2017

Statistical Methods for Multivariate and Correlated Data

Xinling Xu
University of South Carolina

STATISTICAL METHODS FOR MULTIVARIATE AND CORRELATED DATA

by

Xinling Xu

Bachelor of Science
Dalian University of Technology, 2011

Master of Science
University of Minnesota–Twin Cities, 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosphy in

Biostatistics

The Norman J. Arnold School of Public Health

University of South Carolina

2017

Accepted by:

James W. Hardin, Major Professor

Bo Cai, Major Professor

Alexander Charles Mclain, Committee Member

Orgul Ozturk, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Dedication

I dedicate this to my loving and encouraging parents, Shouxiang Xu (father) and Renfeng Xu (mother).

Acknowledgments

I would like to first thank my co-advisors Dr. James W. Hardin , Dr. Bo Cai, and my research assistantship supervisor Dr. Suzanne W. McDermott.

Dr. Hardin is as patient, supportive, and encouraging as a lot of his previous students said. Thank you for giving me so much help on many manuscript edits, random statistical questions, and providing me the opportunity of presenting at conferences. And thank you for always being very positive, especially when I was struggling with my dissertation.

Dr. Cai has given me great guidance on all my Bayesian projects. Thank you for being so patient with me during the methodology development process. Thank you for discussing new ideas with me during long session of regular meetings. And thank you for being supportive and encouraging when I was looking for a job.

Dr. McDermott has been a great mentor for me. Thank you for letting me work as a research assistant for you. I have learned so much and gained so much experience during this two and a half years, which I believe is the stepping stones for me to finally landing a job. I really appreciate all the help you've given me.

I would also like to thank Dr. Alexander C McLain and Dr. Orgul D. Ozturk for agreeing to be part of this dissertation process. Thank you for taking the time to edit my dissertation and give me constructive suggestions for my projects.

Abstract

A commonly encountered data type in real life is count data, especially in self-reported behavioral studies. One issue of the self-reported count data is the inaccuracy. In the first part of the dissertation, we are going to address one specific type of inaccuracy in bivariate count data–heaping. Copula functions are used for the formulation of the bivariate distribution. Using copula functions for solving data inaccuracy problems is still a new area, which we are going to explore in this dissertation.

We also discuss the methods for variable selection when the explanatory variables are highly correlated. In particular, our method is based on the sparse Bayesian infinite factor models (Bhattacharya and Dunson, 2011). The classic Bayesian variable selection priors are integrated into the factor analysis method. The proposed method can accommodate both binary and continuous variables.

In the last part of this dissertation, we extend the Bayesian factor models into the nonparametric setting. As sometimes the normality assumption can be too strict for the data, or there are outliers that might affect the model performance, our proposed method relaxes the normality assumption, while simultaneously groups the correlated explanatory variables. Our proposed method is one of the first explorations of allowing nonparametric assumption for in a Bayesian factor analysis setting.

Table of Contents

Dedication iii

Acknowledgments iv

Abstract v

List of Tables viii

List of Figures x

Chapter 1 Introduction 1

Chapter 2 Copula-based regression models for a bivariate zero-inflated heaped count outcome 3

2.1 Introduction 3

2.2 Methods 6

2.3 Simulation Studies 13

2.4 Real Data Analysis 15

2.5 Conclusion and Discussion 18

Chapter 3 Bayesian variable selection for correlated data: a factor analysis approach 21

3.1 Introduction 21

3.2 Methods 25

3.3 Simulation Studies 30
3.4 Real Data Analysis 32
3.5 Conclusion and Discussion 38

Chapter 4 Nonparametric Bayesian latent factor modeling . . . 40
4.1 Introduction 40
4.2 Methods 42
4.3 Simulation Study 46
4.4 Real Data Analysis 49
4.5 Conclusion and Discussion 51

Chapter 5 Conclusions and future work 53
5.1 Conclusions 53
5.2 Future work 54

Bibliography 55

List of Tables

Table 2.1 Commonly Used Copula Functions 8

Table 2.2 Simulation results with 30 simulations, interval-censored heaping. . 14

