{\beta}_{S_m,n}| > \frac{\epsilon}{2}) + \lim_{n\to\infty} P(|\hat{\beta}_{S_m,n} - \beta^\star_{S_m}| > \frac{\epsilon}{2}) = 0.
\end{aligned}$$

based on Eq. (6.10) and Eq. (6.12). That is,

$$\beta_{S_m}|S_m, Y_m \xrightarrow{P} \beta^\star_{S_m}. \tag{6.13}$$

According to Eq. (6.8), we further can obtain

$$\sqrt{n}(E[\beta_{S_m}|S_m, Y_m] - \beta^*_{S_m}) \xrightarrow{d} N(\mathbf{0}, \sigma^2 C^{-1}_{S_m}), \tag{6.14}$$

□

Lemma 2. *Let $\boldsymbol{Y} = [Y_1, \cdots, Y_M]$ be the $n \times M$response matrix in the multi-response linear regression model, $Y_m = (Y_{m,1}, \cdots, Y_{m,n})'$ be the m-th response vector, and $\boldsymbol{X}$ be the $n \times p$ shared design matrix, where p is fixed. Assume $(\boldsymbol{X}'\boldsymbol{X})/n \to C$ when $n \to \infty$, where C is positive definite. Assume the set that is the collection of all possible models, S, contains the true model S_0, and the median model S_{me}, which is defined with median criterion. Fix $\theta_m > 0, \rho_{j,m} > 0$ and $\tau_{j,m} > 0$, $j = 1, \cdots, p, m = 1, \cdots, M$. If the prior distribution of $\beta_{j,m}|\delta_m, \eta_{j,m}$ is defined as Eq. (2.13), we have*

$$\lim_{n\to\infty} P(S_0|\boldsymbol{Y}) = 1.$$

Proof. Let the prior distribution of $\beta_{j,m}$ is defined as

$$\beta_{j,m}|\delta_j, \eta_{j,m} = (1 - \delta_j\eta_{j,m})N(0, a_n^2) + \delta_j\eta_{j,m}N(0, \tau_{j,m}^2), \tag{6.15}$$

where $0 < a_n < \tau_{j,m}$, $\forall m, j$, and $a_n \to 0$ when $n \to \infty$. In this prior assumption, we need prove that

$$P(\delta_j = 1|\boldsymbol{Y}) \to \begin{cases} 1, & \exists\ m \in \{1, \cdots, M\} : j \in S_{m,0}, \\ 0, & \forall\ m \in \{1, \cdots, M\} : j \notin S_{m,0}, \end{cases} \tag{6.16}$$

and

$$P(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y}) \to \begin{cases} 1, & j \in S_{m,0}, \\ 0, & j \notin S_{m,0}, \end{cases} \tag{6.17}$$

when $a_n \to 0$, where $S_{m,0}$ is the true model for the m-th singular regression model.

From Eq. (6.15), given any $Q \in S$ for Lemma 1, we have $\beta_Q|Q \sim N(\mathbf{0}, \sigma^2 I_d)$, where $d = \#Q$, is chosen from Θ that is an open set of R^d and the prior density of $\beta_Q|Q$ is continuous and positive on Θ. Therefore, we can obtain $\beta_Q|Q, Y_m \xrightarrow{P} \beta_Q^\star, \forall\ Q \in S$. That is, for each response vector Y_m

$$\beta|Y_m \xrightarrow{P} \beta_m^\star. \tag{6.18}$$

If we consider all M response vectors together at the same time, Eq. (6.18) becomes

$$B|\boldsymbol{Y} \xrightarrow{P} B^{\star}. \tag{6.19}$$

Hence

$$\begin{aligned}
\lim_{n\to\infty} P(\delta_j = 1|\boldsymbol{Y}) &= \lim_{n\to\infty} \frac{P(\delta_j = 1, \boldsymbol{Y})}{P(\boldsymbol{Y})} \\
&= \lim_{n\to\infty} \int \frac{P(\beta^{(j)}, \delta_j = 1, \boldsymbol{Y})}{P(\boldsymbol{Y})} d\beta^{(j)} \\
&= \lim_{n\to\infty} \int \frac{P(\beta^{(j)}, \delta_j = 1, \boldsymbol{Y})}{P(\beta^{(j)}, \boldsymbol{Y})} \frac{P(\beta^{(j)}, \boldsymbol{Y})}{P(\boldsymbol{Y})} d\beta^{(j)} \\
&= \lim_{n\to\infty} \int P(\delta_j = 1|\beta^{(j)}, \boldsymbol{Y}) dF(\beta^{(j)}|\boldsymbol{Y}) \\
&= P(\delta_j = 1|\beta^{(j)\star}), \qquad (6.20)
\end{aligned}$$

where Eq. (6.20) holds by Eq. (6.19), and $\beta^{(j)\star} = (\beta^{\star}_{j,1}, \cdots, \beta^{\star}_{j,M})$ is the vector of true coefficients for variable X_j in the multi-response linear regression model. Then the posterior probability of δ_j converges to

$$\begin{aligned}
P(\delta_j = 1|\beta^{(j)\star}) &= \frac{P(\beta^{(j)\star}, \delta_j = 1)}{P(\beta^{(j)\star})} \\
&= \frac{\sum_{\eta^{(j)}} P(\beta^{(j)\star}, \eta^{(j)}, \delta_j = 1)}{\sum_{\eta^{(j)}} P(\beta^{(j)\star}, \eta^{(j)}, \delta_j = 1) + P(\beta^{(j)\star}, \eta^{(j)} = \mathbf{0}, \delta_j = 0)} \\
&= \frac{1}{1 + \frac{\theta_j}{1-\theta_j} C_1}, \qquad (6.21)
\end{aligned}$$

where

$$
\begin{aligned}
C_1 &= \frac{P(\beta^{(j)\star}, \eta^{(j)} = \mathbf{0}|\delta_j = 0)}{\sum_{\eta^{(j)}} P(\beta^{(j)\star}, \eta^{(j)}|\delta_j = 1)} \\
&= \frac{\prod\limits_{m=1}^{M} P(\beta^{\star}_{j,m}|\eta_{j,m} = 0, \delta_j = 0)}{\sum_{\eta^{(j)}} \prod\limits_{m=1}^{M} P(\beta^{\star}_{j,m}|\eta_{j,m}, \delta_j = 1) P(\eta_{j,m}|\delta_j = 1)} \\
&= \frac{\prod\limits_{m=1}^{M} \frac{1}{a_n} \exp\left(-\frac{\beta^{\star 2}_{j,m}}{2a_n^2}\right)}{\sum_{\eta^{(j)}} \prod\limits_{m=1}^{M} \left(\rho_{j,m} \frac{1}{a_n} \exp\left(-\frac{\beta^{\star 2}_{j,m}}{2a_n^2}\right)\right)^{1-\eta_{j,m}} \left((1-\rho_{j,m}) \frac{1}{\tau_{j,m}} \exp\left(-\frac{\beta^{\star 2}_{j,m}}{2\tau^2_{j,m}}\right)\right)^{\eta_{j,m}}} \\
&= \left[\sum_{\eta^{(j)}} \prod_{m=1}^{M} \rho_{j,m} \left(\frac{1-\rho_{j,m}}{\rho_{j,m}} \frac{a_n}{\tau_{j,m}} \exp\left(-\frac{\beta^{\star 2}_{j,m}}{2} \frac{a_n^2 - \tau^2_{j,m}}{a_n^2 \tau^2_{j,m}}\right)\right)^{\eta_{j,m}}\right]^{-1}.
\end{aligned}
$$

If j should not be included in any $S_{m,0}$, i.e. $\beta^{\star}_{j,m} = 0$, $\forall$ m, then we have

$$
P(\delta_j = 1|\beta^{(j)\star}) = \frac{1}{1 + \frac{\theta_j}{1-\theta_j} \left[\sum_{\eta^{(j)}} \prod\limits_{m=1}^{M} \rho_{j,m} \left(\frac{1-\rho_{j,m}}{\rho_{j,m}} \frac{a_n}{\tau_{j,m}}\right)^{\eta_{j,m}}\right]^{-1}}. \tag{6.22}
$$

It converges to 0 when $a_n \to 0$. That is, $P(\delta_j = 1|\boldsymbol{Y}) \to 0$, when $a_n \to 0$, if j should not be included in any $S_{m,0}$.

