L1-Penalized Ordinal Regression and Bayesian Variable Selection for Linear Models with Multiple Responses

by

Ya-Ting Chang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2015

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Polychotomous ordinal response data are often analyzed by first introduce a latent continuous variable which can be modeled as an ordinary regression problem with the presence of covariates by using Markov chain Monte Carlo techniques. For variable selection purpose, we modified this approach by using the idea of Stochastic EM algorithm to infuse L-1 penalized regression in estimating the parameter of interest. This allows us to rank the variables in their order of significance based on posterior selection probabilities. We make comparisons with univariate Bayesian variable selection in the simulation and applied the proposed algorithm on data obtained from the MovieLens Project and the World Value Survey.

Given the convenience of using Gibbs sampler to sample from the posterior distributions and choosing prior distributions based on the problems of our interest in Bayesian analyses, we extended the variable selection problem to consider multiple response data by allowing different sets of variables to be selected for different response variables through the infusion of additional information into the prior distribution of the selection variables. This contrasts with the usual approach to multiple response variable selection that selects a common set of variables for all of the response variables. In the simulation, we compared our proposed method against univariate Bayesian variable selection and it shows that the performance is improved after the infusion of relationship information.

Acknowledgements

I would like to thank my supervisor, Professor Zhu for always being there to answer all my questions. He can always help me find solutions and new ideas. Without him, this thesis would not be possible. I would also like to thank Professor Martin Lysy for his valuable inputs and being on my thesis committee along with Professor Chong Zhang.

Special thanks to my classmates, especially Mi, Qin, and Honglei for bringing laughter into my life. Finally, to my parents and siblings for their supports and understandings.

Table of Contents

Chapter 1

Introduction

Polychotomous ordinal response data arise often in surveys that involve rankings, where the answers are categorical but have an underlying order. This type of data can be analyzed by first introduce a latent continuous variable. The presence of covariates allows us to impose an ordinary regression model on the latent variable that can be transformed back into corresponding ordinal levels. Such models can be easily fitted with standard Markov chain Monte Carlo techniques. In this thesis, we consider one of the most important problems in statistical modeling: variable selection, for polychotomous ordinal response data.

Variable selection has been a popular problem in statistical modeling. As such, many methods have been proposed. A number of methods such as L-1 penalized regression and Bayesian variable selection have received great attention recently. In analyzing ordinal response data, the method known as data augmentation has made it easy to incorporate existing variable selection techniques in the latent regression step. Historically, extended Bayesian variable selection techniques are used to analyze dichotomous ordinal response data. We propose to incorporate an existing constrained optimization problem in the MCMC algorithm as a selection technique for the analysis of polychotomous response data. Doing so allows us to rank the importance of the covariates naturally by their respective selection probabilities. Other than simulations, the proposed algorithms have been applied to data obtained from the MovieLens Project and the World Value Survey.

We extended the variable selection problem to consider multiple response data: while the majority of work in multivariate variable selection focuses on selecting a common set of variables for each of the responses, we are interested in selecting different sets of variables for different responses. This is because of the possibility that the sets of variables with the greatest influence on the responses might be different for each response. We allowed the process of selecting variables for each response to communicate with one another through the infusion of the relationship information into the conditional prior distribution. Furthermore, we showed that the resulting posterior distribution has some desirable properties. The proposed method has been applied on the World Value Survey data.

1.1 Contributions

This thesis first focuses on analyzing polychotomous ordinal response data, the techniques proposed can be seen as a generalization and are easily applied to dichotomous ordinal response data, as well as multinomial response data by making some modifications. Other than polychotomous ordinal response data, a method is developed for the analysis of multiple continuous response data. After all, we show that the proposed methods are computationally convenient and also address the importance of prior specification in Bayesian analysis.

1.2 Outline

The remainder of the thesis is organized as follows. Chapter 2 reviews some basic concepts which build the foundation of our proposed methods. Chapter 3 introduces the proposed algorithm, presents some simulation results, the comparison with other methods, and the applications to real datasets. Chapter 4 shows an extension of Bayesian variable selection to analyzing multiple response data, some simulation results, and real data analysis. Finally, we summarize our findings, draw conclusions and suggest future research opportunities and challenges that might be encountered in Chapter 5.

Chapter 2

Background

In this chapter, we briefly review the Gibbs sampler and some variations and applications of Gibbs sampler that build the foundation for the analysis of ordinal response data and our proposed method, which include Expectation Maximization (EM), Stochastic EM algorithms, Maximum A Posteriori, and Collapsed Gibbs.

2.1 Gibbs Sampling

With the increase in complexity of models in statistical analyses, the joint distribution of the parameters are usually intractable hence difficult to sample from and make inference directly. However, an approximation algorithm that uses the idea of a Markov chain has been proposed by [10] and [9] has demonstrated its application in calculating Bayesian posterior densities. The procedure can be summarized as follows. We are interested in the joint distribution defined by $f(\Theta)=f(\theta_1, \theta_2, \ldots, \theta_p)$. Gibbs sampler can be implemented as follows

- We begin with some initial value $\Theta^{(0)}$
- For each $i \in \{1, \ldots, p\}$, sample $\theta_i^{(t)}$ from $\theta_i^{(t)} \sim f(\theta_i^{(t)} | \theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_p^{(t-1)})$

until convergence. The samples approximates the joint distribution.

2.2 Expectation Maximization (EM) algorithm

It is often encountered in applications of statistics where the data are incomplete or cannot be analyzed directly with information available. However, in the case of data being incomplete, we cannot simply dispose of the observations with missing values since it might contain important information. Expectation Maximization [7] iterates between two steps as suggested by its name. The Expectation step calculates the expectation of the loglikelihood of the complete data with respect to the conditional distribution of the augmented data given observed data under the current estimate of the parameters at the iteration. The Maximization step then updates the estimate of the parameters so that the expectation calculated from the previous step is maximized.

Let $X = (Y, Z)$ be a complete set of data where Y is observed and Z is augmented (latent).

Expectation Step:

$$\text{Compute } Q(\theta|\theta^{(t-1)}) = E_{Z|Y,\theta^{(t-1)}}[logf(X|\theta)] = E_{Z|Y,\theta^{(t-1)}}[logf(Y,Z|\theta)]$$

where $\theta^{(t-1)}$ is the current estimate of θ.

Maximization Step:

$$\theta^{(t)} = \arg\max_{\theta} Q(\theta|\theta^{(t-1)})$$

The algorithm is proven to increase the observed data likelihood function at each iteration, that is,

$$L(\theta^{(t+1)}) = logf(Y|\theta^{(t+1)}) \geq L(\theta^{(t)})$$

However, when several stationary points are present, there is no guarantee that $\theta^{(t)}$ will converge to a maximum likelihood estimate. Moreover, the EM algorithm has been observed to be extremely slow in some applications.

2.3 Stochastic EM algorithm

In cases where it is difficult to compute the expectation in the EM algorithm or it converges slowly, an alternative is to use stochastic imputation for the Maximization step in EM. Stochastic EM algorithm [5] draws a sample of the augmented data from its conditional distribution to form a complete data with the observed and updates the estimate of the parameters based on the complete data. Given the same setup where $X = (Y, Z)$ is a complete set of data with Y observed and Z latent, the Stochastic

Expectatation Step:

$$\text{Sample } Z^{(t)} \sim f(Z|Y, \theta^{(t-1)})$$

where $\theta^{(t-1)}$ is the current estimate of θ. Let $X^{(t)} = (Y, Z^{(t)})$

Maximization Step:

$$\theta^{(t)} = \arg\max_{\theta} L(X^{(t)}|\theta) = \arg\max_{\theta} L(Y, Z^{(t)}|\theta)$$

which calculates the maximum likelihood estimate of θ based on $X^{(t)}$.

2.4 Maximum A Posterior (MAP) Estimation

In Bayesian statistics, a maximum a posteriori estimate is a mode of the posterior distribution, or

$$\hat{\theta} = \arg\max_{\theta} f(\theta|X) = \arg\max_{\theta} f(X|\theta)\pi(\theta)$$

as opposed to a maximum likelihood estimate of

$$\hat{\theta} = \arg\max_{\theta} f(X|\theta).$$

2.5 Collapsed Gibbs

Consider the fact that Gibbs sampling procedure can be computationally expensive if the model contains a reasonably large number of variables and many of them are irrelevant in the analysis, collapsed Gibbs can be used to approximate the marginal distribution of any subsets of the variables by integrating out the rest. The idea can be illustrated as follows. Suppose we are interested in approximating $f(\alpha, \beta, \theta)$, the usual approach would be to sample

1. $\alpha \sim f(\alpha|\beta, \theta)$
2. $\beta \sim f(\beta|\alpha, \theta)$,
3. $\theta \sim f(\theta|\alpha, \beta)$

for a number of iterations; however, suppose now that we can integrate out θ, we are left with $f(\alpha, \beta)$ which takes fewer steps to approximate and we can obtain θ later on by sampling from $\theta \sim f(\theta|\alpha, \beta)$.

Chapter 3

L-1 Penalized Univariate Ordinal Response Data Analysis

In this chapter, we will first review a popular approach to analyzing ordinal response data that is to introduce a latent continuous variable consider there exist an underlying order on the level of categories. The presence of covariates allows us to impose an ordinary regression model on the latent variable. By the presence of covariates, we consider the problem of variable selection where the response variable is ordinal. We will also introduce some alternatives for variable selection purposes and introduce our proposed method.

3.1 Analysis of Ordinal Response Data

Polychotomous response data arise often in social science applications where the methodology limitations in collecting data force the researchers to report grouped categorical results [14]. Therefore, the categories are thought to have an underlying order. Moreover, when the true "distance" between each successive level does not seem to be constant, it does not make sense to treat them as continuous. Given a sample of n observations on response variable $Y_{n\times 1}$, and p independent variables $X_1, X_2, ..., X_p$, the usual linear model assumes that these data satisfy

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + u_{n\times 1}, \quad u \sim N(0, \sigma^2 I)$$

However, in the case of Y being ordinal, some of the assumptions for the model might be violated. To overcome such difficulties, a popular approach is to introduce a latent continuous variable by the method of data augmentation [11]. In the context of analyzing ordinal response data, it is natural to assume an underlying continuous variable (Z) and breakpoints (γ) so that when the latent continuous variable (Z_i) falls in the interval ($[\gamma_{j-1}, \gamma_j]$) defined by the breakpoints, the observation (Y_i) is in the corresponding category (j).

$$Y_i \in R_j \iff \gamma_{j-1} < Z_i < \gamma_j \qquad 1 \le j \le K$$

Since Z is continuous, we can assume a normal regression structure on Z that is given by

$$Z_{n\times 1} = X_{n\times p}\beta_{p\times 1} + u_{n\times 1}, \quad u \sim N(0, \sigma^2 I)$$

So,

$$\begin{aligned} Y_i \in R_j &\iff \gamma_{j-1} < X_i\beta + u_i < \gamma_j \\ &\iff \frac{\gamma_{j-1} - X_i\beta}{\sigma} < \frac{u_i}{\sigma} < \frac{\gamma_j - X_i\beta}{\sigma} \end{aligned}$$

where γ_k are breakpoints for the ordinal levels $0 \le k \le K$ and $-\infty = \gamma_0 \le \gamma_1 \le \cdots \le \gamma_K = \infty$. Since $u \sim N(0, \sigma^2 I)$

$$Pr(Y_i \in R_j) = \Phi(\frac{\gamma_j - X_i\beta}{\sigma}) - \Phi(\frac{\gamma_{j-1} - X_i\beta}{\sigma})$$

where

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

For identifiability reason, σ is usually taken to be 1 and $\gamma_1 = 0$.

When introducing Z_i's and γ, the joint posterior distribution is given by

$$\pi(\beta, \gamma, Z|y) \propto \prod_{i=1}^{n} [\frac{1}{\sqrt{2\pi}} \exp\{-\frac{(Z_i - X_i\beta)^2}{2}\} \{\sum_{j=1}^{J} I(Y_i = j) I(\gamma_{j-1} < Z_i < \gamma_j)\}]$$

by assuming a diffuse prior for (β, γ). The complete data likelihood is hard to evaluate and sample from directly. However, it can be seen that

$$\beta|Z \sim N(\hat{\beta}, (X^T X)^{-1}))$$
$$Z_i|\beta, Y_i = j, \gamma \sim N(X_i\beta, 1) \text{ truncated at the left (right) by } \gamma_{j-1}(\gamma_j), \text{ and}$$
$$\gamma_j|Z, Y \sim \text{ Unif}[\max\{Z_i : Y_i = j\}, \min\{Z_i : Y_i = j+1\}].$$

3.2 Variable Selection

Variable selection is the process of selecting variables for the purpose of constructing statistical models to help us understand the relationships among variables. It has been one of the most popular and important topics in statistical modeling since the data collected usually contain redundant information that should be excluded. The problem has been examined from both frequentist and Bayesian perspectives and a large number of techniques have been proposed in the literatures. In this section, we will review some of the most commonly seen techniques.

3.2.1 Frequentist

The frequentist approach assumes that each parameter has a true (unique) value and that given sufficient information (data), we should be able to draw conclusions about the parameters so the techniques proposed are usually deterministic in nature.

Recall that the ordinary least square solution of a variable selection problem given by

$$Y \sim N(X\beta, \sigma^2)$$

is obtained by solving

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2.$$

However, not every predictor should be included in the model; therefore, variable selection is needed. LASSO, SCAD and MCP are some of the most commonly seen

variable selection techniques that solve the problem of the form

$$\arg\min_{\beta} |Y - X\beta|^2 + p_\lambda(\beta) \tag{3.1}$$

with different penalty functions of the form $p_\lambda(\beta)$.

LASSO Least Absolute Shrinkage and Selection Operator [16] places a constraint on the sum of the absolute value of the coefficients in a regression problem, which can be written as

$$\arg\min_{\beta} |Y - X\beta|^2 \quad \text{subject to } |\beta| \leq t$$

where t is a tuning parameter. In terms of β,

$$\begin{aligned}\arg\min_{\beta} |Y - X\beta|^2 &= \arg\min_{\beta} \beta^T X^T X\beta - 2\beta X^T Y + constant \\ &= \arg\min_{\beta} \beta^T X^T X\beta - 2\beta X^T X\hat{\beta}_{ols} = \arg\min_{\beta}(\beta - \hat{\beta}_{ols})^T X^T X(\beta - \hat{\beta}_{ols})\end{aligned}$$

The criterion $|Y - X\beta|^2$ for β is equivalent as $(\beta - \hat{\beta}_{ols})^T X^T X(\beta - \hat{\beta}_{ols})$ where $\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y$ is the lease square estimate of β. The possible solutions occur at the intersections of the elliptical contours and the constraint and the lasso solution occur at the first point of intersection. The constraint shrinks the estimate of the coefficients from their least square estimates to 0 as t increases so it produces sparse solutions that achieves variable selection automatically. Adding the constraint is equivalent as placing a Lagrangian penalty to the residual sum of squares, with λ depending on t. The problem is equivalent as

$$\arg\min_{\beta} |Y - X\beta|^2 + \lambda|\beta|.$$

This problem has solutions for any given λ, so we need to fix λ to obtain estimate for the coefficient β.

MCP MCP [17] also solves the problem of the form 3.1 with

$$p_\lambda(\beta) = \lambda \int_0^t (1 - \frac{\beta}{\gamma\lambda})_+ d\beta, \qquad \gamma > 0.$$

or

$$p_\lambda(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2\gamma} & \text{if } \beta \leq \gamma\lambda \\ \frac{\lambda^2\gamma}{2} & \text{if } \beta > \gamma\lambda \end{cases}$$

which can be easier to understand by looking at the first derivative

$$p'_\lambda(\beta) = \begin{cases} \lambda - \frac{\beta}{\gamma} & \text{if } \beta \leq \gamma\lambda \\ 0 & \text{if } \beta > \gamma\lambda \end{cases}$$

SCAD Smoothly Clipped Absolute Deviation Penalty [8] is defined by

$$p_\lambda(\beta) = \lambda \int_0^t \min\{1, \frac{(\gamma\lambda - \beta)_+}{(\gamma - 1)\lambda}\} d\beta, \qquad \gamma > 2$$

or

$$p'_\lambda(\beta) = \lambda\{I(\beta \leq \lambda) + \frac{(\gamma\lambda - \beta)_+}{(\gamma - 1)\lambda} I(\beta > \lambda)\}, \qquad \gamma > 2$$

Both MCP and SCAD have rate of penalty shrinks to 0 as the size of the coefficient increases [3].

3.2.2 Bayesian

Bayesian models are specified by distinctive prior distributions where the prior acts as a penalty for models with a smaller number of observations. Bayesian variable selection usually involves introduction of a latent indicator variable for the inclusion of the predictors. In this section, we briefly review two of the most popular choices for prior distributions of the coefficient β.