Table 2.3 Simulation results with 100 simulations, scaled heaping. 14

Table 2.4 Simulation results under scaled heaping setting 1, comparing IFM with MLE. 14

Table 2.5 Frequencies of the two responses. 16

Table 2.6 Bivariate Negative Binomial model with Frank copula and corresponding independent models, Global Adult Tabacco Survey Bangladesh 2009. 17

Table 2.7 Bivariate Negative Binomial model with Clayton copula and corresponding independent models, Eban Study. 19

Table 3.1 Simulation study results based on $N = 100$ simulation data sets, mean squared prediction error (MSPE), average absolute prediction error (AAPE), and maximum absolute prediction error (MAPE) of Y. 32

Table 3.2 Simulation study results based on $N = 100$ simulations, mean square error (MSE), average absolute bias (AAB), and maximum absolute bias (MAB) of the original β. 33

Table 3.3 Percentages of false positives and true positives from Variable selection under three settings, with selected quantiles reported. . . 34

Table 3.4 Summary of the colon cancer probability of successfully classification results for training set and testing set. 35

Table 3.5 Top 32 selected genes for classification, based on 99.8th quantile cut-off point. 35

Table 4.1 Simulation study results from setting 1 based on $N = 50$ simulations with $p = 50, k = 4$, mean square error (MSE), absolute average bias (AAB), and maximum average bias (MAB) of standardized Y. 47

Table 4.2 Simulation study results from setting 2 based on $N = 50$ simulations with $p = 100, k = 10$, mean square error (MSE), absolute average bias (AAB), and maximum average bias (MAB) of standardized Y. 49

Table 4.3 Mean squared error results from four methods, using bodyfat data. 50

List of Figures

Figure 2.1 Heaping feature in a smoking behavior study. 4

Figure 3.1 Selected factor loading elements, from left to right, first row to second row, $\Lambda_{1,1}, \Lambda_{1,5}, \Lambda_{1,8}$, and $\Lambda_{1903,5}$. 36

Figure 3.2 Posterior means of factor loading for genes 1 to 100 on the latent factors. 36

Figure 3.3 Posterior means of factor loading for genes 501 to 600 on the latent factors. 37

Figure 3.4 Posterior means of factor loading for genes 1001 to 1100 on the latent factors. 37

Figure 3.5 Posterior means of factor loading for genes 1501 to 1600 on the latent factors. 38

Figure 4.1 Density of the response Y_1 from the $N = 50$ simulation dataset under setting 1. 48

Figure 4.2 Density of the response Y_1 from the $N = 50$ simulation dataset under setting 2. 48

Figure 4.3 Density plot of the percentage of body fat. 50

Chapter 1

Introduction

Correlated variables scenarios are often encountered in real life data analyses. In the case where the responses are correlated, modeling them simultaneously can be challenging, especially when they are not from the multivariate Gaussian distribution. As in the case when the covariates are correlated, both the estimation and the prediction procedures would be affected, as well as the selection of important predictors.

In this dissertation, I will develop an estimation solution when the response is a bivariate vector, with marginal count distribution (Poisson, zero-inflated Poisson, negative binomial, zero-inflated negative binomial). More specifically, the marginal count data are heaped on certain values. Heaped data commonly result from retrospective studies, where subjects are asked to recall the frequency that which certain events happened over a period of time. Because questions are asked after the events have happened, the reported counts are often an approximation of the underlying true counts. In the literature, there are existing methods for dealing with this type of univariate count data. Inspired by a real life data set, I will extend the existing methods to much more complicated bivariate cases.

In a second research focus, I address the case when the number of covariates is large and correlated. In this case, caution should be taken for the selection of the important groups of covariates. There is a rich literature of methods for variable selection under both frequentist and Bayesian schemes. A classic approach to variable selection is to use a hierarchical normal mixture model with latent variables. Combining

this classic method with Bayesian infinite-factor models, we develop a method for correlated covariates selection under the simplest setting where both the response and the covariates are continuous. I will extend this method to the binary response scenario.