If j should be included in at least one $S_{m,0}$, it means at least one $\beta^{\star}_{j,m} \neq 0$, $m \in \{1, \cdots, M\}$. Because when $\beta_{j,m} \neq 0$, we have $\frac{a_n}{\tau_{j,m}} \exp\left(-\frac{\beta^2_{j,m}}{2} \frac{a_n^2 - \tau^2_{j,m}}{a_n^2 \tau^2_{j,m}}\right) \to \infty$ when $a_n \to 0$. In this case, $C_1 \to 0$ when $a_n \to 0$. Therefore, we get $P(\delta_j = 1|\boldsymbol{Y}) \to 1$ when $a_n \to 0$, if $\exists$ $m \in \{1, \cdots, M\} : j \in S_{m,0}$. Then Eq. (6.16) is proven.

Similarly, we can have

$$
\begin{aligned}
P(\delta_j = 0|\beta^{(j)\star}) &= \frac{P(\beta^{(j)\star}, \eta^{(j)} = \mathbf{0}, \delta_j = 0)}{\sum_{\eta^{(j)}} P(\beta^{(j)\star}, \eta^{(j)}, \delta_j = 1) + P(\beta^{(j)\star} \eta^{(j)} = \mathbf{0}, \delta_j = 0)} \\
&= \frac{1}{\frac{1-\theta_j}{\theta_j} C_2 + 1},
\end{aligned} \tag{6.23}
$$

where

$$C_2 = \frac{\sum_{\eta^{(j)}} P(\beta^{(j)\star}\eta^{(j)}|\delta_j = 1)}{P(\beta^{(j)\star}, \eta^{(j)} = \mathbf{0}|\delta_j = 0)} = C_1^{-1}.$$

Using the same argument, we can get the result:

$$P(\delta_j = 0|\boldsymbol{Y}) \to \begin{cases} 0, & \exists\ m \in \{1, \cdots, M\} : j \in S_{m,0}, \\ 1, & \forall\ m \in \{1, \cdots, M\} : j \notin S_{m,0}, \end{cases} \tag{6.24}$$

Next we consider the active support in each singular model. Suppose we know j is included in at least one $S_{m,0}$, then we want to know on which specific model the variable is active through the posterior conditional probability:

$$P(\eta_{j,m} = 1|\boldsymbol{Y}) = P(\eta_{j,m} = 1, \delta_j = 1|\boldsymbol{Y}) = P(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y})P(\delta_j = 1|\boldsymbol{Y}). \tag{6.25}$$

Then

$$\begin{aligned}
\lim_{n\to\infty} P(\eta_{j,m} = 1, \delta_j = 1|\boldsymbol{Y}) &= \lim_{n\to\infty} \frac{P(\eta_{j,m} = 1, \delta_j = 1, \boldsymbol{Y})}{P(\boldsymbol{Y})} \\
&= \lim_{n\to\infty} \int \frac{P(\beta_{j,m}, \eta_{j,m} = 1, \delta_j = 1, \boldsymbol{Y})}{P(\boldsymbol{Y})} d\beta_{j,m} \\
&= \lim_{n\to\infty} \int \frac{P(\beta_{j,m}, \eta_{j,m} = 1, \delta_j = 1, \boldsymbol{Y})}{P(\beta_{j,m}, \boldsymbol{Y})} \frac{P(\beta_{j,m}, \boldsymbol{Y})}{P(\boldsymbol{Y})} d\beta_{j,m} \\
&= \lim_{n\to\infty} \int P(\eta_{j,m} = 1, \delta_j = 1|\beta_{j,m}, \boldsymbol{Y}) dF(\beta_{j,m}|\boldsymbol{Y}) \\
&= P(\eta_{j,m} = 1, \delta_j = 1|\beta^\star_{j,m}).
\end{aligned} \tag{6.26}$$

Here Eq. (6.26) is held due to Eq. (6.19) and it can be expressed as

$$
\begin{aligned}
&P(\eta_{j,m}=1,\delta_j=1|\beta^\star_{j,m})\\
&=\frac{P(\beta^\star_{j,m},\eta_{j,m}=1,\delta_j=1)}{P(\beta^\star_{j,m})}\\
&=\frac{P(\beta^\star_{j,m},\eta_{j,m}=1,\delta_j=1)}{P(\beta^\star_{j,m},\eta_{j,m}=1,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=0)}\\
&=\frac{(1-\theta_j)(1-\rho_{j,m})}{(1-\theta_j)(1-\rho_{j,m})+((1-\theta_j)\rho_{j,m}+\theta_j)\,\frac{\tau_{j,m}}{a_n}\exp\left(-\frac{\beta^{\star 2}_{j,m}}{2}(\frac{1}{a_n^2}-\frac{1}{\tau^2_{j,m}})\right)}. \qquad (6.27)
\end{aligned}
$$

If j is included in $S_{m,0}$, i.e. $\beta^\star_{j,m}\neq 0$, Eq. (6.27) converges to 1 when $a_n\to 0$. If j is not included in $S_{m,0}$, i.e. $\beta^\star_{j,m}=0$, Eq. (6.27) converges to 0 when $a_n\to 0$. That is, when $a_n\to 0$, we have that Eq. (6.17) is obtained.

Similarly, using the same approach, we can have

$$
\begin{aligned}
&\lim_{n\to\infty}P(\eta_{j,m}=0,\delta_j=1|\boldsymbol{Y})\\
&=P(\eta_{j,m}=0,\delta_j=1|\beta^\star_{j,m})\\
&=\frac{P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=1)}{P(\beta^\star_{j,m},\eta_{j,m}=1,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=0)}\\
&=\frac{(1-\theta_j)\rho_{j,m}}{(1-\theta_j)(1-\rho_{j,m})\frac{a_n}{\tau_{j,m}}\exp\left(-\frac{\beta^{\star 2}_{j,m}}{2}(\frac{1}{\tau^2_{j,m}}-\frac{1}{a_n^2})\right)+(1-\theta_j)\rho_{j,m}+\theta_j},
\end{aligned}
$$

and

$$
\begin{aligned}
&\lim_{n\to\infty}P(\eta_{j,m}=0,\delta_j=0|\boldsymbol{Y})\\
&=P(\eta_{j,m}=0,\delta_j=0|\beta^\star_{m})\\
&=\frac{P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=0)}{P(\beta^\star_{j,m},\eta_{j,m}=1,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=1)+P(\beta^\star_{j,m},\eta_{j,m}=0,\delta_j=0)}\\
&=\frac{\theta_j}{(1-\theta_j)(1-\rho_{j,m})\frac{a_n}{\tau_{j,m}}\exp\left(-\frac{\beta^{\star 2}_{j,m}}{2}(\frac{1}{\tau^2_{j,m}}-\frac{1}{a_n^2})\right)+(1-\theta_j)\rho_{j,m}+\theta_j}.
\end{aligned}
$$

Then we can get the result

$$P(\eta_{j,m}=0,\delta_j=1|\boldsymbol{Y}) \to \begin{cases} 0, & j \in S_{m,0}, \\ \frac{(1-\theta_j)\rho_{j,m}}{(1-\theta_j)\rho_{j,m}+\theta_j}, & j \notin S_{m,0}, \end{cases} \tag{6.28}$$

$$P(\eta_{j,m}=0,\delta_j=0|\boldsymbol{Y}) \to \begin{cases} 0, & j \in S_{m,0}, \\ \frac{\theta_j}{(1-\theta_j)\rho_{j,m}+\theta_j}, & j \notin S_{m,0}. \end{cases} \tag{6.29}$$