Spike and slab [11] uses a Gibbs sampling technique called "stochastic search variable selection" that introduces a latent indicator variable α and places a two-component normal mixture prior on the coefficient β defined by

$$\beta_j | \alpha_j \sim (1 - \alpha_j) N(0, \tau_j^2) + \alpha_j N(0, c_j^2 \tau_j^2)$$

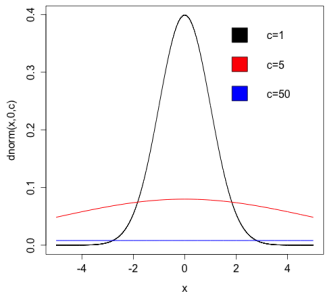
with $p(\alpha_j) = p_j^{\alpha_j}(1 - p_j)^{1-\alpha_j}$ or $\alpha_j \overset{iid}{\sim} \text{Bernoulli}(p_j)$. In matrix form

$$\beta | \alpha \sim N_p(0, D_\alpha R D_\alpha)$$

where $\alpha = (\alpha_1, \ldots, \alpha_p)$, R is the prior correlation matrix for β, and

$$D \equiv \text{diag}[a_1\tau_1, \ldots, a_p\tau_p]$$

with a_i=1 if α_i=0 and $a_i = c_i$ if α_i=1. [11] argues that since the densities of N(0,τ_i^2) and N(0,$c_i^2\tau_i^2$) intersect at $\xi(c_i) = \sqrt{2(\log c_i)c_i^2/(c_i^2-1)}\tau_i$ and c_i is the ratio of the heights of N(0,τ_i^2) and N(0,$c_i^2\tau_i^2$) at 0 so c_i can be interpreted as the prior odds that x_i should be excluded when β_i is very close to 0.



If we consider each β_i separately, since

$$\hat{\beta}_i|\sigma_{\beta_i}, \gamma_i = 0 \sim N(0, \sigma_{\beta_i}^2 + \tau_i^2)$$
$$\hat{\beta}_i|\sigma_{\beta_i}, \gamma_i = 1 \sim N(0, \sigma_{\beta_i}^2 + c_i^2\tau_i^2)$$

Let $t_i\sigma_{\beta_i}$ denote the intersection points of these distributions where σ_{β_i} is the variance of the least square estimate $\hat{\beta}_i$. Then

$$P(\gamma_i = 1|\hat{\beta}_i, \sigma_{\beta_i}) > p_i \quad \text{iff} \quad \hat{\beta}_i/\sigma_{\beta_i} > t_i$$

so the point t_i can be thought of as the threshold at which the t statistics corresponds to an increased marginal probability that X_i should be included in the model. In practice, the performance of this method is highly sensitive to the choice of c and τ where we must be able to obtain the standard errors for the least square estimates before we can determine what to use for them.

Non-informative [12] employed a hierarchical Bayesian model that includes latent variables for analyzing dichotomous ordinal response variable. The prior distribution for the coefficient β is defined as

$$\beta_\alpha|\alpha \sim N(0, c(X_\alpha^T X_\alpha)^{-1})$$

where c is a positive scale factor that need to be pre-specified and X_α the columns of X corresponding to those α's that are nonzero.

3.3 Penalized Ordinal Regression Algorithm

Here, we describe our proposed method and make comparisons with univariate Bayesian variable selection proposed by [12] for binary classification problems.

Ordinal Regression

We use the procedure proposed in [1] for ordinal regression, which can be summarized as follows.

- Initialization
  - Set $\gamma^{(0)}$
  - Draw $Z|Y, \gamma$ from Unif$[\gamma_{j-1}, \gamma_j]$
  - Set $\beta^{(0)} = (X^T X)^{-1} X^T Z$
- At the k^{th} iteration

  Expectation Step:

  1. Sample γ from $\gamma_j|Y, Z \sim$ Unif$[\max\{Z_i : Y_i = j\}, \min\{Z_i : Y_i = j+1\}]$
  2. Sample Z from $Z_i|Y_i = j, \beta, \gamma \sim \mathrm{N}(X\beta, 1)$ truncated at the left (right) by γ_{j-1} (γ_j)

  Maximization Step:

  - Update β by setting $\beta \sim \arg\max_\beta p(\beta|Z)$

  We infuse LASSO in the β sampling step to achieve automatic variable selection by letting

  $$\beta = \arg\min_\beta |Y - X\beta|^2 + \lambda|\beta|$$

  and select appropriate λ based on the number of predictors we want to retain by using an algorithm that stops after certain number of steps. Other variable selection techniques such as SCAD, and MCP can also be used here to achieve automatic variable selection.

3.4 Simulation Studies

To test the performance of our proposed method, we consider 4 design structures for generating independent variable X. The latent (response) variable Z is distributed $N_n(\eta, \sigma^2 I)$ where η is a linear combination of X as in [6, 15]. Here, we only consider one model

$$\eta = 3X_1 + 1.5X_2 + 2X_5$$

and $\sigma = 1.5$. Once we have Z, Y is obtained by partitioning Z into 5 levels based on their percentiles so each level has approximately the same number of observations, which contrasts with real data sets where the number of observations are not uniform in each level. We use fixed number of observations ($n = 50$), predictors ($p = 20$), number of iterations at 100, Gibbs sampling cycles within each iteration at 400 and burnins at 200. For each design, we show the selection probabilities of the predictors under 3 different variable selection techniques and compare them against the result obtained by Bayesian variable selection with β's prior distribution given by $\beta|\alpha \sim N(0, c(X_\alpha^T X_\alpha)^{-1})$ where c approaches infinity and $P(\alpha_j = 1) =$ maximum steps$/p$ for all j. The rows show results when the maximum steps (number of variables to retain) used when fitting the models are at 5 and 10, respectively. The selection probabilities presented are in the order of the variables for all of the design structures for ease of comparison and in consideration of the effects of design structures.

- Design 1

  $$X_i \overset{iid}{\sim} \mathrm{N}_n(0, I_n),\ i = 1, \ldots, p$$

  We consider the simplest design for selecting variables where all the predictors are independently distributed.

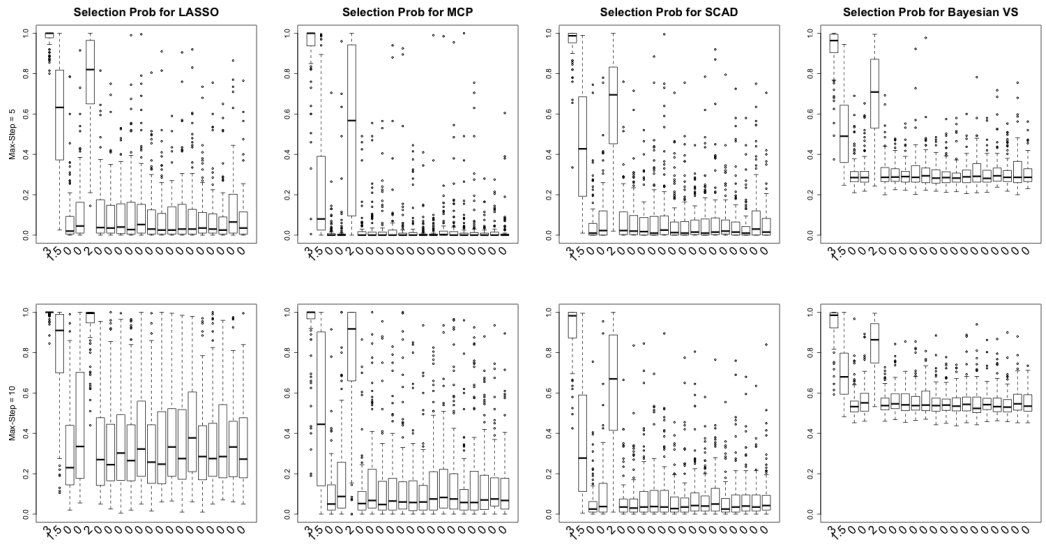


Figure 3.1: Comparing the selection probabilities for LASSO, MCP, SCAD against Bayesian variable selection under Design 1

- Design 2

$X_i \overset{iid}{\sim} \mathrm{N}_n(0, I_n), i = 1, \ldots, 10$ and $X_i \overset{iid}{\sim} \mathrm{N}_n(0.5X_1 + X_2 + 1.5X_3, I_n),\ i = 11, \ldots, 15$

We consider a design for a harder variable selection problem where there are five predictors that are correlated with the first three.

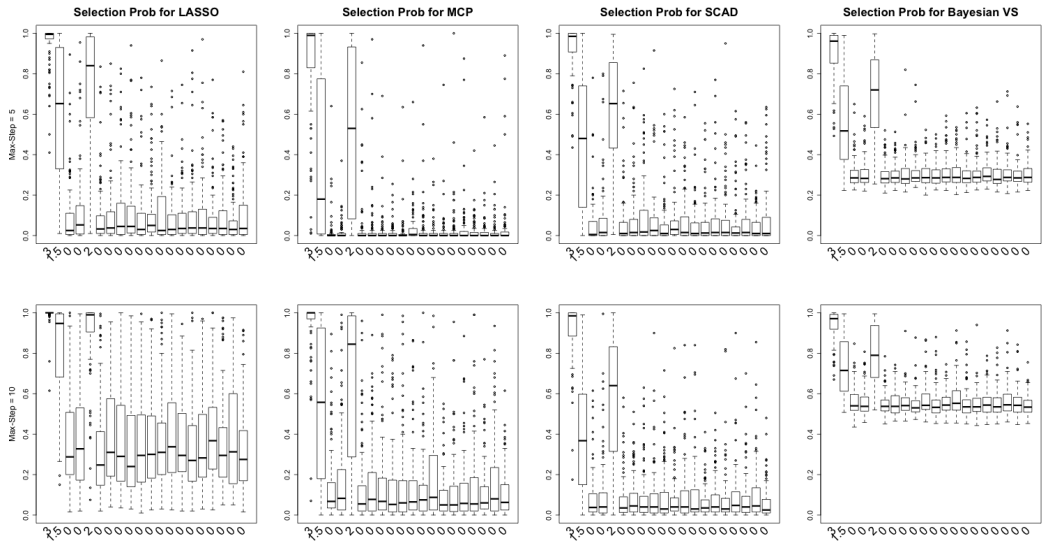


Figure 3.2: Comparing the selection probabilities for LASSO, MCP, SCAD against Bayesian variable selection under Design 2

- Design 3

$$X_i \sim \mathrm{N}_n(0, I_n),\ i = 1, \ldots, p \text{ and } \rho(X_i, X_j) = 0.7 \text{ for all } i \neq j$$

Again, we consider a design for a hard variable selection problem where there are strong correlations between each pair of predictors.

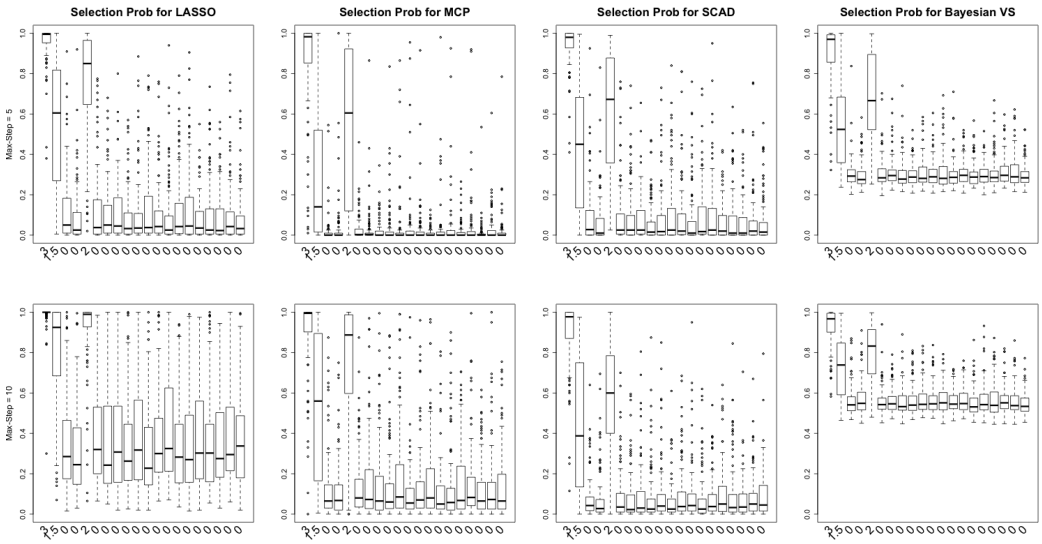


Figure 3.3: Comparing the selection probabilities for LASSO, MCP, SCAD against Bayesian variable selection under Design 3

- Design 4

$$X_i \sim \mathrm{N}_n(0, I_n),\ i = 1, \ldots, p \text{ and } \rho(X_i, X_j) = 0.5^{|i-j|}$$

We consider a different design structure that results in correlation among predictors in a different way.

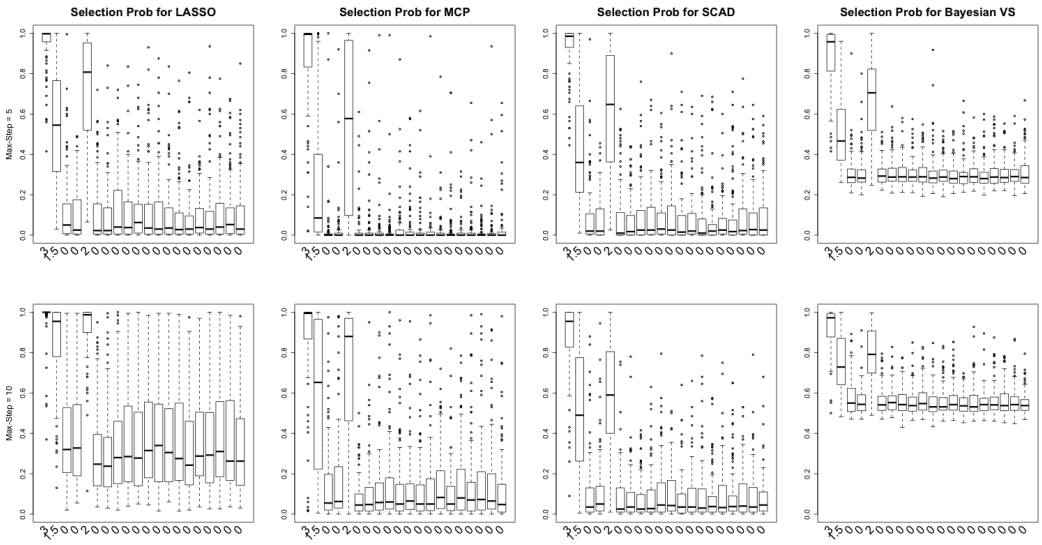


Figure 3.4: Comparing the selection probabilities for LASSO, MCP, SCAD against Bayesian variable selection under Design 4

For variable selection purpose, LASSO is pretty good at selecting the true variables with high probabilities and Bayesian variable selection can at the same time picks the right variables but it also has higher selection probabilities for noises compared to LASSO. We see that LASSO is successful for selecting variables when the response is ordinal under the augmented framework. Moreover, comparing to Bayesian variable selection where we have to update α one at a time, LASSO is more computationally efficient. Although the estimates of coefficients under LASSO might be biased, for variable selection purpose, as long as the true variables are selected, we can always obtain the least square solution using only the variables selected.

3.5 Data Analysis

We applied the proposed method on two different data sets, one is consumer preference data taken from the MovieLens project and the other one is a survey data. By applying the method to data that are different in nature, we show that our method is easily applicable to different types of data to obtain the ranking of selection probabilities.

3.5.1 MovieLens

The data set (http://movielens.org) consists of 100,000 movie ratings from users, the genres of the movies in the form of indicator vectors as well as simple demographic information of the users such as gender, age and occupation. In the data set, each user included has at least 20 movie ratings. We use the genres as predictors to construct linear models on predicting the ratings. We treat each genre independently as if they are different covariates and do variable selection to find out the most and least favoured genres by groups of individuals. The results are presented by groups of individuals and individuals within the same group are not differentiated. The bar-plots below show the selection probabilities and direction for frequently selected covariates (genres) for chosen groups of users with similar demographic information. We only include up to two-way interactions since it makes it easier to compare the results between different groups of individuals.