To relax the normal assumption for the continuous response, I will explore nonparametric Bayesian latent factor models, with correlated variable selection. Although the idea of nonparametric latent factor model has been suggested in the Bayesian literature, most of the developments tried to relax the dimension of latent factors based on a nonparametric distribution. I assume a Dirichlet process as the underlying distribution for the response, and examine the performance of my variable selection method, as well as the resulting factor loading matrix.

In this dissertation, each of chapters 2 through 4 can be seen as a separate targeted manuscript. Chapter 2 describes the methods for bivariate heaped count data, using frequentist methods. In Chapter 3, Bayesian factor models will be introduced. Correlated variable selection under parametric assumptions will be developed and then applied to real data. Chapter 4 extends the methods in Chapter 3 to the nonparametric scenario, to allow heavy tails and multimodal densities. I conclude the dissertation in Chapter 5 with future research ideas and applications.

Chapter 2

Copula-based regression models for a bivariate zero-inflated heaped count outcome

We present a new approach to modeling bivariate zero-inflated count outcomes which are heaped on certain values. We discuss heaping under two assumptions, interval-censored heaping and scaled heaping. Multiple imputation is used for interval-censored heaping to obtain an estimate of the probability of heaping. Mixture modeling is applied when the data are from scaled heaping. We adopt several copula functions to account for correlation between paired outcomes. Simulation studies are presented to illustrate the performance of our methods. We also briefly compare estimation via inference functions for margins (IFM) with maximum likelihood estimation. This chapter was motivated by a desire to analyze data from a real-life study on the intervention for risk reduction among HIV African-American serodiscordant couples, on which we apply our methods.

2.1 Introduction

Heaped data commonly exist in self-reported studies. This type of data exhibits "inflated" frequencies at certain outcomes, resulting from, for example, rounded responses or digital preferences. One can observe resulting high frequencies on certain intervals, as shown in Figure 2.1 with taller bars for more commonly reported outcomes. Heaping is a special case of data coarsening (Heitjan and Rubin, 1991). Ignoring the heaping feature of the data can lead to biased estimates and incorrect

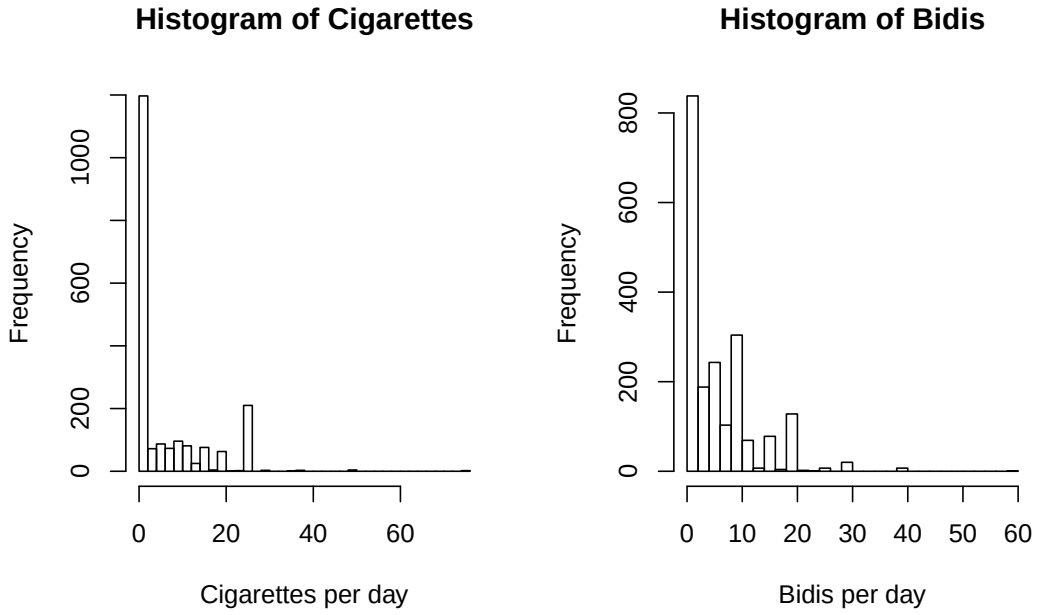

Figure 2.1 Heaping feature in a smoking behavior study.

inference.