When $a_n \to 0$, it implies that the prior distribution of $\beta_{j,m}$ converges to Eq. (2.13). Therefore, Eq. (6.16), Eq. (6.17), Eq. (6.24), Eq. (6.28), and Eq. (6.29) hold if the prior distribution of $\beta_{j,m}$ is defined as Eq. (2.13). Assume the true model S_0 is identified by the inclusion $A_{j,m}=\{(j,m):\eta_{j,m}=1\}$ and exclusion of the remaining $B_{j,m}=\{(j,m):\eta_{j,m}=0\}$. Let S be the model chosen from all 2^p possible models, then

$$\begin{aligned}\lim_{n\to\infty} P(S=S_0|\boldsymbol{Y}) &= \lim_{n\to\infty} P\left(\bigcap A_{j,m}\bigcap B_{j,m}|\boldsymbol{Y}\right) \\ &= \lim_{n\to\infty}\left[\prod P(A_{j,m}|\boldsymbol{Y})\prod P(B_{j,m}|\boldsymbol{Y})\right] \\ &= 1\end{aligned}$$

Because

$$\begin{aligned}\lim_{n\to\infty} P(A_{j,m}|\boldsymbol{Y}) &= \lim_{n\to\infty} P(\eta_{j,m}=1|\boldsymbol{Y}) = 1, \\ \lim_{n\to\infty} P(B_{j,m}|\boldsymbol{Y}) &= \lim_{n\to\infty} P(\eta_{j,m}=0|\boldsymbol{Y}) \\ &= \lim_{n\to\infty} P(\eta_{j,m}=0,\delta_j=1|\boldsymbol{Y}) + \lim_{n\to\infty} P(\eta_{j,m}=0,\delta_j=0|\boldsymbol{Y}) \\ &= 1.\end{aligned}$$

through Eq. (6.17), Eq. (6.28), and Eq. (6.29). Therefore, we can obtain

$$\lim_{n\to\infty} P(S_0|\boldsymbol{Y}) = 1. \tag{6.30}$$

□

Proof. Proof of Theorem 1

To accomplish the proof, based on the median probability criterion, we need to show that the two sets of indicator variables should follow the equations as below

$$P(\tilde{\delta}_j = 1|\boldsymbol{Y}, S_0) \to P_j \begin{cases} \geqslant 1/2, & \exists\, m \in \{1, \cdots, M\} : j \in S_{m,0}, \\ < 1/2, & \forall\, m \in \{1, \cdots, M\} : j \notin S_{m,0}, \end{cases} \tag{6.31}$$

$$P(\eta_{j,m} = 1|\tilde{\delta}_j = 1, \boldsymbol{Y}, S_0) \to P_{j,m} \begin{cases} \geqslant 1/2, & j \in S_{m,0}, \\ < 1/2, & j \notin S_{m,0}, \end{cases} \tag{6.32}$$

where $\tilde{\delta}_j = \delta_j \times \mathbf{1}(\eta^{(j)} \neq \mathbf{0})$.

From Lemma 1, by conditional on model $S_{m,0}$, we have

$$\beta_{m,0}|Y_m, S_{m,0} \xrightarrow{P} \beta^{\star}_{m,0}. \tag{6.33}$$

Let $S^c_{m,0}$ be the complement of $S_{m,0}$, i.e. $S^c_{m,0}$ is the set with inactive variables for the m-th singular model, then

$$\begin{aligned}
\lim_{n\to\infty} P(|\beta^c_{m,0}| > \epsilon|Y_m, S_{m,0}) &= \lim_{n\to\infty} \frac{P(|\beta^c_{m,0}| > \epsilon, Y_m|S_{m,0})}{P(Y_m|S_{m,0})} \\
&= \lim_{n\to\infty} \frac{\int P(\beta_{m,0}, |\beta^c_{m,0}| > \epsilon, Y_m|S_{m,0}) d\beta_{m,0}}{\int P(\beta_{m,0}, Y_m|S_{m,0}) d\beta_{m,0}} \\
&= \lim_{n\to\infty} \frac{P(|\beta^c_{m,0}| > \epsilon|S_{m,0}) \int P(\beta_{m,0}, Y_m||\beta^c_{m,0}| > \epsilon, S_{m,0}) d\beta_{m,0}}{\int P(\beta_{m,0}, Y_m||\beta^c_{m,0}| > \epsilon, S_{m,0}) d\beta_{m,0}} \\
&= \lim_{n\to\infty} P(|\beta^c_{m,0}| > \epsilon|S_{m,0}) \\
&= 0,
\end{aligned}$$

that is,

$$\beta^c_{m,0}|S_{m,0}, Y_m \xrightarrow{P} \beta^{\star c}_{m,0} = 0. \tag{6.34}$$

Using Eq (6.33) and Eq. (6.34), we have

$$\beta|S_{m,0}, Y_m \xrightarrow{P} \beta^\star_m, \forall\ m = 1, \cdots, M. \tag{6.35}$$

Hence

$$\begin{aligned}
\lim_{n\to\infty} P(\tilde{\delta}_j = 1|\boldsymbol{Y}, S_0) &= \lim_{n\to\infty} \int P(\beta^{(j)}, \tilde{\delta}_j = 1|\boldsymbol{Y}, S_0) d\beta^{(j)} \\
&= \lim_{n\to\infty} \int \frac{P(\beta^{(j)}, \tilde{\delta}_j = 1, \boldsymbol{Y}|S_0)}{P(\boldsymbol{Y}|S_0)} d\beta^{(j)} \\
&= \lim_{n\to\infty} \int \frac{P(\beta^{(j)}, \tilde{\delta}_j = 1, \boldsymbol{Y}|S_0)}{P(\beta^{(j)}, \boldsymbol{Y}|S_0)} \frac{P(\beta^{(j)}, \boldsymbol{Y}|S_0)}{P(\boldsymbol{Y}|S_0)} d\beta^{(j)} \\
&= \lim_{n\to\infty} \int P(\tilde{\delta}_j = 1|\beta^{(j)}, \boldsymbol{Y}, S_0) dF(\beta^{(j)}|\boldsymbol{Y}, S_0) \\
&= P(\tilde{\delta}_j = 1|\beta^{(j)\star}, S_0), \qquad (6.36)
\end{aligned}$$

where the Eq. (6.36) holds by Eq. (6.35). Then the posterior probability of $\tilde{\delta}_j$ is

$$\begin{aligned}
&P(\tilde{\delta}_j = 1|\beta^{(j)\star}, S_0) \\
&= \frac{P(\beta^{(j)\star}, \tilde{\delta}_j = 1|S_0)}{P(\beta^{(j)\star}|S_0)} \\
&= \frac{\sum_{\eta^{(j)}\neq \mathbf{0}} P(\beta^{(j)\star}, \eta^{(j)}, \tilde{\delta}_j = 1|S_0)}{\sum_{\eta^{(j)}\neq \mathbf{0}} P(\beta^{(j)\star}, \eta^{(j)}, \tilde{\delta}_j = 1|S_0) + P(\beta^{(j)\star}, \eta^{(j)} = \mathbf{0}, \tilde{\delta}_j = 0|S_0)}. \qquad (6.37)
\end{aligned}$$

If $\exists\ m \in \{1, \cdots, M\}$ such that $j \in S_{m,0}$, there is at least one model m such that $\beta^\star_{j,m} \neq 0$. Therefore, $P(\beta^{\star(j)}, \tilde{\delta}_j = 0, \eta^{(j)} = \mathbf{0}|S_0)$ should be equal to 0 and then we obtain that

$$P(\tilde{\delta}_j = 1|\beta^{(j)\star}, S_0) = 1 \tag{6.38}$$

based in Eq. (6.37). That is $\lim_{n\to\infty} P(\tilde{\delta}_j = 1|\boldsymbol{Y}, S_0) \geqslant 1/2$ if variable X_j is in at least one true singular model $S_{m,0}$.