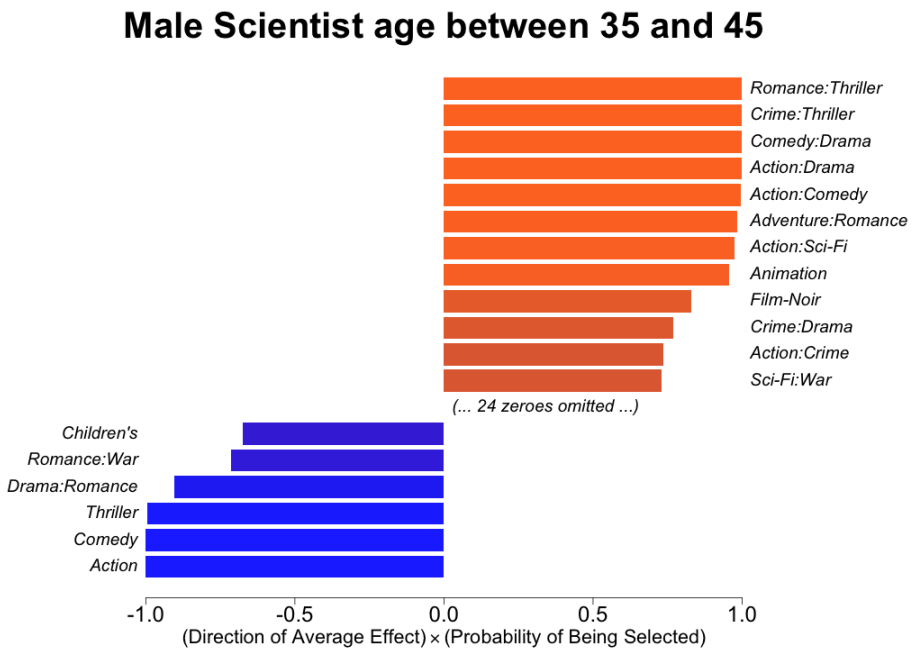


Figure 3.5: Selection probability by the direction of average effect of top ranked genres for male scientists age between 35 and 45

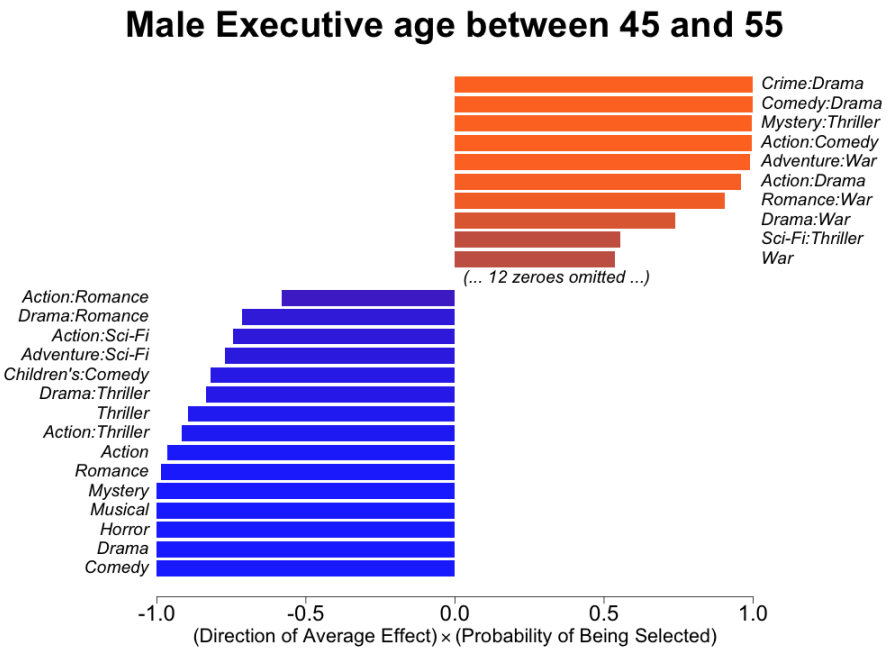


Figure 3.6: Selection probability by the direction of average effect of top ranked genres for male executives age between 45 and 55

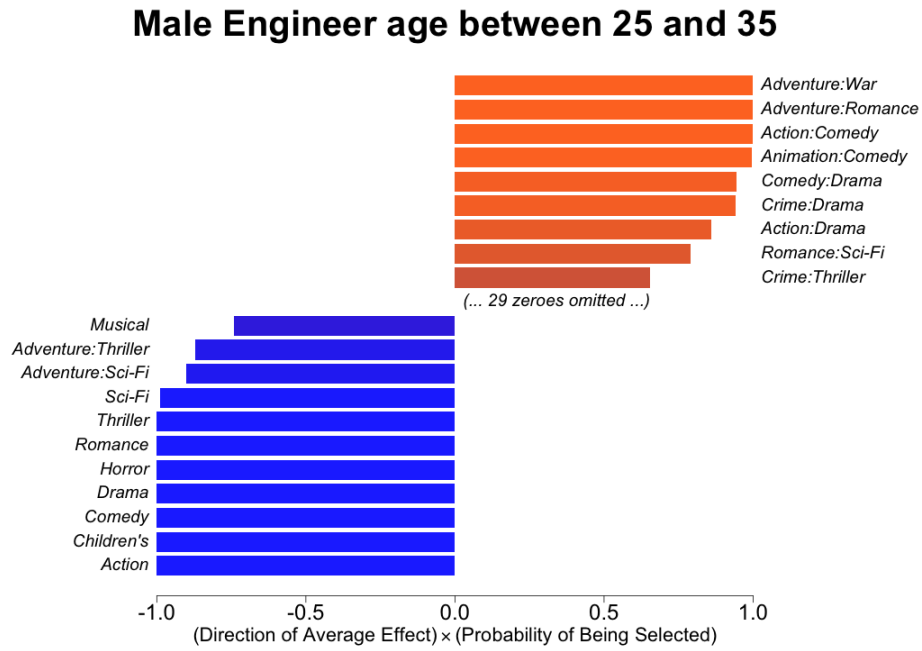


Figure 3.7: Selection probability by the direction of average effect of top ranked genres for male engineers age between 25 and 35

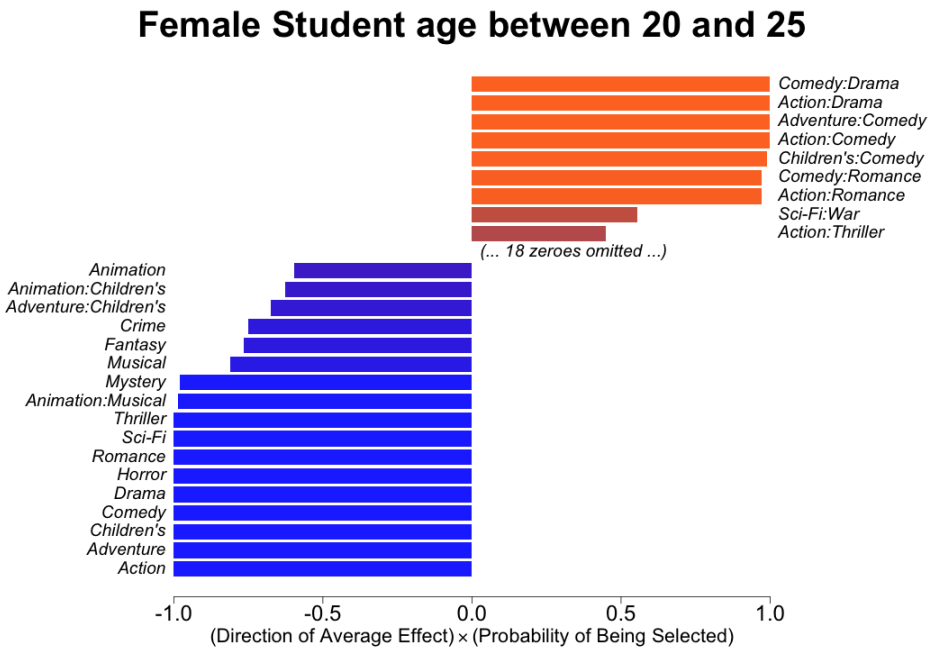


Figure 3.8: Selection probability by the direction of average effect of top ranked genres for female students age between 20 and 25

It is interesting to see that some genres are not favored by itself but receive better feedback when combined with other ones.

3.5.2 World Values Survey

World Values Survey is a global research project (accessed at www.worldvaluessurvey.org) that explores people's beliefs and values. We made use of the 2005-2006 Wave data which consists of 67,268 incomplete responses across 112 countries on 260 questions. We selected a subset of the questions for our analysis, picked "life satisfaction" as our response variable and analysed within countries that have complete data for the selected questions. Most of the questions are in the forms of "multiple choice" or "rating" and the orders of the answers are inconsistent between different questions so we transformed the data to ensure consistency for the ease of analysis and treat the predictors as continuous variables.

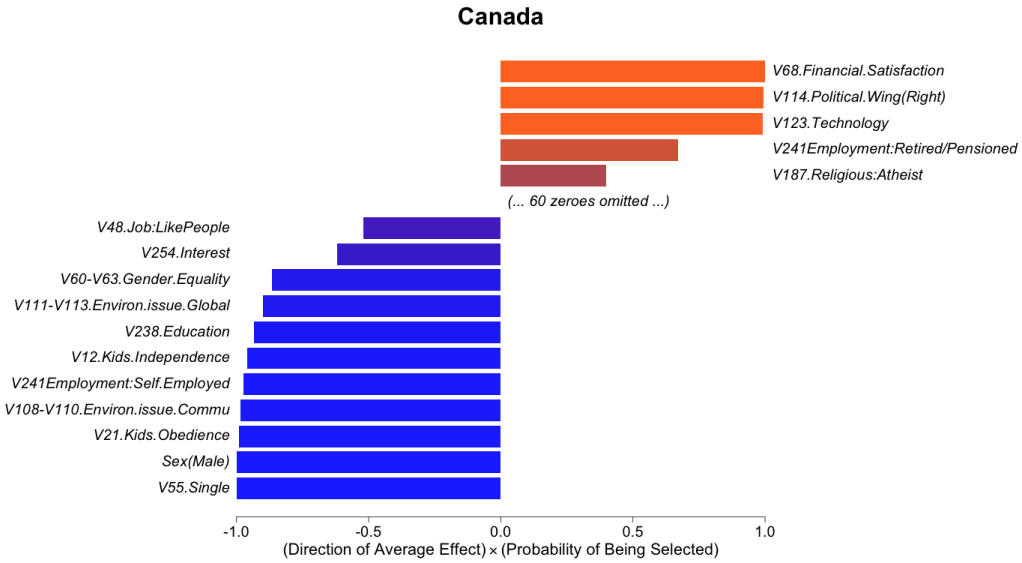


Figure 3.9: Selection probability by the direction of average effect of top ranked factors for life satisfaction in Canada

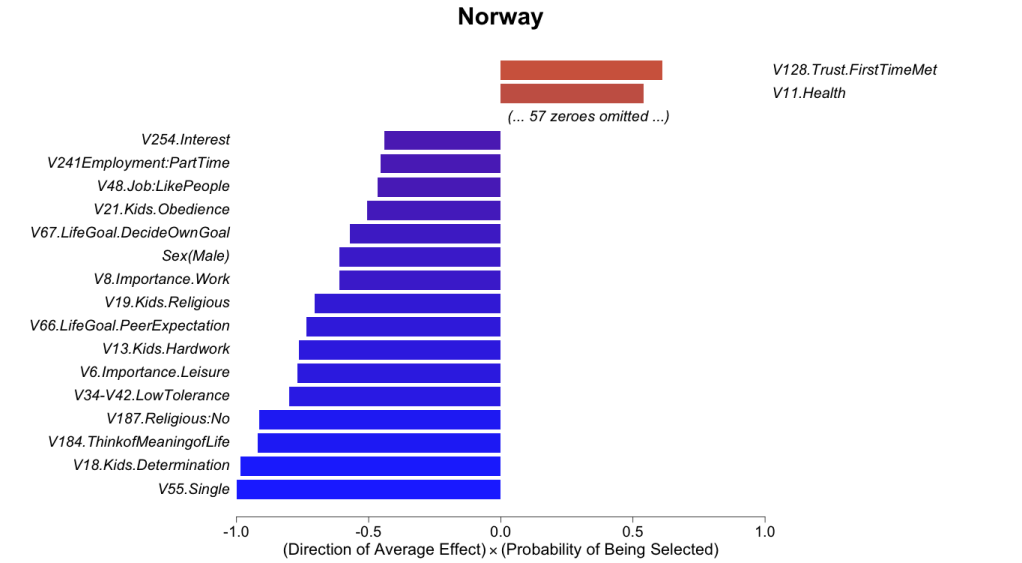


Figure 3.10: Selection probability by the direction of average effect of top ranked factors for life satisfaction in Norway

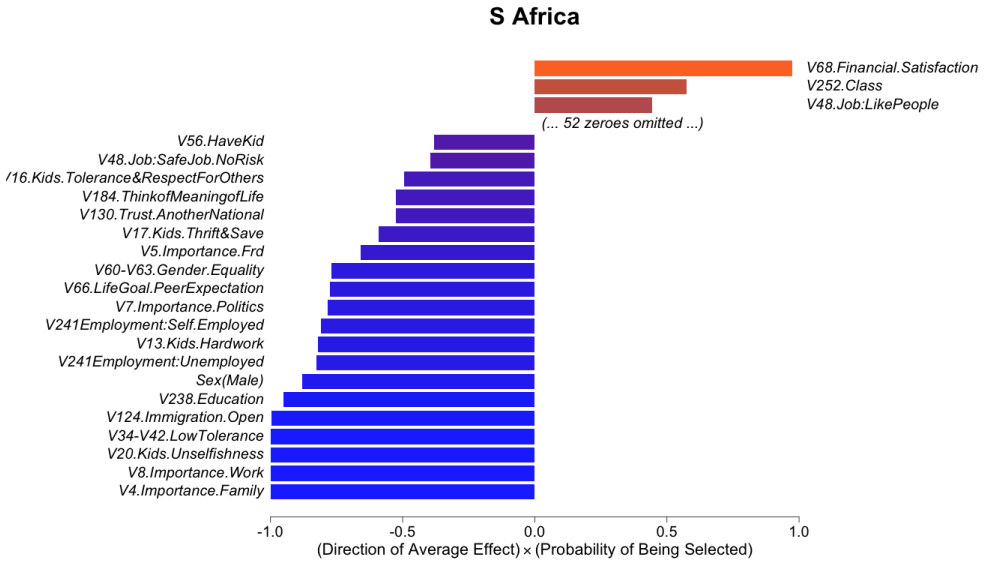


Figure 3.11: Selection probability by the direction of average effect of top ranked factors for life satisfaction in South Africa

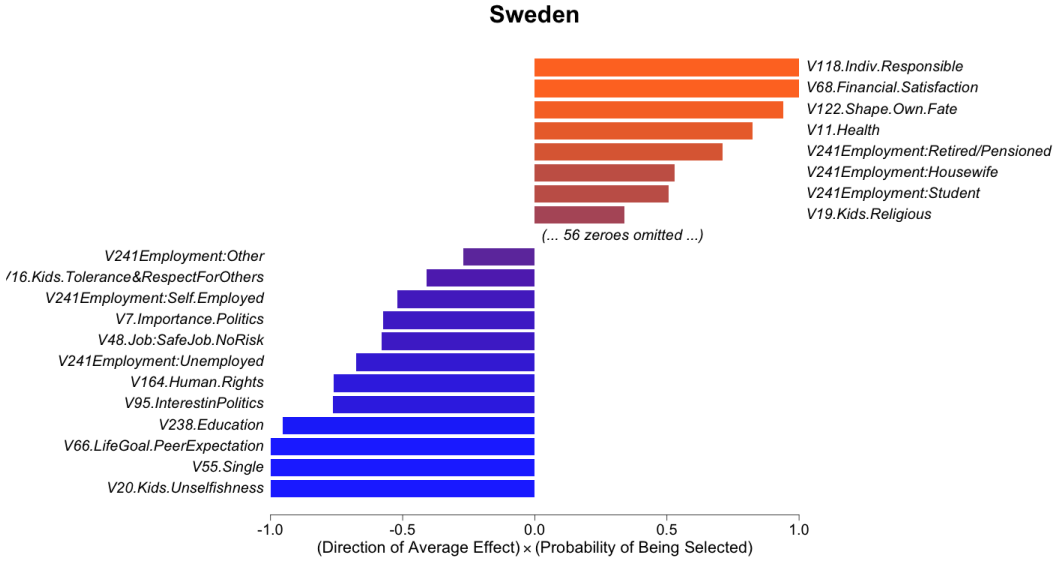


Figure 3.12: Selection probability by the direction of average effect of top ranked factors for life satisfaction in Sweden

It can be seen that there are universal factors like financial satisfaction, health, being a student or retired, believing that wealth is enough for all, and having control over own life that contribute the most to a satisfied life. On the other hand, the universal factors that contribute to least satisfied life include thinking about the

meaning of life, being single, and having life goals as pleasing parents or meeting others' expectations. Most interestingly, there exist cross-country variations in the key factors that affect people's life satisfaction. In Canada, educated individuals and people who think women are as good as men are less satisfied whereas right wing people and atheist are. In Norway, those who trust people easily are satisfied with their lives. In South Africa, people in upper classes tend to be more satisfied. In Sweden, people who believe they can shape their own fate or they are the one who is responsible for their own lives are more satisfied with their lives. The highly educated individuals and those who are interested in politics are less. These results are interesting since they coincide with our understand of the countries and also show the hidden problems in the country.