Heaping in univariate responses has been studied extensively in the literature. One paper proposed a mixture-model approach for continuous smoking cession data (Bar and Lillard, 2011). Heaping in self-reported income data has also been addressed (Zinn and Würbach, 2016). Bayesian mixture models for heaping were developed for longitudinal count data (Suchard et al., 2015). Most of these researcher studied heaping under continuous distributions, such as the *gamma* distribution for annual income data and the *normal* distribution for quit smoking age. Heaping in count data has been discussed as well, with available commands for univariate outcomes in Stata (Cummings et al., 2015). However, to our knowledge, heaping in the bivariate responses has not been studied.

One study targeting risk reduction intervention for HIV-serodiscordant African American couples motivated the methodology development in this chapter (NIMH Multislice HIV/STD Prevention Trial for African American Couples Group, 2008).

In this study, HIV-serodiscordant couples were recruited and randomly assigned to the intervention group or the active comparison group. One aim was to reduce the number of unprotected vaginal sex episodes within couples, as well as reducing the probability of having unprotected vaginal sex completely. Questions were asked regarding the subjects' sexual behaviors. When asked to report the number of unprotected vaginal sex episodes (with their study partner) in the last 90 days, the answers from the female and male members were generally different, and exhibited heaping. One way to address the question of whether the intervention was effective was to model the correlated responses simultaneously, which is our goal in this chapter.

To address the analysis of heaped bivariate heaped data, the first issue we face is to choose which of two scenarios could lead to the heaped responses. The first scenario is that data are being rounded to the nearest observed heaping point, i.e. the heaping mechanism is interval-censoring. The second scenario assumes that data are being scaled down or up, i.e. the heaping mechanism is scaling. Scaling occurs, for example, when a person is asked to recall the number of events that occurred in the last month, but the respondent considers the number of events in a week and then multiplies that count by 4 to give as a response. In the case of scaling, the probability of heaping can be estimated directly using mixture models. However, the estimation of the heaping probability when the data are interval-censored is less straightforward. By using multiple imputation, we provide one crude way of estimating the heaping probability.

The second issue we face is how to take the correlation into account between the two responses of paired data. Unlike the normal distribution, there are no standard forms for the probability density function of multivariate count distributions. One parameterization for bivariate Poisson and two parameterizations for the negative binomial distributions have been presented (Famoye, 2010; Marshall and Olkin, 1985). Another well-known method for modeling multivariate distribution is to use copula

functions. Copula functions have been extensively used for multivariate continuous responses, for which case the copula function is unique. When the responses are discrete, the copula is unique on the range of the marginal distribution. Hence, it makes utilizing copula functions possible for discrete variables. Bivariate copula constructions were developed for multivariate discrete data (Panagiotelis et al., 2012). Joint regression analysis of correlated data was studied using Gaussian copulas (Song et al., 2009), as well as some other applications (Zhao and Zhou, 2012; Nikoloulopoulos and Karlis, 2010).

In this chapter, we combine the techniques of addressing univariate heaping count data with copula functions, and more importantly extend the univariate techniques to the bivariate cases. We are going to use multiple imputation for data with interval-censored heaping. For scaled heaping, we are going to illustrate an application of mixture models using copula functions. Maximum likelihood estimation (MLE) is used for both point estimates. Bootstrapping is used for the variance estimation under the interval-censored heaping assumption. Simulation studies are done to evaluate the performance of the methods. More specifically, we compare the estimation method using two-step inference function for margins (IFM) with MLE (Joe, 1997). Real data analysis is carried out in the last section before the conclusion for this chapter.