If variable X_j is inactive in any singular model m, i.e. $\beta^\star_{j,m} = 0, \forall\ m$. Then Eq. (6.37) can be expressed as

$$\begin{aligned}
&\frac{1}{1+\frac{P(\beta^{\star(j)},\eta^{(j)}=\mathbf{0},\tilde{\delta}_j=0|Q_0)}{\sum_{\eta^{(j)}\neq\mathbf{0}}P(\beta^{(j)\star},\eta^{(j)},\tilde{\delta}_j=1|Q_0)}}\\
&=\frac{1}{1+\frac{\theta_j}{(1-\theta_j)\sum_{\eta^{(j)}\neq\mathbf{0}}\prod_{m=1}^{M}\left((1-\rho_{j,m})\frac{1}{\sqrt{2\pi}\tau_{j,m}}\right)^{\eta_{j,m}}(\rho_{j,m})^{1-\eta_{j,m}}}}\\
&=\frac{1}{1+\frac{\theta_j}{(1-\theta_j)}\left[\sum_{\eta^{(j)}\neq\mathbf{0}}\prod_{m=1}^{M}\left(\frac{1}{\sqrt{2\pi}\tau_{j,m}}\right)^{\eta_{j,m}}(\rho_{j,m})^{1-\eta_{j,m}}\left(1-\rho_{j,m}\right)^{\eta_{j,m}}\right]^{-1}}. \qquad (6.39)
\end{aligned}$$

Let $\tau_{(j,1)} = \min\{\tau_{j,1}, \cdots, \tau_{j,M}\}$, then we have

$$\sum_{\eta^{(j)}\neq\mathbf{0}}\left(\prod_{m=1}^{M}\left(\frac{1}{\sqrt{2\pi}\tau_{j,m}}\right)^{\eta_{j,m}}(\rho_{j,m})^{1-\eta_{j,m}}\left(1-\rho_{j,m}\right)^{\eta_{j,m}}\right)\leqslant\frac{1}{\sqrt{2\pi}\tau_{(j,m)}}.$$

Therefore, based on the assumption in Eq. (6.1), we can show that Eq. (6.39) is less than 1/2. That is, $\lim_{n\to\infty} P(\tilde{\delta}_j = 1|\boldsymbol{Y}, S_0) < 1/2$ if X_j is an inactive variable in all singular models. Therefore, Eq. (6.31) is obtained.

Next we consider the consistency of the support in each singular model. Suppose variable X_j is active in the m-th singular model, then the following equation

$$P(\eta_{j,m} = 1|\boldsymbol{Y}, S_0) = P(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y}, S_0)P(\delta_j = 1|\boldsymbol{Y}, S_0) \qquad (6.40)$$

holds with prior assumption. Hence we can get

$$\lim_{n\to\infty} P(\eta_{j,m} = 1|\boldsymbol{Y}, S_0) = \lim_{n\to\infty} P(\eta_{j,m} = 1|\delta_j = 1, \boldsymbol{Y}, S_0) \qquad (6.41)$$

via Eq. (6.36) and Eq. (6.38). Then

$$\begin{aligned}
\lim_{n\to\infty} P(\eta_{j,m}=1|\boldsymbol{Y},S_0) &= \lim_{n\to\infty} \frac{P(\eta_{j,m}=1,\boldsymbol{Y},S_0)}{P(\boldsymbol{Y},S_0)} \\
&= \lim_{n\to\infty} \int \frac{P(\beta_{j,m},\eta_{j,m}=1,\boldsymbol{Y},S_0)}{P(\boldsymbol{Y},S_0)} d\beta_{j,m} \\
&= \lim_{n\to\infty} \int \frac{P(\beta_{j,m},\eta_{j,m}=1,\boldsymbol{Y},S_0)}{P(\beta_{j,m},\boldsymbol{Y},S_0)} \frac{P(\beta_{j,m},\boldsymbol{Y},S_0)}{P(\boldsymbol{Y},S_0)} d\beta_{j,m} \\
&= \lim_{n\to\infty} \int P(\eta_{j,m}=1|\beta_{j,m},\boldsymbol{Y},S_0) dF(\beta_{j,m}|\boldsymbol{Y},S_0) \\
&= P(\eta_{j,m}=1|\beta^{\star}_{j,m},S_0), \qquad (6.42)
\end{aligned}$$

where Eq. (6.42) is held based on Eq. (6.35), and can be expressed as

$$\begin{aligned}
P(\eta_{j,m}=1|\beta^{\star}_{j,m},S_0) &= P(\eta_{j,m}=1,\tilde{\delta}_j=1|\beta^{\star}_{j,m},S_0) \\
&= \frac{P(\beta^{\star}_{j,m},\eta_{j,m}=1,\tilde{\delta}_j=1|S_0)}{P(\beta^{\star}_{j,m}|S_0)} \\
&= \frac{P(\beta^{\star}_{j,m},\eta_{j,m}=1,\tilde{\delta}_j=1|S_0)}{P(\beta^{\star}_{j,m},\eta_{j,m}=1,\tilde{\delta}_j=1|S_0)+P(\beta^{\star}_{j,m},\eta_{j,m}=0,\tilde{\delta}_j=1|S_0)}, \qquad (6.43)
\end{aligned}$$

because

$$P(\beta^{\star}_{j,m},\eta_{j,m}=0,\tilde{\delta}_j=0|S_0) = P(\beta^{\star}_{j,m},\eta_{j,m}=0|\tilde{\delta}_j=0,S_0)P(\tilde{\delta}_j=0|S_0)=0$$

Suppose j is included in $S_{m,0}$, it means $\beta^{\star}_{j,m} \neq 0$, then Eq. (6.43) is equal to 1 because $P(\beta^{\star}_{j,m},\eta_{j,m}=0,\tilde{\delta}_j=1|S_0)=0$. On the other hand, if X_j is not active in the m-th singular model, that is $\beta^{\star}_{j,m}=0$, Eq. (6.43) becomes

$$\frac{1}{1+\frac{\rho_{j,m}}{1-\rho_{j,m}}\sqrt{2\pi}\tau_{j,m}}. \qquad (6.44)$$

Therefore based on the assumption in Eq. (6.1), we obtain Eq. (6.43) is less than1/2. Thus Eq. (6.32) is proven. Combine Eq. (6.31) and Eq. (6.32)

$$\lim_{n\to\infty} P(S_{me}=S_0|\boldsymbol{Y},S_0)=1. \qquad (6.45)$$

Finally, we have

$$\begin{aligned}\lim_{n\to\infty} P(S_{me}=S_0|\boldsymbol{Y}) = &\lim_{n\to\infty} P(S_{me}=S_0|\boldsymbol{Y},S_0)P(S_0|\boldsymbol{Y})+\\ &\lim_{n\to\infty}\sum_{k=1}^{2^p-1} P(S_{me}=S_0|\boldsymbol{Y},S_k)P(S_k|\boldsymbol{Y})\\ \geqslant &\lim_{n\to\infty} P(S_{me}=S_0|\boldsymbol{Y},S_0)P(S_0|\boldsymbol{Y})\\ =&1,\end{aligned}$$

according to Eq. (6.45) and Lemma 2, and the proof of Theorem 1 is completed. □

CHAPTER 7

Empirical Results

In this chapter, with the proposed two-layer Gibbs sampler for learning the multi-response linear regression, we conduct a number of numerical simulations to evaluate the performance of support union recovery on different finite sample size. Depending on the different number of tasks and on the different sparsities of the regression vectors, we study how the sample size affect the accuracy of support union recovery both in the shared model and in individual models. Besides, we also compare the performance of the proposed two-layer Gibbs sampler and the sparse group Lasso. In each replication, we measure the true positive rate (TPR), false positive rate (FPR), and accuracy, of the first and the second set of indicator variables, δ_j and $\eta_{j,m}$, respectively. True positive rate, which is also called the sensitivity, measures the proportion of the variables in the support union which are correctly chosen as active. It is a measure of correct recovery. False positive rate measures the proportion of the variables outside the support union which are mis-identified as active. It measures the prediction errors. Accuracy measures the proportion of true results of all variables. For TPR and accuracy, the higher the better, and for FPR, the lower the better. The performance is evaluated by taking average over 100 replications.