3.6 Summary and Remarks

- When doing ordinal regression, if we are only interested in selecting variables, we can modify the Gibbs Sampling algorithm to a stochastic EM algorithm so less sampling steps are needed. Moreover, when a variable selection technique such as LASSO, SCAD, or MCP is used, we can rank the covariates in their order of significance based on the selection probabilities calculated as $\mathbb{E}[I(\hat{\beta}_{lasso} \neq 0)]$. For the maximization step, conjugate priors can to be used for closed form solutions.

- Under the Bayesian framework, it is convenient to choose priors based on the problems of interest and Gibbs sampler can be used to estimate the posterior distributions.

Chapter 4

Multiple Response Data Analysis

In this chapter, we consider an extension of Bayesian variable selection to handle multiple response data. We will first review multiple response Bayesian variable selection and discuss the drawbacks of the technique to make improvement upon that. We make comparisons with several techniques and explain that our proposed method is comparable to one and show some simulation results to compare their performances.

4.1 Introduction

In analyzing survey data or biological data, people are usually interested in more than one attribute. However, the vast majority of work in multivariate variable selection focuses on selecting a common set of variables for all the responses, which can be too restrictive in that the sets of variables with the greatest influence on the responses might be different for each response. However, considering there are usually "connections" between the attributes of interest, it also does not make sense to consider each response separately [13]. Therefore, we attempt to select different sets of variables for different responses while allowing the process of selecting variables for each response to communicate with one another through the infusion of relationship information into the prior distribution. In this section, we will first review some proposed variable selection techniques, propose a method and show

that our results have some desirable properties and demonstrate the performance of our proposed method by empirical evidences.

4.2 Multiple Response Bayesian Variable Selection

Given multiple response data $Y_{n\times q}$ and a set of independent variables $X_1, \ldots, X_p$, a typical multivariate variable selection problem can be defined as finding the best model of the form

$$Y_{n\times q} = X_1^* B_1^* + \cdots + X_r^* B_r^* + E_{n\times q}$$

where $X_1^*, \ldots, X_r^*$ is a selected subset of $X_1, \ldots, X_p$. In [4], a latent indicator variable is introduced for the inclusion of the p independent variables and spike and slab priors [4] are used for the coefficient B, which can be seen as a generalization of [11] that focused on univariate regression. In summary, [4] calculated the posterior distribution of the latent indicator variable α which is given by

$$\pi(\alpha|Y,X) \propto \pi(\alpha) \int\int f(Y|X,B,\Sigma)\pi(B|\Sigma,\alpha)\pi(\Sigma)dBd\Sigma$$

The integrals can be evaluated if conjugate priors are used, the resulting posterior is then

$$\pi(\alpha|Y,X) \propto (|H_\alpha||K_\alpha|)^{-q/2}|Q_\alpha|^{-(n+\delta+q-1)/2}\pi(\alpha)$$

where

$$\begin{aligned} H_\alpha &= D_\alpha R_\alpha D_\alpha \quad \text{as in [11]} \\ K_\alpha &= X^T X + H_\alpha^{-1} \\ Q_\alpha &= Q + Y^T Y - Y^T X K_\alpha^{-1} X^T Y \end{aligned}$$

under the prior specifications given by

$$\begin{aligned} Y &\sim \mathcal{MN}(XB, I_n, \Sigma) \\ B &\sim \mathcal{MN}(0, H_\alpha, \Sigma) \\ \Sigma &\sim \mathcal{IW}(\delta; Q) \end{aligned}$$

where $\mathcal{MN}_{n,p}(M,U,V)$ stands for a matrix normal variate with mean M, row covariance matrix U, column covariance matrix V, and probability density of

$$f(X) = \frac{1}{(2\pi)^{np/2}|V|^{n/2}|U|^{p/2}} exp\{-\frac{1}{2}tr[V^{-1}(X-M)^T U^{-1}(X-M)]\}$$

This posterior distribution is claimed to enclose the information on the effectiveness of the predictors in explaining the response Y. For the ease of calculation, a form is developed for fast updating as

$$\begin{aligned} g(\alpha) &= \pi(\alpha|Y,X) && (4.1)\\ &\propto (|H_\alpha||K_\alpha|)^{-q/2}|Q_\alpha|^{-(n+\delta+q-1)/2}\pi(\alpha) && (4.2)\\ &\propto (|\tilde{X}^T\tilde{X}|)^{-q/2}|Q_\alpha|^{-(n+\delta+q-1)/2}\pi(\alpha) && (4.3)\\ &\propto (|\tilde{X}^T\tilde{X}|)^{-q/2}|Q+\tilde{Y}^T\tilde{Y}-\tilde{Y}^T\tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}|^{-(n+\delta+q-1)/2}\pi(\alpha) && (4.4) \end{aligned}$$

where

$$\tilde{X} = \begin{pmatrix} XH_\alpha^{1/2} \\ I_p \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$$

Now the posterior distribution is in closed form and can be evaluated easily for any given α. Furthermore, by setting $H_\alpha = c(X_\alpha^T X_\alpha)^{-1}$, Q_α becomes

$$\begin{aligned} Q_\alpha &= Q + Y^TY - Y^TX_\alpha(X_\alpha^TX + (1/c)X_\alpha^TX_\alpha)^{-1}X_\alpha^TY \\ &= Q + \frac{1}{c+1}Y^TY + \frac{c}{c+1}(Y^TY - Y^TX_\alpha(X_\alpha^TX_\alpha)^{-1}X_\alpha^TY) \end{aligned}$$

and $|H_\alpha K_\alpha|$ simplifies to

$$\begin{aligned} |H_\alpha K_\alpha| &= |c(X_\alpha^TX_\alpha)^{-1}(X_\alpha^TX_\alpha + H_\alpha^{-1})| \\ &= |(c+1)I_{p_\alpha}| \\ &= (c+1)^{p_\alpha} \end{aligned}$$

However, the space of α is of dimension $\{0,1\}^p$, for p sufficiently large, the computational cost makes evaluating the posterior distribution at every possible α infeasible; therefore, MCMC is adopted. When p is large (>25), MCMC becomes necessary to explore the posterior distribution of α by sampling α_j one by one from its full conditional distribution, that is $p(\alpha_j = 1, \alpha_{-j}|Y,X) = \theta_j/(\theta_j+1)$ with

$$\theta_j = \frac{g(\alpha_j = 1, \alpha_{-j}|Y,X)}{g(\alpha_j = 0, \alpha_{-j}|Y,X)}$$

and $g(\alpha)$ specified previously in 4.1.

4.3 Bayesian Variable Selection for Linear Models with Multiple Responses

Now, we wish to consider a problem of the form

$$Y_{n\times q} = X_{n\times p}B_{p\times q} + E_{n\times q}$$

where the coefficient matrix B is sparse, which makes it easier to consider q different problems given by

$$Y^{(i)}_{n\times 1} = X_{n\times p}B^{(i)}_{p\times 1} + E^{(i)}_{n\times 1}.$$

with $E^{(i)}_{n\times 1} \sim \mathcal{N}(0, \sigma_i^2 I_n)$ for all i and

$$\sigma_i^2 \sim \mathcal{IG}(\nu, \delta).$$

We are interested in selecting $B^{(i)}_{p\times 1}$'s so that the coefficients being non-zero constitutes a significant corresponding variable for response i. In order to do this, we made use of the popular approach in Bayesian variable selection problems that is to introduce a (latent) indicator variable α for each of the independent variables (j) for each response (i) where $\alpha_{i,j}$ follows a Bernoulli distribution with parameter $p_{i,j}$.

$$\alpha_{i,j} \sim \text{Bernoulli}(p_{i,j})$$

and given α,

$$\beta^{(i)}|\alpha_i \sim N(0, c(X^T_{\alpha_i}X_{\alpha_i})^{-1}) \qquad \text{for } i = 1, \ldots, q$$

where α_i is an indicator vector of length p and X_{α_i} is the columns of X corresponding to those α_i's that are nonzero. For univariate variable selection problems, the variable selection process for each response variable is considered separately as if they are independent. However, this disregards the most important feature in a multiple response problem - the relationship between the response variables.

Given that it is convenient to put a prior on any parameter under the Bayesian framework, we make use of the advantage to incorporate the relationship information into the conditional prior distribution of $p_{i,j}$. Moreover, Gibbs sampler can be used directly under the conditional specification even though the joint distribution is not known. A more thorough discussion can be found in [2].

Consider

$$\begin{aligned} Cor(Y^{(j)}, Y^{(k)}) &= \frac{Cov(XB^{(j)} + E^{(j)}, XB^{(k)} + E^{(k)})}{\sqrt{Var(XB^{(j)} + E^{(j)})Var(XB^{(k)} + E^{(k)})}} \\ &= \frac{Cov(XB^{(j)}, XB^{(k)})}{\sqrt{Var(XB^{(j)})Var(XB^{(k)})}} \end{aligned}$$

$$Cov(XB^{(j)}, XB^{(k)}) = \sum_i B_i^{(j)} B_i^{(k)} Var(X_i) + \sum_{i<l} 2B_i^{(j)} B_l^{(k)} Cov(X_i, X_l)$$

by first assuming X's are random variables. It can be seen that the variances, hence the correlations of the response variables depend on the variance of the data (covariate) matrix, the coefficients B, and most importantly, through the sharing of common predictors. As an example, suppose X_i's are independently distributed with $Var(X_i) = 1$ for all i and

$B^T =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Y is generated by $Y^{(i)} \overset{iid}{\sim} \mathrm{N}_n(XB^{(i)}, 2.5^2 I)$, $i = 1, 2$. This is the first model under Design 1 of our simulations. We will show some results later on in the section. The correlation can be calculated as

$$\begin{aligned} Cor(Y^{(1)}, Y^{(2)}) &= \frac{Cov(XB^{(1)}, XB^{(2)})}{\sqrt{Var(Y_1)Var(Y_2)}} \\ &= \frac{Cov(X_1 + X_2 + X_3 + X_4 + X_6, X_1 + X_5 + X_9 + X_{13} + X_{17})}{\sqrt{Var(Y_1)Var(Y_2)}} \\ &= \frac{Var(X_1) + Cov(X_2 + X_3 + X_4 + X_6, X_5 + X_9 + X_{13} + X_{17})}{\sqrt{Var(Y_1)Var(Y_2)}} = \frac{1}{5}. \end{aligned}$$

The correlation between the two response variables is based on the sharing of common predictors. With this in mind, we consider the problem of variable selection

where we take the correlations between the response variables into account when selecting variables given other sets have already been selected for other response variables by placing a conditional prior on $p_{i,j}$ that depends on $\alpha_{l,j}$ and $\rho_{i,l}$ for all other responses (l). Therefore, α is q by p. The prior is defined as follows,

$$p(\alpha_{i,j}|p_{i,j}) = p_{i,j}^{\alpha_{i,j}}(1-p_{i,j})^{(1-\alpha_{i,j})}$$

with

$$\pi(p_{i,j}|\alpha^c, R) \sim \text{Beta}(a, b)$$

where a and b are functions of α^c (α except for $\alpha_{i,j}$), more specifically, $\alpha_{-i,j}$, and $R \equiv \{\rho\}_{i,j}$. We chose Beta distribution for its properties

1. it's the conjugate prior for Bernoulli, and
2. it's parametrized by two positive shape parameters which can be functions of our parameters of interest (α and ρ or R)

Then,

$$\begin{aligned}
p(\alpha_{i,j}|\alpha_{-i,j}) &= \int p(\alpha_{i,j}|p_{i,j})\pi(p_{i,j}|\alpha_{-i,j}, R)dp_{i,j} \\
&= \int \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p_{i,j}^{a-1}(1-p_{i,j})^{b-1} p_{i,j}^{\alpha_{i,j}}(1-p_{i,j})^{(1-\alpha_{i,j})}dp_{i,j} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int p_{i,j}^{a-1+\alpha_{i,j}}(1-p_{i,j})^{b-1+1-\alpha_{i,j}}dp_{i,j} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+\alpha_{i,j})\Gamma(b+1-\alpha_{i,j})}{\Gamma(a+b+1)}
\end{aligned}$$

So,

$$\begin{aligned}
p(\alpha_{i,j} = 1|\alpha_{-i,j}, R) &= \frac{a}{a+b}, \\
p(\alpha_{i,j} = 0|\alpha_{-i,j}, R) &= \frac{b}{a+b}
\end{aligned}$$

specifies the posterior distribution of $\alpha_{i,j}$. Since we would like

$$p(\alpha_{i,j} = 1|\alpha_{-i,j}, R) = \frac{a}{a+b} = \begin{cases} \text{large} & \text{if } |\rho_{i,l}| \approx 1 \text{ and } \alpha_{l,j} = 1 \\ \pi_j & \text{if } \rho_{i,l} \approx 0 \\ \text{small} & \text{if } |\rho_{i,l}| \approx 1 \text{ and } \alpha_{l,j} = 0 \end{cases}$$

the selection probability for the covariate be high if the two response variables are highly correlated and the covariate is selected for the other response variable and vice versa (with a and b depending on on i and j). For two nearly independent response variables, the selection probabilities of any covariates for them depend only on a pre-specified probability. Therefore, we set

$$a_{i,j} = \pi_j k^{(2\alpha_{l,j}-1)|\rho_{i,l}|},\ b_{i,j} = 1 - \pi_j k^{(2\alpha_{l,j}-1)|\rho_{i,l}|}$$

for some predetermined π_j and tuning parameter $1 < k < \frac{1}{\pi_j}$. In the case where there are more than 2 response variables, we use the average effect given by

$$a_{i,j} = \pi_j \frac{1}{q-1} \sum_{l \neq i} k^{(2\alpha_{l,j}-1)|\rho_{i,l}|},\ b_{i,j} = 1 - \pi_j \frac{1}{q-1} \sum_{l \neq i} k^{(2\alpha_{l,j}-1)|\rho_{i,l}|}$$

instead.

4.3.1 The Algorithm

For convenience, we ignore i that indicates which response we are specifying here. Each step in this algorithm is repeated q number of times (one for each response i). Moreover, except for α_i which conditioned on α^c, $\beta^{(i)}$ and σ_i^2 depend only on α_j for $j = i$. The superscripts (k) indicates the *iteration*.

- Initialization
  - Set $\alpha^{(0)}$ from Bernoulli$(\pi_{i,j})$ $i = 1, \ldots, q,\ j = 1, \ldots, p$
  - Set $\beta^{(0)}$ and $\sigma^{2(0)}$ given $\alpha^{(0)}$ to the maximum likelihood estimate of β given $\alpha^{(0)}$ and σ^2 given $\alpha^{(0)}$ and $\beta^{(0)}$.
- At the k^{th} iteration

  Expectation Step:

  We incorporate the idea of integrating out (collapsing down) irrelevant parameters by placing conjugate prior on β to obtain the posterior distribution of α.

  Draw $\alpha^{(k)}$ from $p(\alpha|Y, \sigma^2, \alpha^c)$

  $$p(\alpha_{i,j}|Y, \sigma^2, \alpha^c) \propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2} - (\frac{1}{\sigma^2} + \frac{1}{c})^{-1} \frac{Y^TX_\alpha(X_\alpha^TX_\alpha)^{-1}X_\alpha^TY}{\sigma^4}]\} p_{i,j}^{*\alpha_{i,j}} (1 - p_{i,j}^*)^{(1-\alpha_{i,j})}$$

where

$$p^*_{i,j} = \frac{a_{i,j}}{a_{i,j} + b_{i,j}}$$

with $a_{i,j}$ and $b_{i,j}$ specified in the previous section. Here, one sample of $\alpha_{i,j}$ is drawn one by one from Bernoulli($\frac{A}{A+1}$) for each i and each j by keeping α^c fixed.

$$A = \frac{p(\alpha_{i,j} = 1|Y, \sigma^2, \alpha^c)}{p(\alpha_{i,j} = 0|Y, \sigma^2, \alpha^c)}.$$

This is the stochastic E-Step in Stochastic EM algorithm.

Maximization Step:

Given α, we update β and σ^2 to the mode of their posterior distributions.