2.2 Methods

Copula Functions and Count Marginals

Suppose we have n pair of count observations $(Y_{i1}, Y_{i2}), i = 1, ..., n$, where $Y_{ij}, j = 1, 2$ follows a zero-inflated count distribution with parameters (ω_{ij}, μ_{ij}). Here, ω_{ij} and μ_{ij} are the zero-inflation probability and the mean for Y_{ij}. Note that, the non-zero-inflated count distribution is a special case for zero-inflated count distribution, with $\omega_{ij} = 0$. In this chapter, we only consider the zero-inflated count distributions. When Y_{ij} follows a negative binomial distribution, an additional dispersion parameter α_{ij}

is introduced. It is well-known that the probability distribution function of Y_{ij} is,

$$p(Y_{ij} = y_{ij}) = \begin{cases} \omega_{ij} + (1 - \omega_{ij}) f(y_{ij}|\mu_{ij}) & y_{ij} = 0 \\ (1 - \omega_{ij}) f(y_{ij}|\mu_{ij}) & y_{ij} = 1, 2, ... \end{cases} \tag{2.1}$$

The distribution for Y_{i1} and Y_{i2} are not necessarily the same. They can follow different distributions. Later we will see that they can have different heaping patterns as well.

Sklar's theorem states that for any $q-$dimensional random vector $Y = (Y_1, Y_2, ..., Y_q)$ with marginal cumulative distribution functions (*cdf*), $F_1(\cdot), F_2(\cdot), ..., F_q(\cdot)$, there exists a copula function C such that,

$$F(y_1, ..., y_q) = C(F_1(y_1), ..., F_q(y_q)|\theta),$$

where θ is the dependency parameter vector with length $q - 1$. It also states when the multivariate *cdf*s are continuous, the copula representation is unique. In the case of discrete marginals, the copula is unique on $\prod_{k=1}^{q} Ran(F_k)$, where $Ran(F_k)$ is the range of the marginal distribution F_k.

In our case, $q = 2$ and the joint *cdf* of (Y_{i1}, Y_{i2}) can be written as,

$$F(y_{i1}, y_{i2}) = C(F_1(y_{i1}), F_2(y_{i2})|\theta), \tag{2.2}$$

where $F_1(\cdot)$ and $F_2(\cdot)$ are the *cdf*s of the marginal count distributions, and scalar θ is the dependency parameter. Table 2.1 shows some commonly used copula functions. The Frank and Normal copulas can accommodate positive or negative correlation, while the Clayton copula only allows positive correlation. The Product, Frank, and Clayton copulas all belong to the Archimedean copulas family, which admits an explicit formula. Although the normal copula does not have an explicit formula, it has the most straightforward interpretation for the dependency parameter, which is interpreted the same as for the Pearson correlation parameter.

The joint *cdf* of Y_{i1} and Y_{i2} is expressed using a chosen copula function as shown in Equation 2.2. In order to obtain the joint probability distribution function (*pdf*) of Y_{i1}, and Y_{i2}, one needs to write it into four parts,

Table 2.1 Commonly Used Copula Functions

Copula type	Function $C(u, v)$	Dependence
Product	uv	NA
Frank	$-r^{-1}\log((\eta-(1-e^{-ru})(1-e^{-rv}))/\eta)$	$r \in \mathbb{R}\backslash\{0\}$
Normal	$\Phi_2\left[\Phi^{-1}(u), \Phi^{-1}(v); r\right]$	$-1 \leq r \leq +1$
Clayton	$(u^{-r}+v^{-r}-1)^{-1/r}$	$r \in [0, \infty)$

Note: $\eta = 1 - exp(-\theta)$, u and v are the marginal *cdf* of each response.