7.1 Simulation on Different Number of Tasks

In the first study, we consider the scenario when the number of tasks M varies. We set the predicator vector X_j as

$$X_j = G_j + G,$$

where $G_j s$ and G are independently generated from multivariate normal distribution with mean zero and covariance matrix I_n. Then the correlation between any two predictors is 0.5. The sparsity of the linear regression vector is linear proportional to the dimension p, i.e., $s = \alpha p$, where α is the parameter that controls the sparsity of the model. We set $\alpha = 1/16$ and choose two different sizes of regressors $p = \{128, 256\}$. In the setting of the coefficients, half of the active variables in the support union have nonzero coefficients over all tasks, one quarter of the active variables in the support union have nonzero coefficients in half tasks, and the remaining quarter of the active variables in the support union have nonzero coefficient in quarter of the tasks. The values of the nonzero coefficients are chosen randomly from $\{0.5, 1, 2, 3\}$. We apply the sample version of two-layer Gibbs sampler for support union recovery with $M = \{4, 8, 12, 16\}$. The setting of prior parameters are: $\theta_{j,m} = \rho_{j,m} = 0.5, \tau_{j,m}^2 = 20, a = b = 0.001$ for all $m = 1, \cdots, M, j = 1, \cdots, p$.

Fig. 7.1 shows results of support union recovery of the shared model with two different number of predictors $p = \{128, 256\}$, which are displayed in two different rows respectively. After 100 replications, the mean value of the true positive rate (TPR), the false positive rate (FPR), and the accuracy versus the rescaled sample size $r = n/[2slog(p - s)]$, where $s = |S_S|$ is the number of variables in the support union of the shared model, are shown in the left, the middle, and the right column, respectively. It shows the increase in the number of tasks do improve the performance of support union recovery of the shared model, no matter in true positive rate, in false positive rate, or in the accuracy. Given the same rescaled sample size, when the number of tasks increases, the true positive rate, i.e. the sensitivity, and the accuracy increase, and the false positive rate, i.e. the prediction errors, decrease. Besides, the results show that with the two-layer Gibbs sampler, the support union recovery in the shared model rapidly reaches very good performance when the rescaled sample size is equal or bigger than 0.5. There are sharp increases or sharp decreases when the rescaled sample size increase from 0.2 to 0.5. Therefore, by pooling data across tasks, two-layer Gibbs sampler can efficiently help related tasks collaborate with each other to detect the true active variables. Furthermore, once the rescaled sample size is big enough, the proposed Bayesian method can achieve high precision on support union recovery in the shared model.

Fig. 7.2 shows results of support union recovery of the individual models in two different dimensions $p = \{128, 256\}$. Generally, the performance gets better when the number of tasks increase. And when the rescaled sample size get bigger, there is no much difference between models with different number of tasks, because all achieve high performance of recovery, i.e, the TPR and accuracy are close to 1, and the FPR is close to 0. Therefore, with the proposed two-layer Gibbs sampler, we can successfully and efficiently recover the support for each individual model simultaneously.

7.2 Simulation on Different Sparsity Ratios

In this section, we study how the sparsity ratio affect the results of support union recovery. We fix the number of tasks $M = 8$, and study three linear sparsity ratios: $\alpha = \{1/8, 1/16, 1/32\}$. We set the dimension $p = \{128, 256\}$. The setting of the

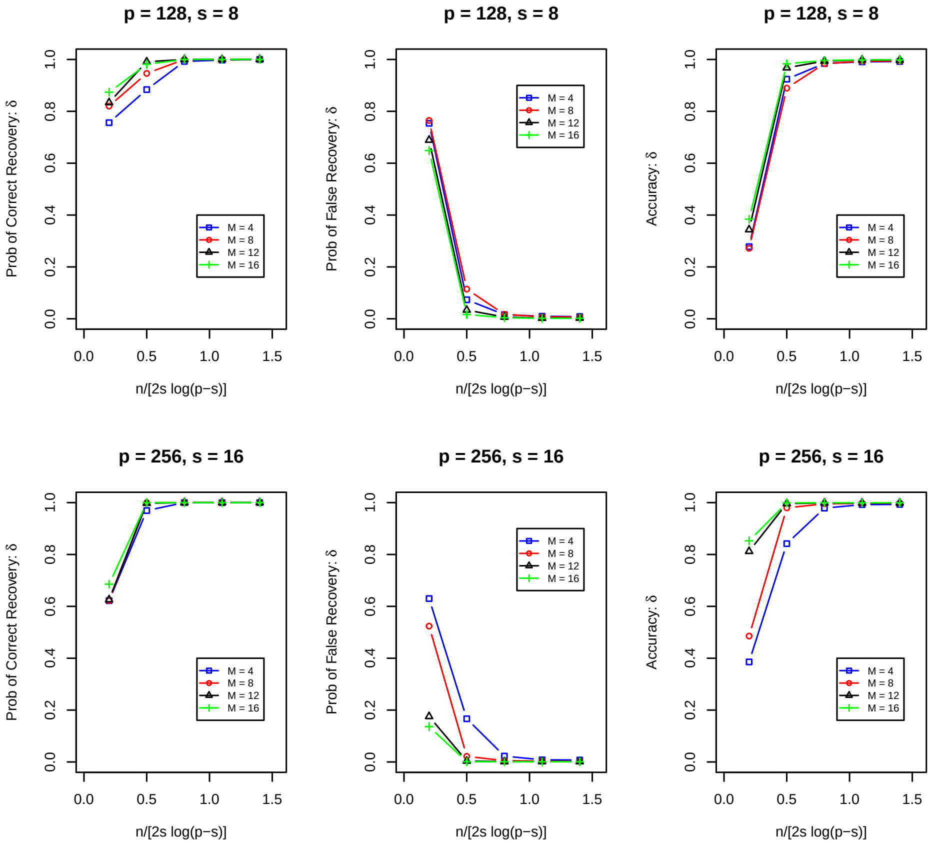


Figure 7.1: Plots of support union recovery of the shared models, S_S, versus the control parameter $r = n/[2s\log(p-s)]$ with different number of tasks: $M = \{4, 8, 12, 16\}$. The two rows present results for the number of regressors $p = 128, 256$, respectively.

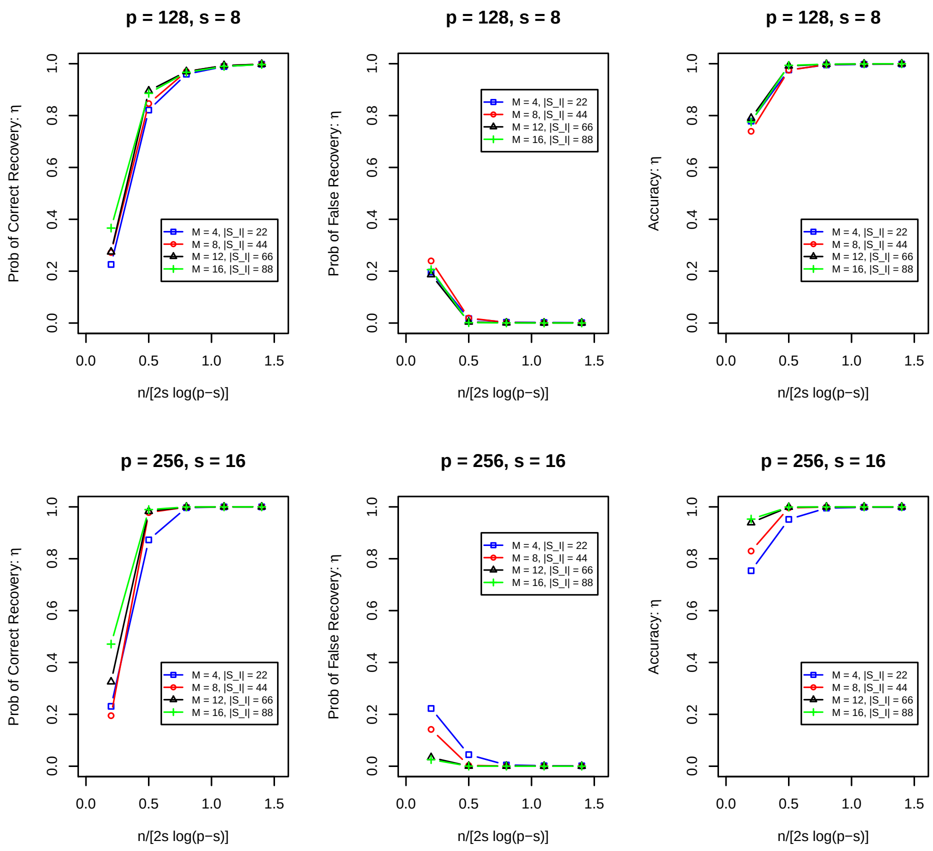


Figure 7.2: Plots of support union recovery of the individual models, S_I, versus the control parameter $r = n/[2s\log(p-s)]$ with different number of tasks: $M = \{4, 8, 12, 16\}$. The two rows present results for the number of regressors $p = 128, 256$, respectively.