- Update

$$\beta^{(k)} = \arg\max_{\beta} p(\beta|Y, \alpha, \sigma^2) = \frac{c}{c + \sigma^{2(k-1)}}(X^T_{\alpha^{(k)}} X_{\alpha^{(k)}})^{-1} X^T_{\alpha^{(k)}} Y$$

- Update

$$\sigma^{2(k)} = \arg\max_{\sigma^2} p(\sigma^2|Y, \beta) = \frac{\delta + (Y - X_{\alpha^{(k)}}\beta^{(k)})^T (Y - X_{\alpha^{(k)}}\beta^{(k)})/2}{\nu + \frac{n}{2} + 1}$$

4.4 Simulation Studies

We use the same design structures as in Chapter 3 to test the performance of our proposed method. The designs are

- Design 1

$$X_i \overset{iid}{\sim} N_n(0, I_n),\ i = 1, \ldots, p$$

- Design 2

$$X_i \overset{iid}{\sim} N_n(0, I_n),\ i = 1, \ldots, 10 \text{ , and}$$
$$X_i \overset{iid}{\sim} N_n(0.5X_1 + X_2 + 1.5X_3, I_n),\ i = 11, \ldots, 15$$

- Design 3

$$X_i \sim N_n(0, I_n),\ i = 1, \ldots, p \text{ and } \rho(X_i, X_j) = 0.7 \text{ for all } i \neq j$$

- Design 4

$$X_i \sim N_n(0, I_n),\ i = 1, \ldots, p \text{ and } \rho(X_i, X_j) = 0.5^{|i-j|}$$

For each design, we consider 4 models of the form $Y \sim \mathcal{MN}_{n,q}(\eta,\ \sigma^2 I,\ I)$ with different number of response variables and the results are summarized in plots. The following results are obtained by 50 simulations for each design under each model; however, since the variable selection problems for designs 2 and 3 are harder due to the existence of correlation among the predictors, the number of steps used in each iteration of simulation is twice that for designs 1 and 4 at 400 and so are burnins at 200. The number of observations and covariates are fixed at n=50 and p=20. The prior used for $p_{i,j}$ ($\pi_{i,j}$) is 0.3 for all i and j. Since k can be any number within the range $(1, \frac{1}{\pi})$, we show results for 3 different choices of k's. The columns of the plots show the box-plots of the selection probabilities obtained from

1. Univariate Bayesian variable selection
2. Our proposed method with k=1.5
3. Our proposed method with k=2
4. Our proposed method with k=3,

and the rows are in the order of the response variables. We only compare our results with those obtained by univariate Bayesian variable selection but not the ones obtained by multivariate variable selection since our method is similar to univariate Bayesian variable selection and results are better than those obtained by using multiple response Bayesian variable selection due to the allowance of selecting different predictors for different response variables. The univariate Bayesian variable selection used for comparison is obtained by using the same prior distribution for coefficient B that is $B|\alpha \sim \mathrm{N}(0{,}100(X_\alpha^T X_\alpha)^{-1})$ and constant $p_{i,j}$ for all i and j so α's are drawn from

$$p(\alpha|Y) \propto \exp\{-\frac{1}{2}[\frac{Y^T Y}{\sigma^2} - (\frac{1}{\sigma^2} + \frac{1}{c})^{-1} \frac{Y^T X_\alpha (X_\alpha^T X_\alpha)^{-1} X_\alpha^T Y}{\sigma^4}]\} p_{i,j}^{\alpha_{i,j}} (1 - p_{i,j})^{(1-\alpha_{i,j})}$$

4.4.1 Study 1

Consider the following model, $\eta = X\beta$ and $\sigma = 2.5$ where

$B^T =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

In this study, we wish to know the selection probabilities for equally strong signals with 1/5 common variables.

Design 1

The correlation matrix for Y under Design 1 is $\text{Cor}(Y) = \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix}$ as derived previously.

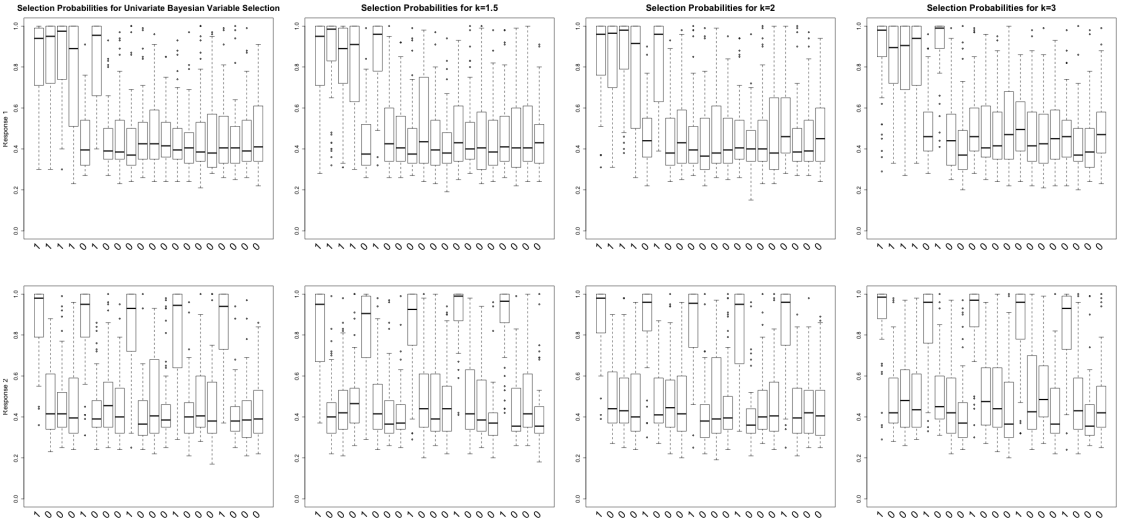


Figure 4.1: Comparing Design 1 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 1

Design 2

The correlation matrix for Y under Design 2 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.614 \\ 0.614 & 1 \end{pmatrix}$

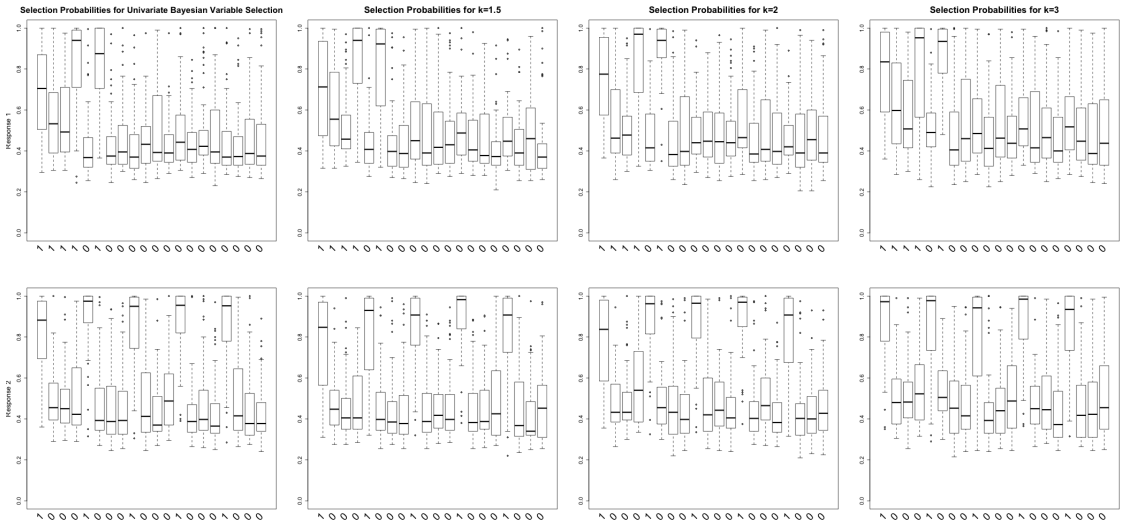


Figure 4.2: Comparing Design 2 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 1

Design 3

The correlation matrix for Y under Design 3 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.9368 \\ 0.9368 & 1 \end{pmatrix}$

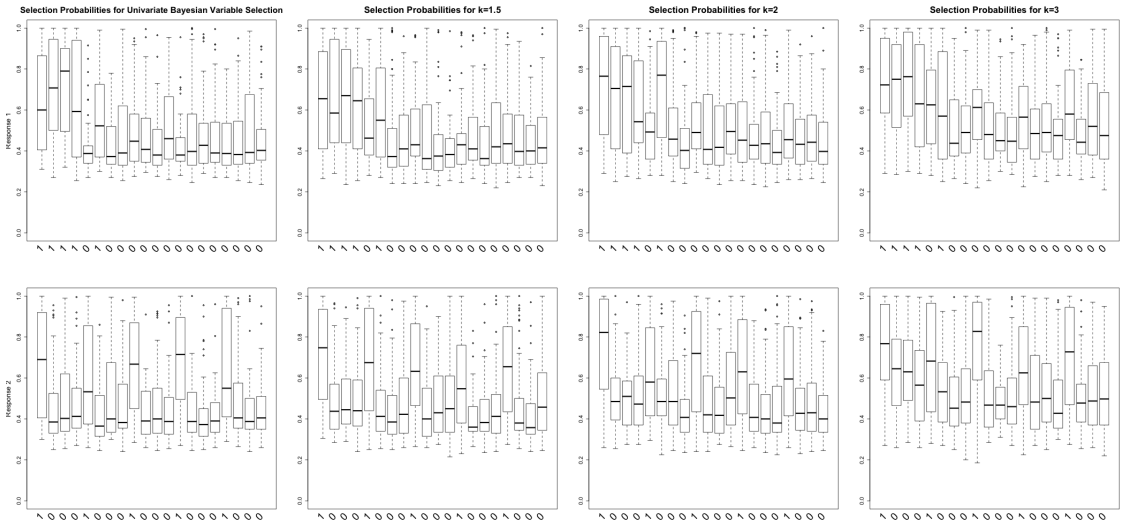


Figure 4.3: Comparing Design 3 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 1

Design 4

The correlation matrix for Y under Design 4 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.4718 \\ 0.4718 & 1 \end{pmatrix}$

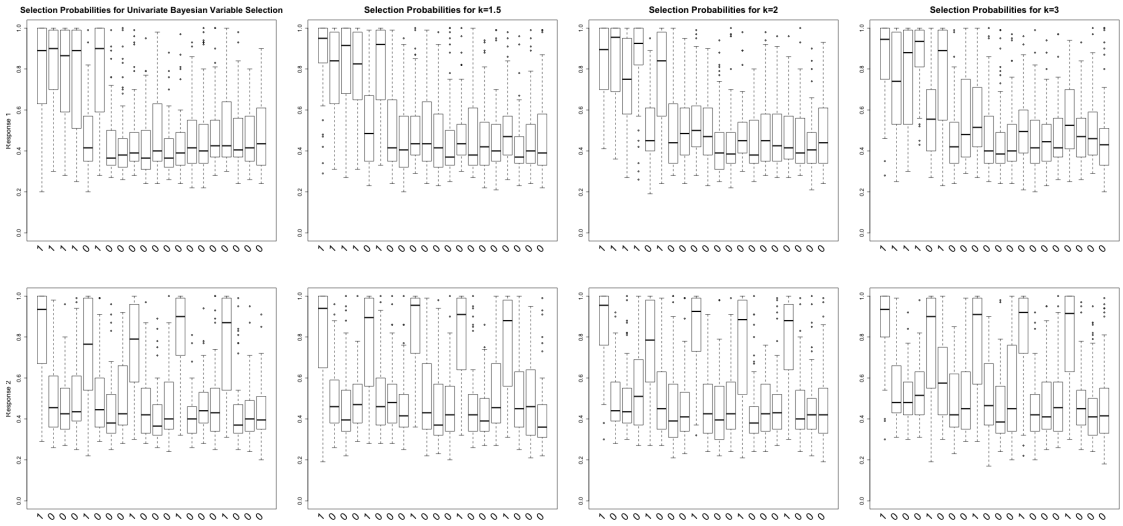


Figure 4.4: Comparing Design 4 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 1

4.4.2 Study 2

Consider the following model, $\eta = X\beta$ and $\sigma = 2.5$ where

$B^T =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

In this study, we wish to investigate the selection probabilities for equally strong signals with each pair of response variables sharing half of the predictors.

Design 1

The correlation matrix for Y under Design 1 is $\text{Cor}(Y) = \begin{pmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$

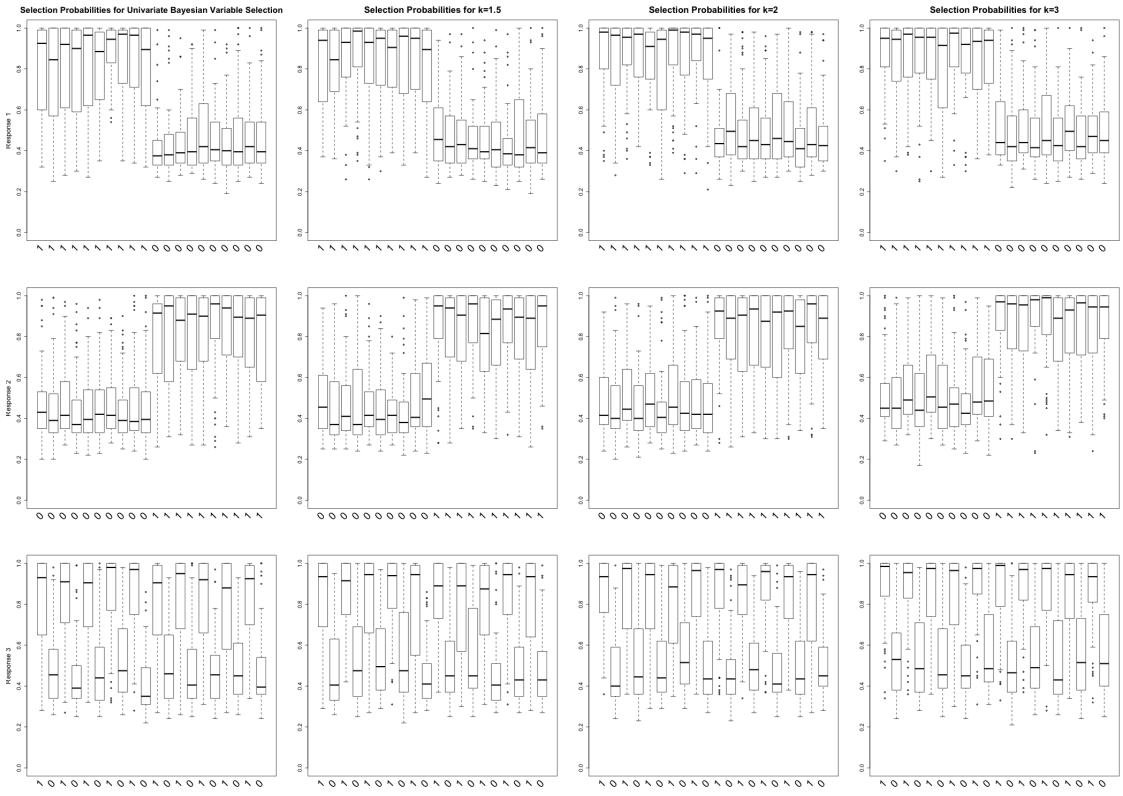


Figure 4.5: Comparing Design 1 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 2

Design 2

The correlation matrix for Y under Design 2 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.493 & 0.623 \\ 0.493 & 1 & 0.944 \\ 0.623 & 0.944 & 1 \end{pmatrix}$

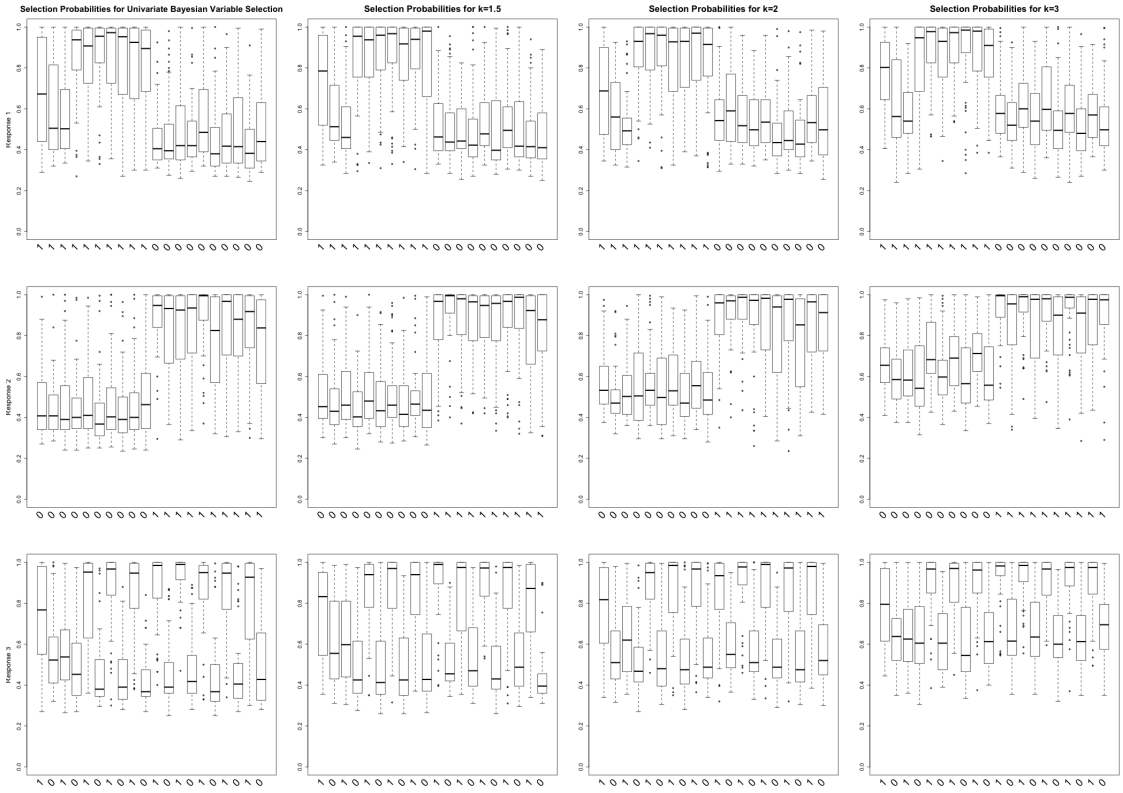


Figure 4.6: Comparing Design 2 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 2