$$
\begin{aligned}
f(y_{i1}, y_{i2}) = & C(F_1(Y_{i1} \leq y_{i1}), F_2(Y_{i2} \leq y_{i2})|\theta) \\
& - C(F_1(Y_{i1} \leq y_{i1}-1), F_2(Y_{i2} \leq y_{i2})|\theta) \\
& - C(F_1(Y_{i1} \leq y_{i1}), F_2(Y_{i2} \leq y_{i2}-1)|\theta) \\
& + C(F_1(Y_{i1} \leq y_{i1}-1), F_2(Y_{i2} \leq y_{i2}-1)|\theta), \qquad (2.3)
\end{aligned}
$$

where $F_1(\cdot), F_2(\cdot)$ are the *cdf*s of the marginal zero-inflated count variables, and θ is the dependence parameter between Y_{i1} and Y_{i2} . Once the likelihood for each observation is known, the estimation can be done using MLE on Equation 2.4.

$$
l(y_1, y_2) = \sum_{i=1}^{n} log\ l_i = \sum_{i=1}^{n} log\ f(y_{i1}, y_{i2}) \qquad (2.4)
$$

One estimation method named *inference functions for margins* (IFM) was proposed to obtain the estimated parameters when the response is high-dimensional (Joe, 1997). This approach consists of estimating univariate parameters from separately maximizing each univariate likelihood, and then estimating dependency parameters from the multivariate likelihood. We will carry out a short simulation study in Chapter 2.3 to compare the performance of this method versus MLE (which simultaneously estimates all parameters).

Suppose $f_1(y_{i1})$ and $f_2(y_{i2})$ are the two univariate *pdf*s of Y_{i1} and Y_{i2}, the IFM estimation is obtained by first maximizing the two marginal likelihood separately, i.e.,

$$
l_j(\mu_j) = \sum_{i=1}^{n} log\ f_j(y_{ij}|\mu_j), \qquad j = 1, 2,
$$

where μ_1 and μ_2 are the parameters corresponding to the two responses respectively. Given $\hat{\mu_1}$ and $\hat{\mu_2}$, then maximizing the joint log likelihood

$$l(\theta|y_{i1}, y_{i2}, \hat{\mu}_1, \hat{\mu}_2) = \sum_{i=1}^{n} log f(\theta|y_{i1}, y_{i2}, \hat{\mu}_1, \hat{\mu}_2),$$

we obtain the estimation for the dependency parameter $\hat{\theta}$. The parameter estimates asymptotically follow a multivariate normal distribution. The standard error estimation can be carried out using Jackknife or bootstrapping methods.

Heaping and Multiple Imputation

The copula functions mentioned in the previous section can be applied to any multivariate standard count distributions. When data exhibit excessive heaping features, directly applying copula functions would lead to biased estimation of the parameters. In this chapter, we assume the heaping points are defined ex ante. The identification of heaping points was discussed in a previous study, where the authors used the comparison between the empirical *cdf* and the hypothesized *cdf* to identify the heaping points (Zinn and Würbach, 2016). There are two commonly encountered heaping mechanisms (Cummings et al., 2015), heaping due to mixture of scaled distributions and interval-censored responses. We will apply the same heaping mechanisms for the bivariate cases.

We consider the mixture of scaled distributions first. There are two behaviors in this type of heaping data, subjects who report an exact count and subjects who recall based on the frequency over $1/k$ th of a specified period of time and then report k times that amount (Cummings et al., 2015). Here we are assuming that there is only one k. The generalization of multiple different ks is straightforward.