predicator vector X_j, the coefficients, and the prior parameters are the same with those in Sec. 7.1.

The results of support union recovery in the shared model and in the individual models are shown in Fig. 7.3 and Fig. 7.4. In both figures, the influence of sparsity ratio happens apparently when the rescaled sample size is small. When rescaled sample size is small, there is significant different between models with different sparsity ratio. The lower the sparsity ratio, the lower the TPR and accuracy, and the higher the FPR. However, once the rescaled sample size is big enough, the influence of sparsity decreases, because the results show that all models achieve high precision in recovery rapidly. The only exception case is when the sparsity ratio is 1/32, and the rescaled sample size is as low as 0.2. In this situation, the sample size used to do union support recovery is just 8 and 17 when the corresponding predictor number is 128 and 256 respectively. Therefore, although the TPR is higher than the other two sparsity ratios 1/8 and 1/16, the FPR is also the highest. It means in this situation, with the lack of sample size, it can't correctly recover the true active variables, and has serious over-selection problem. Once the sample size gets larger, the problem disappears.

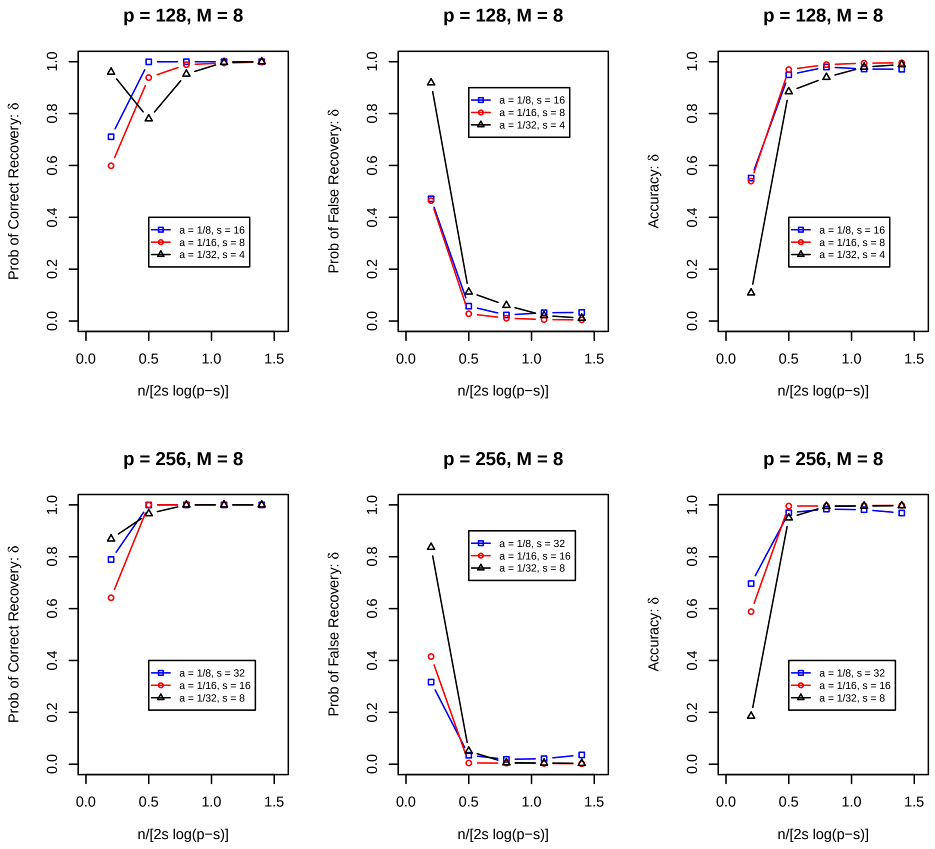


Figure 7.3: Plots of support union recovery of the shared models, S_S, versus the control parameter $r = n/[2s \log(p - s)]$ with different sparsity ratios: $\alpha = \{1/8, 1/16, 1/32\}$. The two rows present results for the number of regressor $p = 128, 256$, respectively.

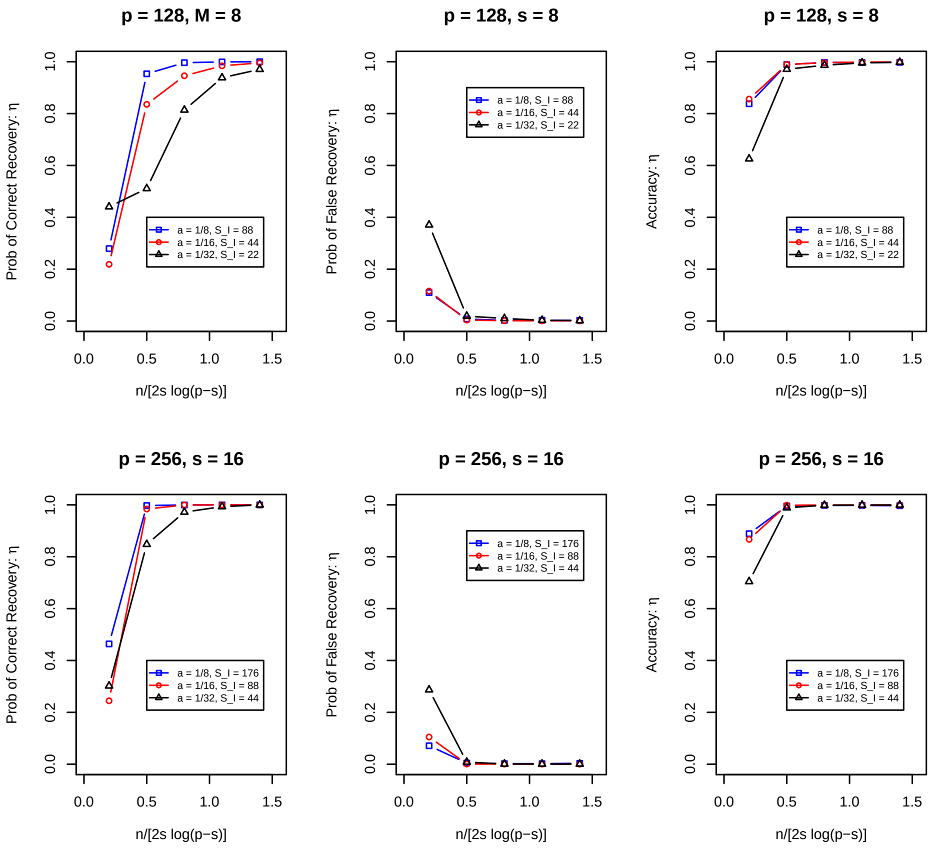


Figure 7.4: Plots of support union recovery of the individual models, S_I, versus the control parameter $r = n/[2s\log(p-s)]$ with different sparsity ratios: $\alpha = \{1/8, 1/16, 1/32\}$. The two rows present results for the number of regressor $p = 128, 256$, respectively.