Design 3

The correlation matrix for Y under Design 3 is $\text{Cor}(Y) = \begin{pmatrix} 1 & 0.9589 & 0.9795 \\ 0.9589 & 1 & 0.9795 \\ 0.9795 & 0.9795 & 1 \end{pmatrix}$

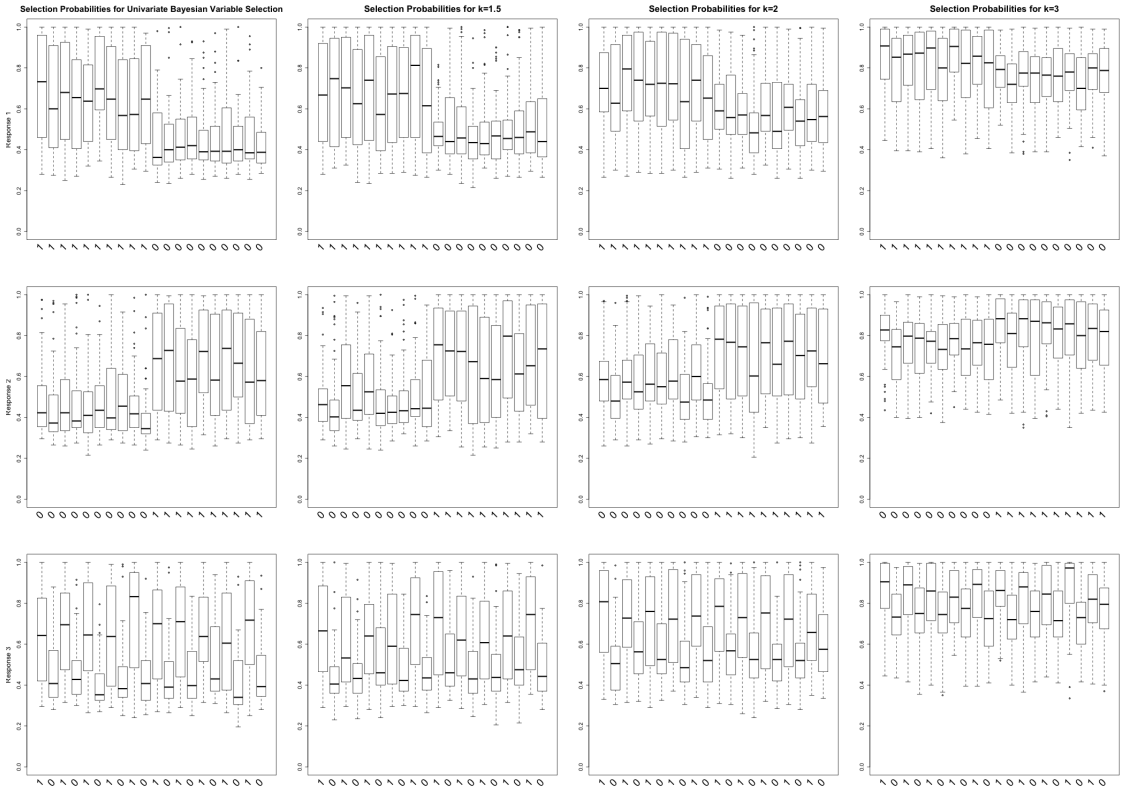


Figure 4.7: Comparing Design 3 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 2

Design 4

The correlation matrix for Y under Design 4 is $\text{Cor}(Y) = \begin{pmatrix} 1 & 0.07676 & 0.7076 \\ 0.07676 & 1 & 0.67475 \\ 0.7076 & 0.67475 & 1 \end{pmatrix}$

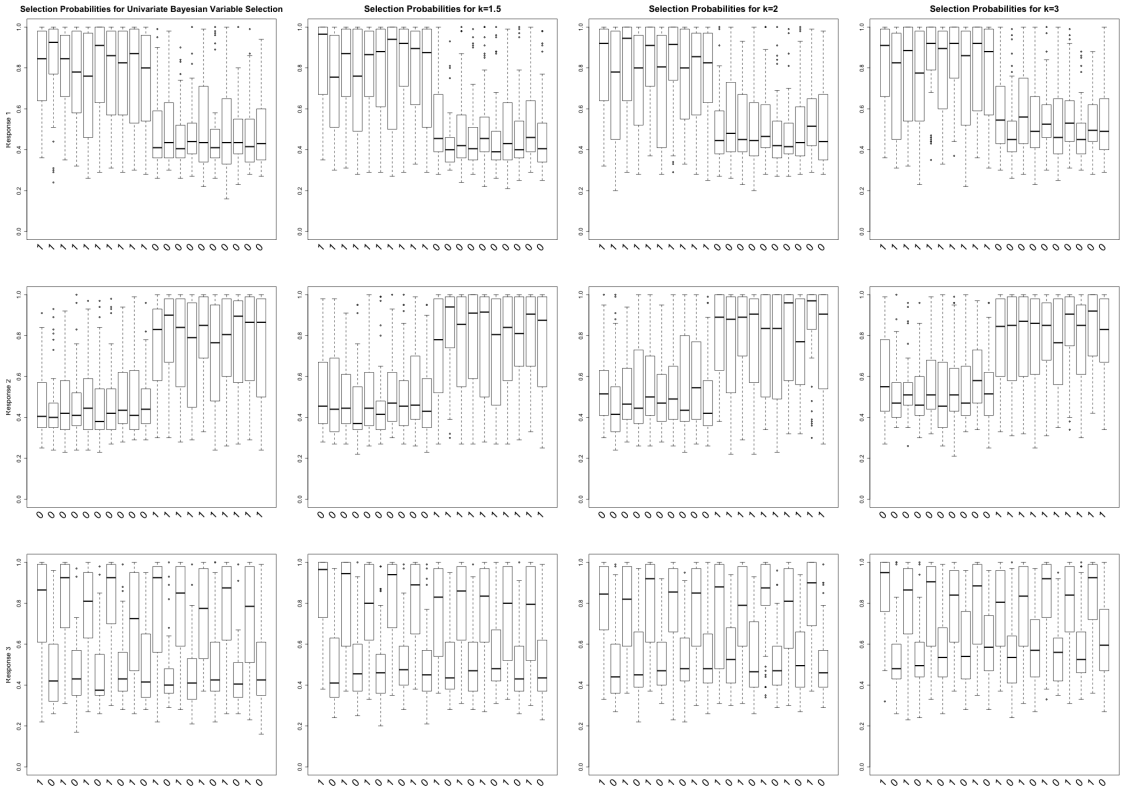


Figure 4.8: Comparing Design 4 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 2

For reasonable sized true signals (1), our proposed method performs similar to univariate Bayesian variable selection but with slightly smaller variances in the selection probabilities, which also seem to depend on the choice of k.

4.4.3 Study 3

Consider the following model, $\eta = X\beta$ and $\sigma = 2.5$ where

$B^T =$

$$\begin{pmatrix} 1 & 0.5 & 0 & 0 & 1 & 0.6 & 0.6 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.7 & 0 & 0.5 & 1 & 0 & 1.2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this study, we wish to investigate the effect of sharing of weak signals on the selection probabilities.

Design 1

The correlation matrix for Y under Design 1 is $\text{Cor}(Y) = \begin{pmatrix} 1 & 0.6746 \\ 0.6746 & 1 \end{pmatrix}$

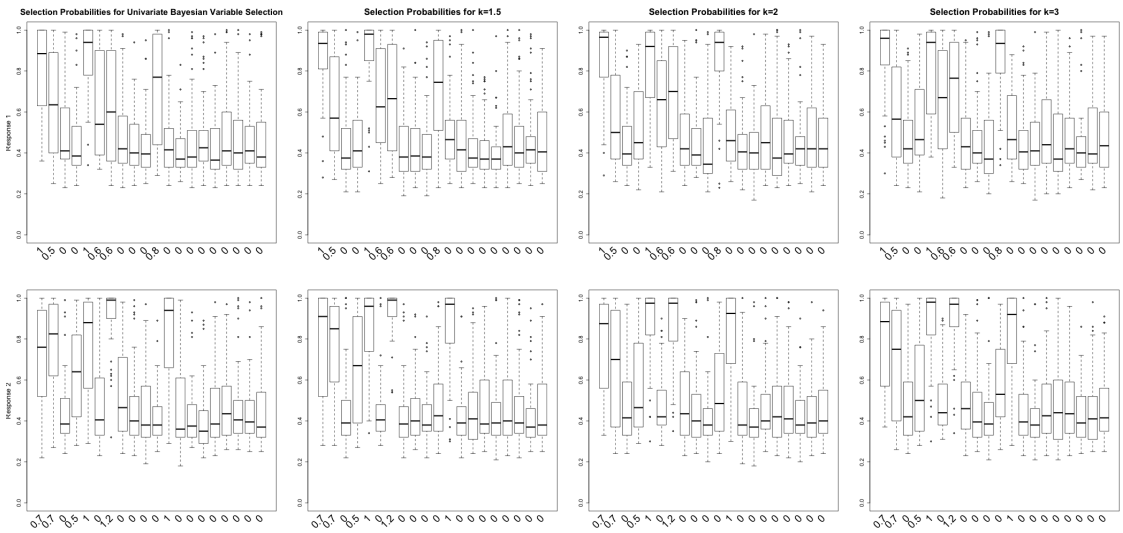


Figure 4.9: Comparing Design 1 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 3

Notice that the selection probabilities for the 2nd variable of Response 2 is smaller than that for the 1st variable on average for our proposed method although the true signals are identical. This can be explained by the size of the true signals (and hence and selection probabilities) of the corresponding variables of Response 1.

Design 2

The correlation matrix for Y under Design 2 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.9326 \\ 0.9326 & 1 \end{pmatrix}$

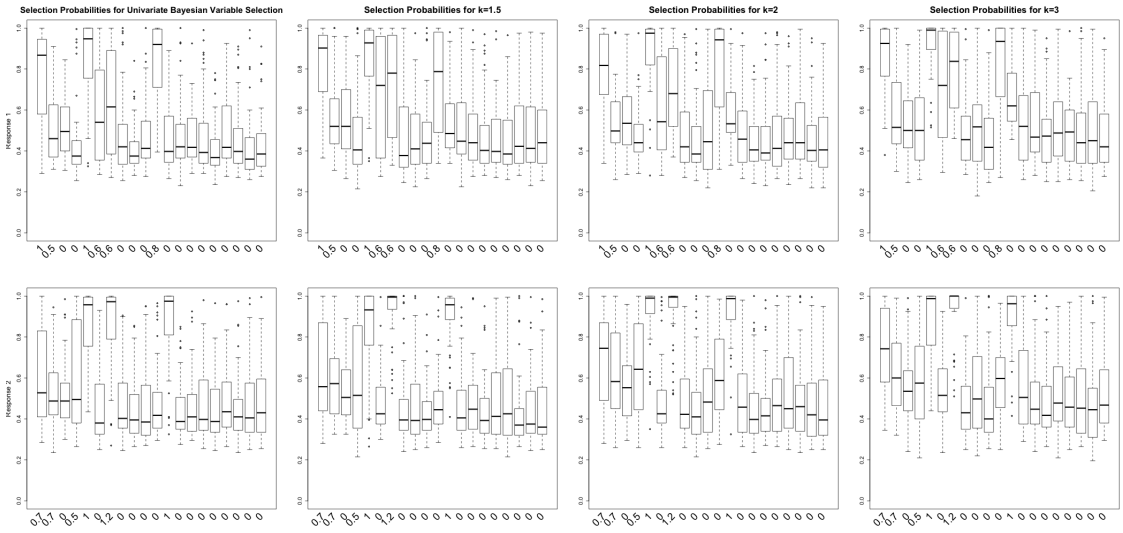


Figure 4.10: Comparing Design 2 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 3

The selection probabilities of our proposed method have smaller variances compared to univariate Bayesian variable selection when the true signals are large ($\geq$1).

Design 3

The correlation matrix for Y under Design 3 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.9768 \\ 0.9768 & 1 \end{pmatrix}$

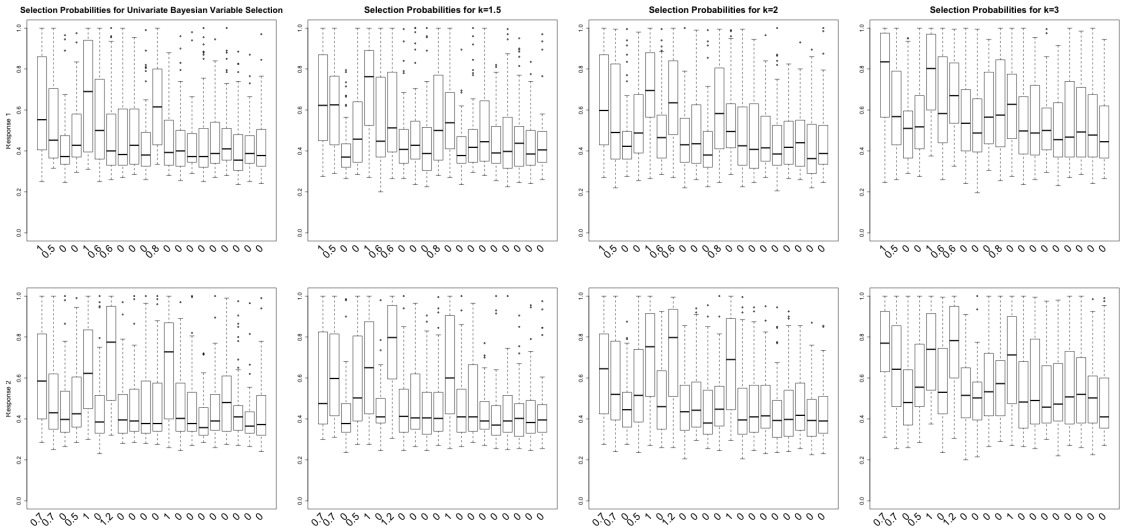


Figure 4.11: Comparing Design 3 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 3

When there exists strong correlation among the response variables and the true signals are alternating, our proposed method has a harder time picking the right variables.

Design 4

The correlation matrix for Y under Design 4 is $\text{Cor}(Y)= \begin{pmatrix} 1 & 0.89297 \\ 0.89297 & 1 \end{pmatrix}$

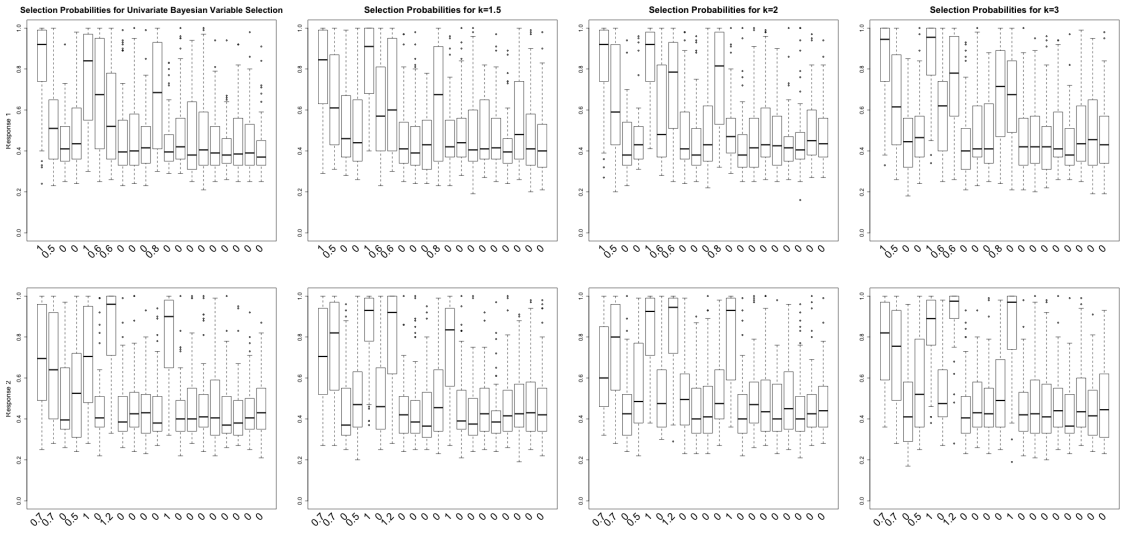


Figure 4.12: Comparing Design 4 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 3

4.4.4 Study 4

Consider the following model, $\eta = X\beta$ and $\sigma = 2.5$ where

$B^T =$

$$\begin{pmatrix} 0.5 & 0 & 0 & 0 & 1 & 0.8 & 1 & 0 & 0 & 0 & 0.6 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0.7 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.6 & 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Design 1

The correlation matrix for Y under Design 1 is $\text{Cov}(Y) = \begin{pmatrix} 1 & 0.2743 & 0.4905 \\ 0.2743 & 1 & 0.1405 \\ 0.4905 & 0.1405 & 1 \end{pmatrix}$

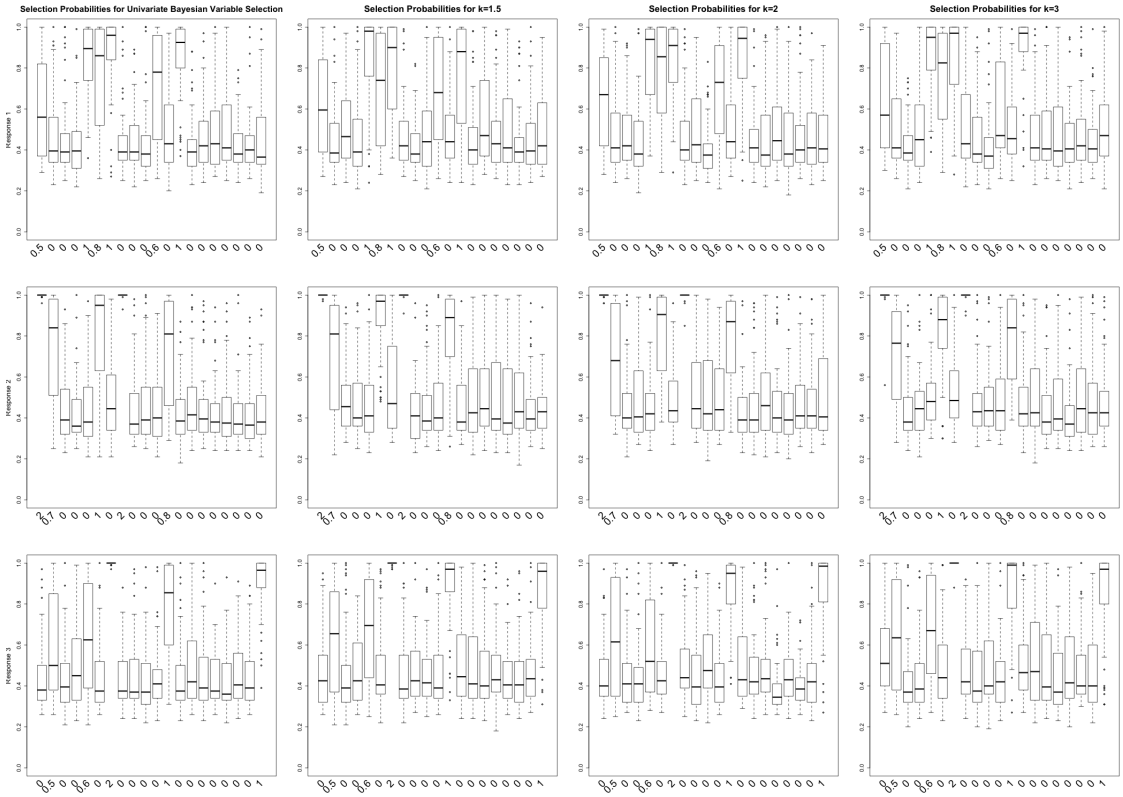


Figure 4.13: Comparing Design 1 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 4

The 12th variable of Response 2 has smaller variances in the selection probabilities under our proposed method, which can be justified by the sharing of variable with

Response 3 and in return results in smaller variances in the selection probabilities for the same variable of Response 3.

Design 2

The correlation matrix for Y under Design 2 is $\text{Cov}(Y)= \begin{pmatrix} 1 & 0.6805 & 0.8179 \\ 0.6805 & 1 & 0.4344 \\ 0.8179 & 0.4344 & 1 \end{pmatrix}$

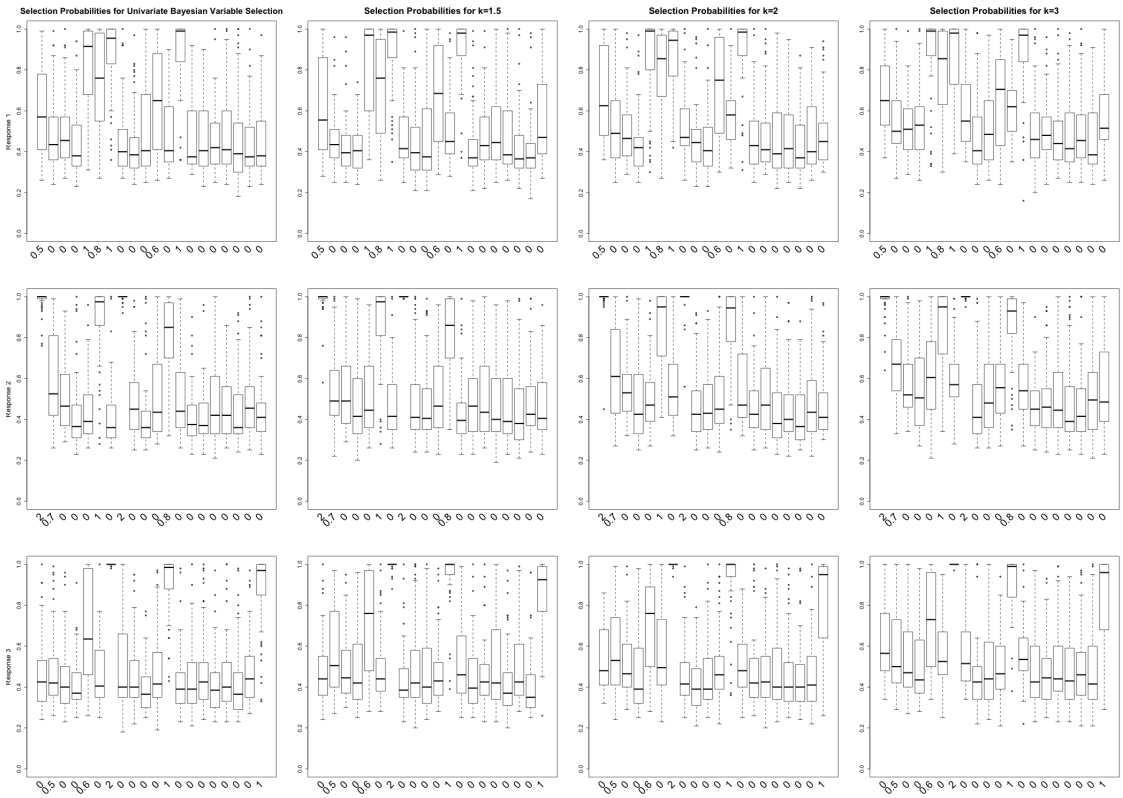


Figure 4.14: Comparing Design 2 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 4

Design 3

The correlation matrix for Y under Design 3 is $\text{Cov}(Y) = \begin{pmatrix} 1 & 0.9403 & 0.9564 \\ 0.9403 & 1 & 0.9177 \\ 0.9564 & 0.9177 & 1 \end{pmatrix}$

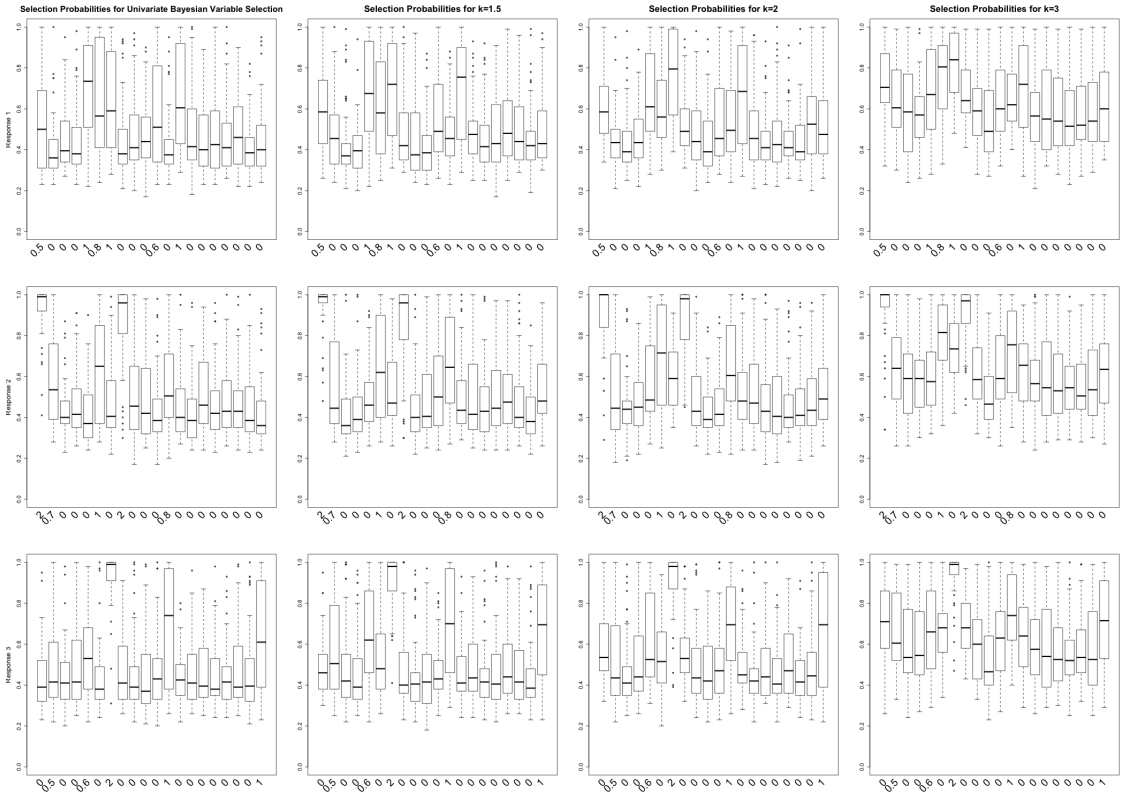


Figure 4.15: Comparing Design 3 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 4

For highly correlated response variables, the probabilities of selecting certain variables are lower if they are not selected for other response variables.

Design 4

The correlation matrix for Y under Design 4 is $\text{Cov}(Y) = \begin{pmatrix} 1 & 0.6222 & 0.7673 \\ 0.6222 & 1 & 0.5639 \\ 0.7673 & 0.5639 & 1 \end{pmatrix}$

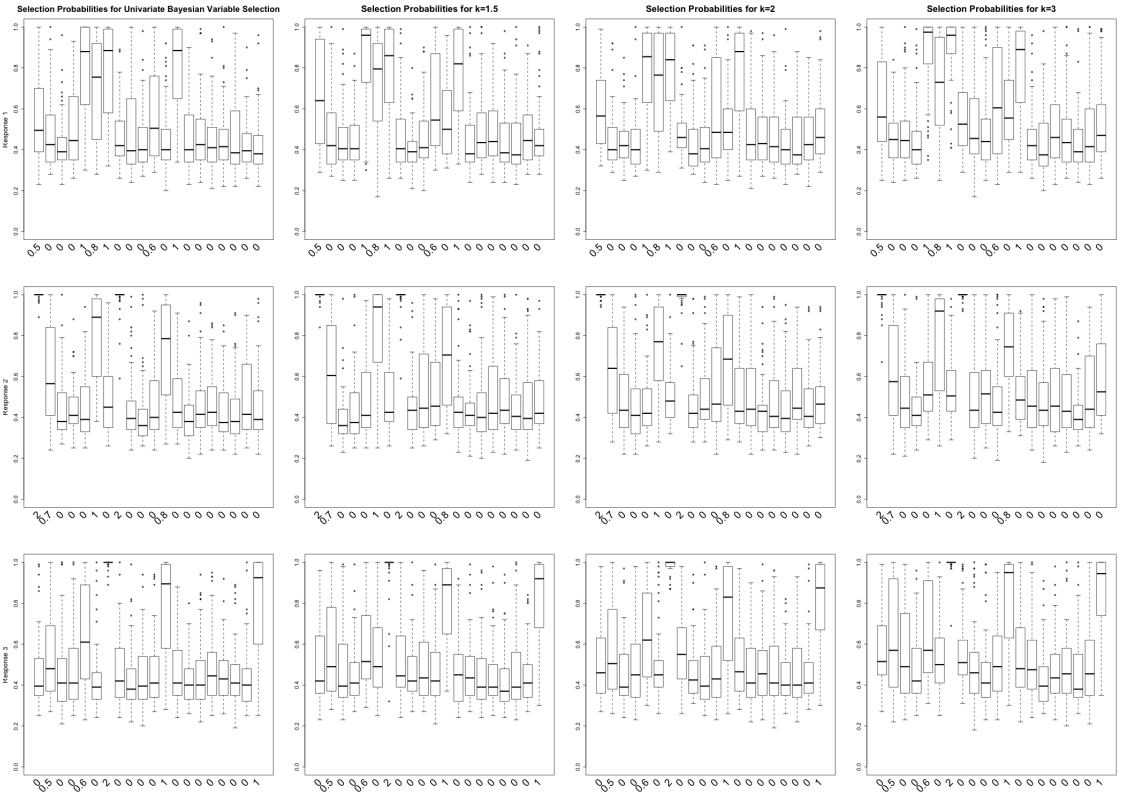


Figure 4.16: Comparing Design 4 selection probability box plots for each response variables with univariate Bayesian variable selection results under Model 4

4.4.5 Summary

Overall, it can be seen that the selection probabilities of our proposed method depend highly on the correlations among the response variables and whether the variables are selected for other response variables as expected. This also shows the influence of prior distribution in Bayesian analysis.

4.5 Data Analysis

In this section, we consider World Value Survey again but with an extra response variable "Happiness". We know that happiness is closely linked with life satisfaction; however, it is interesting to look at what are the differences in factors that contribute to a satisfied life and happiness. Here we treat the response variables as continuous.

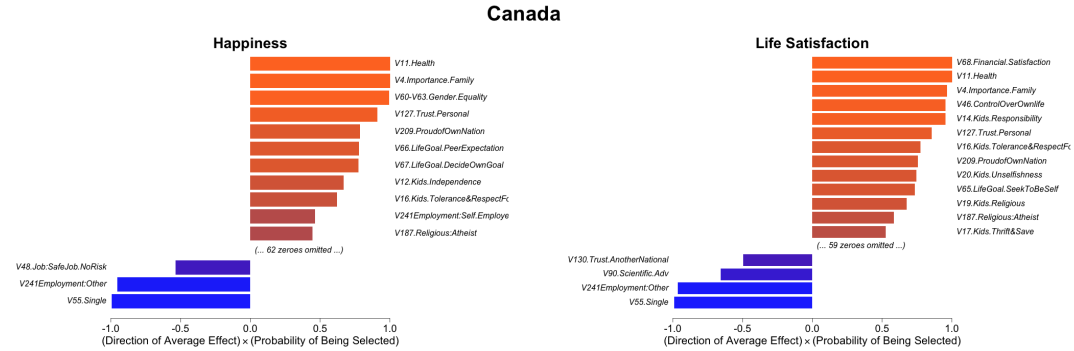


Figure 4.17: Selection probability by the direction of average effect of top ranked factors for happiness and life satisfaction in Canada

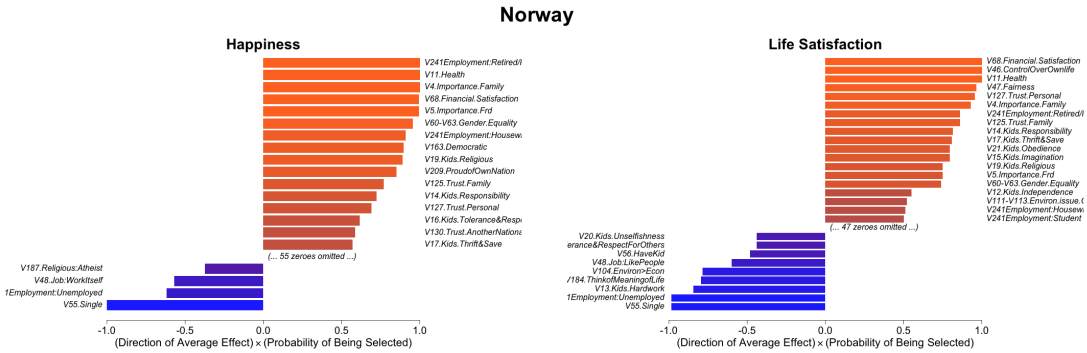


Figure 4.18: Selection probability by the direction of average effect of top ranked factors for happiness and life satisfaction in Norway