7.3 Simulation on Different Methods

In this simulation, we compare the performance between the proposed sample version of two-layer Gibbs sampler and sparse group Lasso. We set the number of tasks $M = 8$, sparsity ratio $\alpha = 1/8$, and the dimension of predictors $p = \{128, 256, 512\}$. Fig. 7.5 shows results of support union recovery in the shared model with three different number of predictors $p = \{128, 256, 512\}$, which are displayed in three different rows respectively. The performance of true positive rate is shown in the left column. Two lines by the two different methods seem to very close, but the proposed Bayesian method achieves high precision faster than sparse group Lasso when the rescaled sample size increases.

As for the false positive rate and accuracy, when the rescaled sample size is as small as 0.2, the proposed Bayesian method has worse performance, due to the problem of over-selection. However, once the rescaled sample size equal or bigger than 0.5, the situation is reversed. In all three cases with different dimension of predictors, the proposed Bayesian method has a sudden sharp decrease in the false positive rates, and a sudden sharp increase in accuracy, when the rescaled sample size increase from 0.2 to 0.5. However, the influence of rescaled sample size on the sparse group Lasso is less apparent.

Fig. 7.6 shows results of support union recovery in individual models. In the three different measures, TPR, FPR and accuracy, the performance of the two-layer Gibbs sampler is better than that in sparse group Lasso, except when the rescaled sample size is 0.2. With the proposed Bayesian method, when the rescaled sample size is equal or bigger than 0.5, the values of true positive rate and the accuracy are close to 1, and the value of false positive rate is close to 0. Combine the results together, the proposed two-layer Gibbs sampler do have better recovery performance on the support union in the shared model and in the individual models simultaneously.

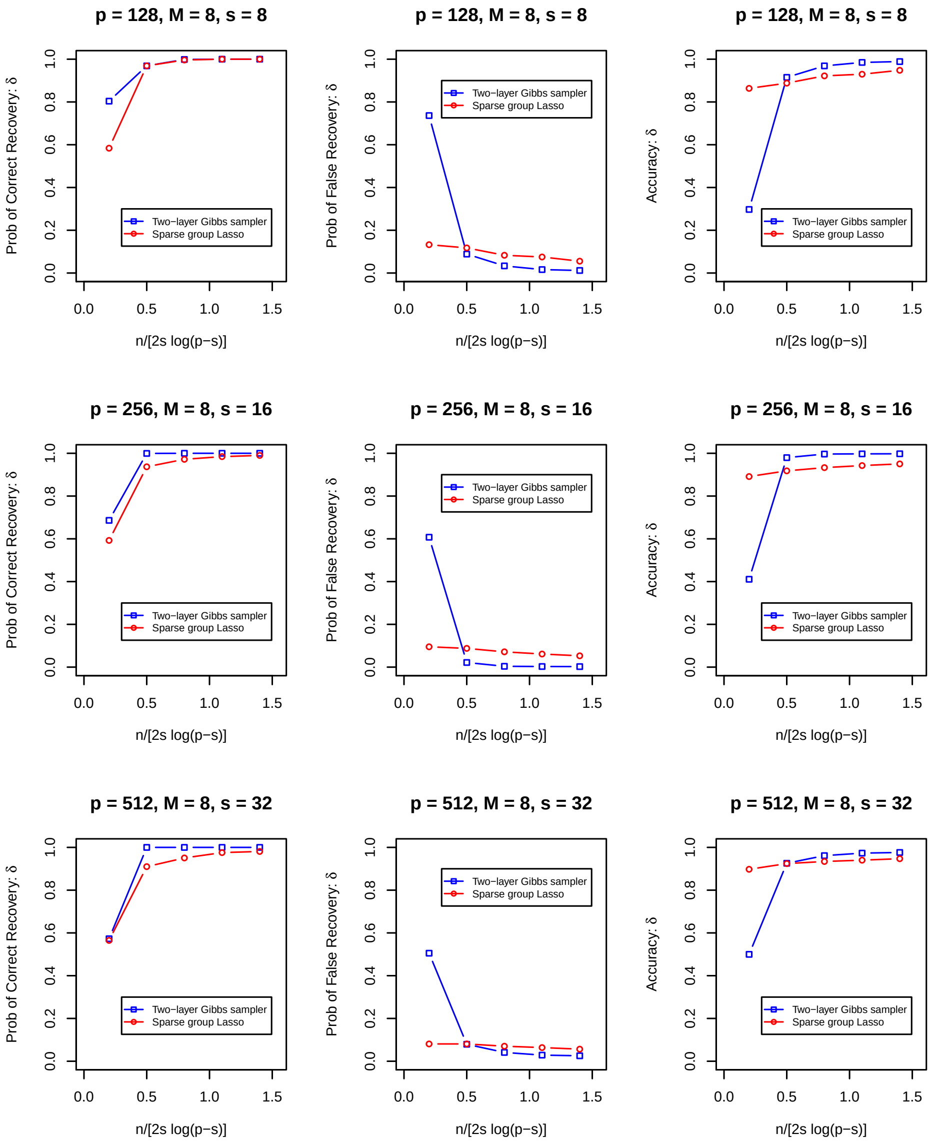


Figure 7.5: Plots of support union recovery of the shared models, S_S, versus the control parameter $r = n/[2s\log(p-s)]$ by two-layer Gibbs sampler and sparse group Lasso. The three rows present results for the number of regressor $p = 128, 256, 512$, respectively.

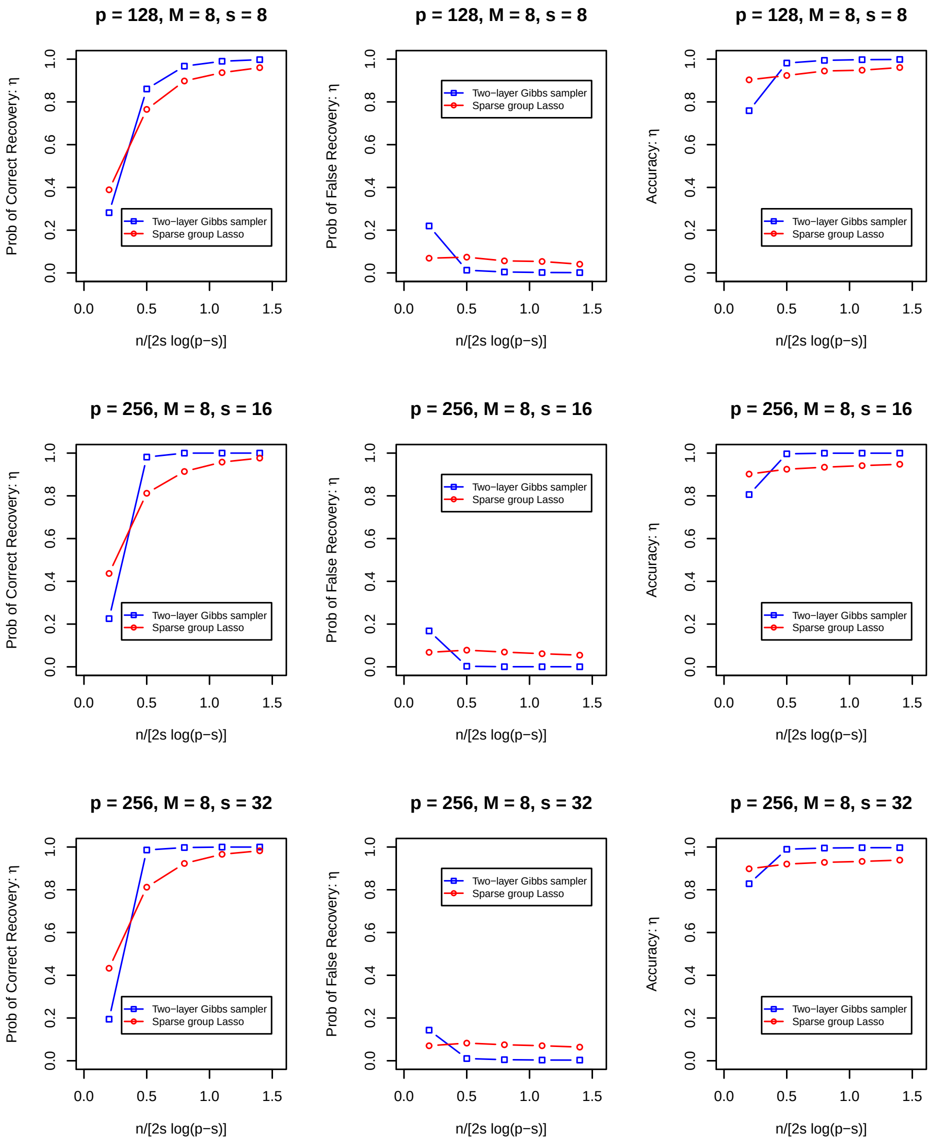


Figure 7.6: Plots of support union recovery of the individual models, S_I, versus the control parameter $r = n/[2s\log(p-s)]$ by two-layer Gibbs sampler and sparse group Lasso. The three rows present results for the number of regressor $p = 128, 256, 512$, respectively.