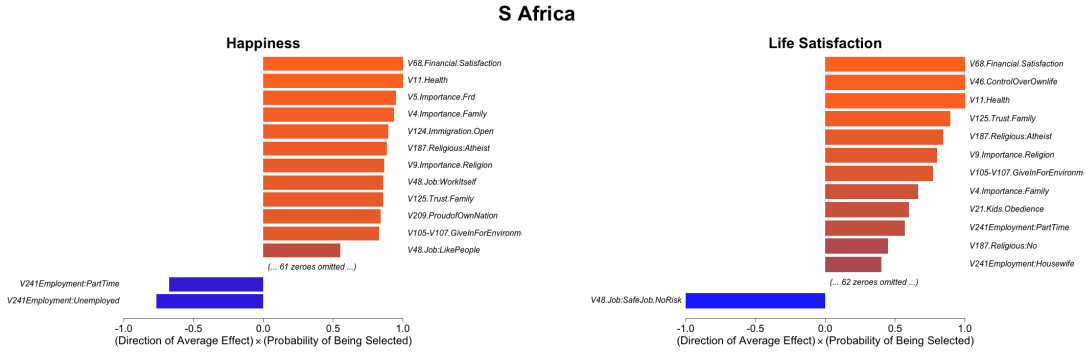


Figure 4.19: Selection probability by the direction of average effect of top ranked factors for happiness and life satisfaction in South Africa

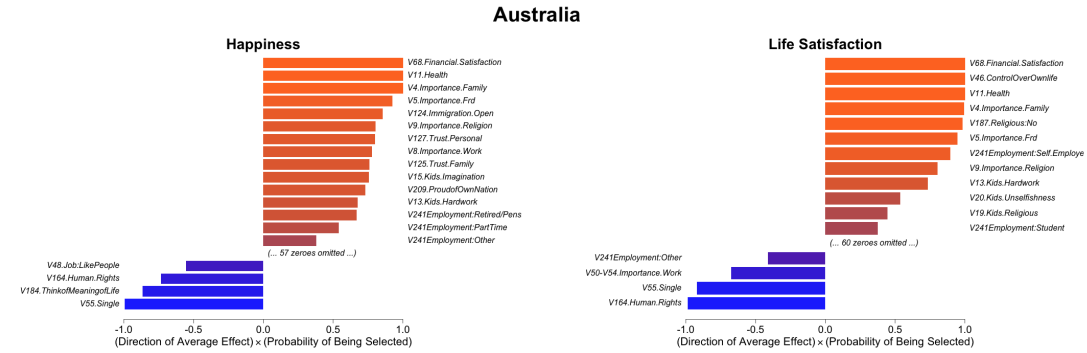


Figure 4.20: Selection probability by the direction of average effect of top ranked factors for happiness and life satisfaction in Australia

The results in this section are slightly different from those in Chapter 3 which uses LASSO to select variables. Furthermore, the response variables are treated as continuous which disregards the "true" distances between each successive levels or ratings of our response variables in the analysis. As we can see from the plots, health and financial satisfaction are among the most influential factors for both happiness and life satisfaction for most countries. However, quite interestingly we see that financial satisfaction is not listed as one of the most influential factors for happiness for Canadians, nor is it as highly ranked as it is for life satisfaction for Norwegians neither.

Chapter 5

Conclusion

While there are already lots of work devoted to analyzing ordinal response data, most of them focuses on analyzing binary data and using Bayesian variable selection to select the variables. We focused our work on analyzing polychotomous ordinal response variable by adapting the data augmented framework that is typically used for analyzing ordinal response data. Under such framework, Gibbs sampler is typically used to sample Z the latent variable, β the parameter of interest, and γ the cutpoints that are used to transform Z back to Y the response variable. For variable selection purposes, it is natural to use Bayesian variable selection since it also uses Gibbs sampling procedure. Bayesian variable selection usually involves introduction of a latent variable for indicating whether the corresponding variables should be entered so Gibbs sampling is required for approximating the joint distribution of the augmented data. While Bayesian variable selection is a convenient technique that utilizes Gibbs sampling to select the variables, for univariate variable selection problems, we proposed to use Stochastic EM and infuse LASSO in the estimation step for β the parameter of interest since it is more computationally efficient. Since we are interested in the selection probabilities, the fact that LASSO shrinks the estimates of the parameters has no impact in our analyses. However, for multiple response data, we still make use of Bayesian variable selection since it is convenient to place priors on any parameters. Therefore, we chose to infuse the relationship information into the conditional prior distribution of α (the latent variable that indicates which corresponding variables should be entered) considering that the

correlation between the response variables is one of the most important information available when analyzing multiple response data. The simulation results show that our proposed method look optimistic when being compared against univariate Bayesian variable selection in that it shows the influence of prior distributions in Bayesian analysis and how we can specify the prior distributions based on the problem of our interest. Overall, the performance of Bayesian variable selection can be modified by adding information through the prior distributions.

5.1 Future Work

1. **Choice of k**

   So far we have not proven the difference in performance of our proposed method for different choices of k; however, we can tell there is no substantial difference based on empirical evidences but it is worthwhile to know the sensitivity of the performance based on k for choosing an optimal solution to a variable selection problem.

2. **Extension to Multiple Ordinal Response Variable Selection**

   Our proposed method seem to be working really well on continuous response data so one future research possibility would be to extend this method to analyzing multiple ordinal response data. The major challenge is the estimation of correlations - currently we use sample correlation for the conditional prior distribution of α; however, for ordinal response variable Y, correlation need to be estimated with caution.

3. **High Dimensional Data Analysis**

   More and more data are high dimensional in nature (with $p \gg n$) but we have not been focusing our work on analyzing such data. To apply our proposed methods to analyzing high dimensional data we might need to make some adjustments on the prior distribution of α that penalizes or restricts the number of variables to be selected.

APPENDICES

Appendix A

The Algorithm

For convenience, we ignore i that indicates which response we are specifying here. Each step in this algorithm is repeated q number of times (one for each response i). Moreover, except for α_i which conditioned on α^c, $\beta^{(i)}$ and σ_i^2 depend only on α_j for $j = i$. The superscripts (k) indicates the *iteration*.

- Initialization
  - Set $\alpha^{(0)}$ from Bernoulli$(\pi_{i,j})$ $i = 1, \ldots, q,\ j = 1, \ldots, p$
  - Set $\beta^{(0)}$ and $\sigma^{2(0)}$ given $\alpha^{(0)}$ to the maximum likelihood estimate of β given $\alpha^{(0)}$ and σ^2 given $\alpha^{(0)}$ and $\beta^{(0)}$
    - $$\beta^{(0)} = (X_\alpha^T X_\alpha)^{-1} X_\alpha^T Y$$
    - $$\sigma^{2(0)} = \frac{|Y - X_\alpha \beta^{(0)}|^2}{n-1}$$
- At the k^{th} iteration

  Expectation Step:

  We incorporate the idea of integrating out (collapsing down) irrelevant parameters by placing conjugate prior on β to obtain the posterior distribution of α.

Draw $\alpha^{(k)}$ from $p(\alpha|Y,\sigma^2,\alpha^c)$

$$p(\alpha|Y,\sigma^2,\alpha^c) \propto \pi(\alpha|\alpha^c)\int p(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta$$
$$\propto \int f(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta \prod_{i,j} p_{ij}^{*\alpha_{ij}}(1-p_{ij}^*)^{1-\alpha_{ij}}$$

$$\int p(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta \propto \int_\beta \exp\{-\frac{1}{2}(Y-X_\alpha\beta)^T\sigma^{-2}(Y-X_\alpha\beta)\}\exp\{-\frac{1}{2}\beta^T\frac{X_\alpha^TX_\alpha}{c}\beta\}d\beta$$
$$\propto \int_\beta \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-2\frac{Y^TX_\alpha\beta}{\sigma^2}+\frac{\beta^TX_\alpha^TX_\alpha\beta}{\sigma^2}+\frac{\beta^TX_\alpha^TX_\alpha\beta}{c}]\}d\beta$$
$$\propto \int_\beta \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-2\frac{Y^TX_\alpha\beta}{\sigma^2}+\beta^T(\frac{X_\alpha^TX_\alpha}{\sigma^2}+\frac{X_\alpha^TX_\alpha}{c})\beta]\}d\beta$$

Let $\Sigma^{-1}=\frac{X_\alpha^TX_\alpha}{\sigma^2}+\frac{X_\alpha^TX_\alpha}{c}$

$$\int_\beta p(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta \propto \int_\beta \exp\{-\frac{1}{2}[\beta^T\Sigma^{-1}\beta-2\beta^T\Sigma^{-1}\Sigma\frac{X_\alpha^TY}{\sigma^2}+\frac{Y^TY}{\sigma^2}]\}d\beta$$
$$\propto \int_\beta \exp\{-\frac{1}{2}[\beta^T\Sigma^{-1}\beta-2\beta^T\Sigma^{-1}(\Sigma\frac{X_\alpha^TY}{\sigma^2})+\frac{Y^TY}{\sigma^2}$$
$$+(\Sigma\frac{X_\alpha^TY}{\sigma^2})^T\Sigma^{-1}(\Sigma\frac{X_\alpha^TY}{\sigma^2})-(\Sigma\frac{X_\alpha^TY}{\sigma^2})^T\Sigma^{-1}(\Sigma\frac{X_\alpha^TY}{\sigma^2})]\}d\beta$$
$$\propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-(\Sigma\frac{X_\alpha^TY}{\sigma^2})^T\Sigma^{-1}(\Sigma\frac{X_\alpha^TY}{\sigma^2})]\}$$
$$\int_\beta \underbrace{\exp\{-\frac{1}{2}[(\beta-\frac{\Sigma X_\alpha^TY}{\sigma^2})^T\Sigma^{-1}(\beta-\frac{\Sigma X_\alpha^TY}{\sigma^2})]\}}_{\text{Normal kernel}}d\beta$$
$$\propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-(\Sigma\frac{X_\alpha^TY}{\sigma^2})^T\frac{X_\alpha^TY}{\sigma^2}]\}$$
$$\propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-((\frac{X_\alpha^TX_\alpha}{\sigma^2}+\frac{X_\alpha^TX_\alpha}{c})^{-1}\frac{X_\alpha^TY}{\sigma^2})^T\frac{X_\alpha^TY}{\sigma^2}]\}$$
$$\propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-(\frac{1}{\sigma^2}+\frac{1}{c})^{-1}\frac{Y^TX_\alpha(X_\alpha^TX_\alpha)^{-1}X_\alpha^TY}{\sigma^4}]\}$$

$$p(\alpha|Y,\sigma^2,\alpha^c) \propto \pi(\alpha|\alpha^c)\int p(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta$$
$$\propto \int f(Y|\beta,\sigma^2)p(\beta|\alpha)d\beta \prod_{i,j} p_{ij}^{*\alpha_{ij}}(1-p_{ij}^*)^{1-\alpha_{ij}}$$
$$\propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2}-(\frac{1}{\sigma^2}+\frac{1}{c})^{-1}\frac{Y^TX_\alpha(X_\alpha^TX_\alpha)^{-1}X_\alpha^TY}{\sigma^4}]\}\prod_{i,j} p_{ij}^{*\alpha_{ij}}(1-p_{ij}^*)^{1-\alpha_{ij}}$$

So,

$$p(\alpha_{i,j}|Y,\sigma^2,\alpha^c) \propto \exp\{-\frac{1}{2}[\frac{Y^TY}{\sigma^2} - (\frac{1}{\sigma^2}+\frac{1}{c})^{-1}\frac{Y^TX_\alpha(X_\alpha^TX_\alpha)^{-1}X_\alpha^TY}{\sigma^4}]\}p_{i,j}^{*\alpha_{i,j}}(1-p_{i,j}^*)^{(1-\alpha_{i,j})}$$

where

$$p_{i,j}^* = \frac{a_{i,j}}{a_{i,j}+b_{i,j}}$$

with $a_{i,j}$ and $b_{i,j}$ specified in the previous section. Here, one sample of $\alpha_{i,j}$ is drawn one by one from Bernoulli($\frac{A}{A+1}$) for each i and each j by keeping α^c fixed.

$$A = \frac{p(\alpha_{i,j}=1|Y,\sigma^2,\alpha^c)}{p(\alpha_{i,j}=0|Y,\sigma^2,\alpha^c)}.$$

This is the stochastic E-Step in Stochastic EM algorithm.

Maximization Step:

Given α, we update β and σ^2 to the mode of their posterior distributions.

- Update

$$\beta^{(k)} = \arg\max_\beta p(\beta|Y,\alpha,\sigma^2)$$

where

$$\begin{aligned} p(\beta|Y,\alpha,\sigma^2) &\propto p(Y|\beta,\sigma^2)p(\beta|\alpha) \\ &\propto \exp\{-\frac{1}{2}[(\beta - \frac{\Sigma X_\alpha^TY}{\sigma^2})^T\Sigma^{-1}(\beta-\frac{\Sigma X_\alpha^TY}{\sigma^2})]\} \\ \beta^{(k)} = \arg\max_\beta p(\beta|Y,\alpha,\sigma^2) &= \frac{\Sigma X_\alpha^TY}{\sigma^{2(k-1)}} = (\frac{X_\alpha^TX_\alpha}{\sigma^{2(k-1)}}+\frac{X_\alpha^TX_\alpha}{c})^{-1}\frac{X_{\alpha^{(k)}}^TY}{\sigma^{2(k-1)}} \\ &= \frac{c}{c+\sigma^{2(k-1)}}(X_{\alpha^{(k)}}^TX_{\alpha^{(k)}})^{-1}X_{\alpha^{(k)}}^TY \end{aligned}$$

When c approaches infinity, $\beta^{(k)}$ is the maximum likelihood estimate.

- Update

$$\sigma^{2(k)} = \arg\max_{\sigma^2} p(\sigma^2|Y,\beta)$$

where

$$
\begin{aligned}
p(\sigma^2|Y,\beta) &\propto p(Y|\beta,\sigma^2)p(\sigma^2) \\
&\propto \exp\{-\frac{1}{2\sigma^2}(Y-X\beta)^T(Y-X\beta)\}\sigma^{2(-\nu-1)}\exp\{-\frac{\sigma}{\sigma^2}\} \\
&\propto \underbrace{\sigma^{2(-\nu-1-\frac{n}{2})}\exp\{-\frac{1}{\sigma^2}[\delta+\frac{(Y-X\beta)^T(Y-X\beta)}{2}]\}}_{\text{Inverse gamma kernel}}
\end{aligned}
$$

$$
\sigma^{2(k)} = \arg\max_{\sigma^2} p(\sigma^2|Y,\beta) = \frac{\delta+(Y-X_{\alpha^{(k)}}\beta^{(k)})^T(Y-X_{\alpha^{(k)}}\beta^{(k)})/2}{\nu+\frac{n}{2}+1}
$$

When ν and δ are taken to be ≈ 0, $\sigma^{2(k)}$ is the maximum likelihood estimate.

References

[1] Albert, J. H., and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association 88*, 422 (1993), 669–679. 13

[2] Arnold, B. C., Castillo, E., and Sarabia, J. M. Conditionally specified distributions: An introduction (with comments and a rejoinder by the authors). *Statistical Science 16*, 3 (Aug. 2001), 249–274. 28

[3] Breheny, P., and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics 5*, 1 (2011), 232–253. 11

[4] Brown, P., Vannucci, M., and Fearn, T. Multivariate bayesian variable selection and prediction. *Journal of the Royal Statistical Society B 60*, 3 (1998), 627–641. 25

[5] Celeux, G., Chauveau, D., and Diebolt, J. On Stochastic Versions of the EM Algorithm. Research Report RR-2514, 1995. 5

[6] Dellaportas, P., Forster, J. J., and Ntzoufras, I. On bayesian model and variable selection using mcmc. *Statistics and Computing 12*, 1 (Jan. 2002), 27–36. 14

[7] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B 39*, 1 (1977), 1–38. 4

[8] Fan, J., and R., L. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96* (2001), 1348–1360. 11

[9] Gelfand, A., and Smith, A. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 410 (1990), 398–409. 3

[10] Geman, S., and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell. 6*, 6 (Nov. 1984), 721–741. 3

[11] George, E. I., and McCulloch, R. E. Variable selection via Gibbs sampling. *Journal of the American Statistical Association 88* (1993), 881–889. 8, 11, 12, 25

[12] Lee, K., Sha, N., Dougherty, E., Vannucci, M., and Mallick, B. Gene selection: a bayesian variable selection approach. *Bioinformatics 19*, 1 (2003), 90–97. 12, 13

[13] Lee, W., and Liu, Y. Simultaneous multiple response regression and inverse covariance matrix estimation via penalized gaussian maximum likelihood. *J. Multivariate Analysis 111* (2012), 241–255. 24

[14] McKelvey, R. D., and Zavoina, W. A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology 4*, 1 (1975), 103–120. 7

[15] Raftery, A. E., Madigan, D., and Hoeting, J. A. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92* (1997), 179–191. 14

[16] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58* (1994), 267–288. 10

[17] Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist. 38*, 2 (2010), 894–942. 10