CHAPTER 8

Conclusion

A Bayesian variable selection method is studied in this work for recovering the union of support sets, where two nested sets of indicators are augmented into the multi-response linear regression model. Firstly, a two-layer Gibbs sampler based on the two-layer setting is proposed for the posterior sampling. The posterior probabilities of indicators are computed with likelihood ratio functions. By sampling the two sets of indicator variables with the posterior probabilities, the union of the support sets can be recovered, and active variables for specific responses can be identified. Secondly, after learning the multi-response linear model, variable coefficients can be estimated using posterior samples of indicators. Furthermore, a sample version of two-layer Gibbs sampler with the Metropolis-Hastings acceptance rejection rule is introduced to improve the performance. Instead of posterior probability, transition probability is used to check whether variables are kept in the current state.

To evaluate the presented approach, a simulation study is conducted, showing the promise on identifying active variables in multi-response linear regression models. The result shows that using the sample version of the two-layer Gibbs sampler can improve the performance, reducing around 35% cost on computation. The presented sample-version approach also achieves high precision compared to Lasso methods in terms of finding active variables. For instance, in a simulation case that our two-layer approach finds exactly all the 6 active variables out of 200 variables, two Lasso methods identify around 14 active variables with 10 variables that are not active (false positives) and two active variables that are not selected (false negatives). Finally, on the study of sketching images with Gabor basis, we show that a shared sketch of an object can be found effectively from different images that have the same object.

We have observed that the selection results can be sensitive to the set-up of the prior parameters, e.g., values of success probability of indicators, and values of coefficient variance. To address this issue, instead of using fixed values, we propose using *Beta* hyper-prior for the success probabilities and using inverse Gamma hyper-prior for the coefficient variance. The experimental result shows that the proposed approach significantly reduces false positives on both sets of indicators compared to a pre-given value.

Finally, the asymptotical property of the proposed Bayesian method is investigated and proved. An empirical study is also conducted. The result confirms the property.

As for the final remark, in this study, we consider the homoscedastic model, i.e. the covariance matrix of the error vector in each single task shares the same identity matrix. How to extend the proposed method to the heteroscedastic multi-response model can be an interesting project.

References

[AEP08] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. “Convex multi-task feature learning.” *Machine Learning*, **73**(3):243–272, 2008.

[BB04] Maria M. Barbieri and James O. Berger. “Optimal Predictive Model Selection.” *The Annals of Statistics*, **32**(3):870–897, 2004.

[BH03] Bart Bakker and Tom Heskes. “Task Clustering and Gating for Bayesian Multitask Learning.” *Journal of Machine Learning Research*, **4**:83–99, 2003.

[Car97] Rich Caruana. “Multitask Learning.” *Machine Learning*, **28**(1):41–75, 1997.

[CCCnt] R. B. Chen, Y. C. Chen, C. H. Chu, and K. J. Lee. “On the Determinants of the 2008 Financial Crisis: A Bayesian Approach to the Selection of Groups and Variables.” preprint.

[CCL11] Ray-Bing Chen, Chi-Hsiang Chu, Te-You Lai, and Ying Nian Wu. “Stochastic matching pursuit for Bayesian variable selection.” *Statistics and Computing*, **21**(2):247–259, 2011.

[CV12] Ismal Castillo and Aad van der Vaart. “Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences.” *Ann. Statist.*, **40**(4):2069–2101, 08 2012.

[DFN00] P. Dellaportas, J. J. Forster, and I. Ntzoufras. “Bayesian Variable Selection Using the Gibbs Sampler.”, 2000.

[EP04] Theodoros Evgeniou and Massimiliano Pontil. “Regularized multi–task learning.” In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *KDD*, pp. 109–117. ACM, 2004.

[Far10] A. Farcomeni. “Bayesian constrained variable selection.” *Statistica Sinica.*, **20**(3):1043–1062, 2010.

[Fer96] Thomas S. Ferguson. *A Course in Large Sample Theory.* Chapman and Hall, 1996.

[FL01] J. Fan and R. Li. “Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties.” *Journal of the American Statistical Association*, **96**:1348–1360, 2001.

[FWR02] E C Friedberg, R Wagner, and M Radman. “Specialized DNA polymerases, cellular survival, and the genesis of mutations.” *Science*, (296):1627–1630, 2002.

[GM93] E. I. George and R. E. McCulloch. “Variable Selection via Gibbs Sampling.” *Journal of the American Statistical Association*, **88**:881–889, 1993.

[I12] Abhishek Kumar 0001 and Hal Daum III. “Learning Task Grouping and Overlap in Multi-task Learning.” In *ICML*. icml.cc / Omnipress, 2012.

[LJY09] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[LPT09] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. “Taking Advantage of Sparsity in Multi-Task Learning.” In *Proceedings of the 22nd Conference on Information Theory*, pp. 73–82, June 2009.

[OF96] B. Olshausen and D. Field. “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images.” *Nature*, **381**:607–609, 1996.

[OTJ06] Guillaume Obozinski, Ben Taskar, and Michael Jordan. “Multi-task feature selection.” *Statistics Department, UC Berkeley, Tech. Rep*, 2006.

[OWJ11] G. Obozinski, M. J. Wainwright, and M. I. Jordan. “Support union recovery in high-dimensional multivariate regression.” *The Annals of Statistics*, **39**(1):1–47, 2011.

[RFW09] S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth. “The Bayesian group-Lasso for Analyzing Contingency Tables.” In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML ’09, pp. 881–888, New York, NY, USA, 2009. ACM.

[SB10] James G. Scott and James O. Berger. “Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem.” *The Annals of Statistics*, **38**(5):2587–2619, October 2010.

[ST12] N. Simon and R. Tibshirani. “Standardization and the group lasso penalty.” *Statistica Sinica.*, **22**(3):983–1001, 2012.

[Tib96] R. Tibshirani. “Regression shrinkage and selection via the Lasso.” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[Tip01] M. E. Tipping. “Sparse Bayesian Learning and the Relevance Vector Machine.” *Journal of Machine Learning Research*, **1**:211–244, 2001.

[Tro06] J. A. Tropp. “Just relax: convex programming methods for identifying sparse signals in noise.” *IEEE Transactions on Information Theory*, **52**(3):1030–1051, 2006.

[YL06] M. Yuan and Y. Lin. “Model selection and estimation in regression with grouped variables.” *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**(1):49–67, 2006.

[ZH03] H. Zou and T. Hastie. “Regularization and Variable Selection via the Elastic Net.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2):301–320, 2003.

[Zha10] C.-H. Zhang. “Nearly unbiased variable selection under minimax concave penalty.” *The Annals of Statistics*, **38**(2):894–942, 04 2010.

[ZY06] Peng Zhao and Bin Yu. “On Model Selection Consistency of Lasso.” *Journal of Machine Learning Research*, **7**:2541–2563, 2006.