We will use an indicator variable b_i to distinguish the behaviors, with $b_i = 1$ indicating that the subject reports from the scaled behavior, and $b_i = 0$ indicating the subject reports on the original scale. Hence, the marginal *cdf* of $Y_{ij}, i = 1, ..., n, j =$

$1, 2$, is

$$F_j(Y_{ij} \leq y_{ij}) = p(b_{ij} = 0)F_j(Y_{ij} \leq y_{ij}|\mu_{ij}, \omega_{ij}) + I_{(y_{ij} mod\ k=0)}p(b_{ij} = 1)F_j(Y_{ij} \leq \frac{y_{ij}}{k}|\frac{\mu_{ij}}{k}, \omega_{ij}), \tag{2.5}$$

where μ_{ij} is the mean of Y_{ij}, and ω_{ij} is the zero-inflated probability. Similar to logistic regression, one can specify the covariates that predict the probability of having a scaled heaping behavior, i.e. $p(b_{ij} = 1)$. By substituting Equation 2.5 into Equation 2.3, the likelihood of each pair observations is known and estimation can be carried out using either maximum likelihood estimation or IFM.

The straightforward way of estimating the heaping probability does not apply to the interval-censored heaping. When the data exhibit interval-censored feature, the observed data no longer follow a standard count distribution. For example, in studies related to cigarette counts, the observed data might have excessive frequencies on multiples of $5s$, $10s$ or $20s$, as subjects might recall based on packs or half packs. Here we use h to denote the value that data heap on multiples of h, e.g. $h = 5, 10, 20$. If the data only heap on multiples of $5s$, the observed value 40 can be either the true cigarette count or it is resulted from rounding up or down from 36 to 44.

Similar as in the univariate case, given that $y_{ij}\ mod\ h = 0$, we assume that only if the true value falls into a close neighborhood of the observed y_{ij}, $[y_{ij} - \lfloor \frac{h}{2} \rfloor, y_{ij} + \lfloor \frac{h}{2} \rfloor]$, can be reported as the y_{ij}, where we define "close" as $\lfloor \frac{h}{2} \rfloor$, the floor function of $\frac{h}{2}$. This assumption is crucial for the multiple imputation we are going to show. Once the heaping points have been "replaced" with regular count values, the usual procedure using copula functions can be applied directly.

The seconed assumption for interval-censored data is that subjects round up or down at random. That is, each person has the equal probability $P(H_{ij} = 1) = p_j$ of rounding up or down to the nearest multiples of h, where H_i is the indicator variable for heaping and $j = 1, 2$. This implies that variables H_{ij} and the count Y_{ij} are

independent. As suggested in a previous study (Bar and Lillard, 2011), we will use

$$\hat{p}_j = \frac{O_j - E_j}{n} \tag{2.6}$$

to estimate the heaping probability, where O_j is the observed number of y_{ij}s that satisfy $y_{ij}\ mod\ h = 0$, and E_j is the expected number of that y_{ij}s that fall on these value given the estimated parameters, μ_{ij} and ω_{ij}.

Now suppose we observe the data heap at multiples of $h = 5$, e.g. $5, 10, 15, 20, ...$ As stated earlier, only counts in a close neighborhood of $5, 10, 15, 20, ...$ can heap at these values. More specifically, only Y_{ij} falls into $[5 - \lfloor h/2 \rfloor, 5 + \lfloor h/2 \rfloor]$, $[10 - \lfloor h/2 \rfloor$, $10 + \lfloor h/2 \rfloor]$,..., can take values of $5, 10, ...$ respectively.

Given that an observed value satisfies $y_{ij}\ mod\ h = 0$, we use Bayes' rule to get the probability of y_{ij} being heaped,

$$P(H_{ij} = 1 | y_{ij}\ mod\ h = 0) = \frac{P(y_{ij}\ mod\ h = 0 | H_{ij} = 1) P(H_{ij} = 1)}{P(y_{ij}\ mod\ h = 0)}$$

$$= \frac{\hat{p}_j \sum_{y_{ij} - \lfloor \frac{h}{2} \rfloor}^{y_{ij} + \lfloor \frac{h}{2} \rfloor} P(y_{ij} | \tilde{\mu}_{ij}, \tilde{\omega}_{ij})}{P(y_{ij}\ mod\ h = 0 | H_{ij} = 1) P(H_{ij=1}) + P(y_{ij}\ mod\ h = 0 | H_{ij} = 0) P(H_{ij} = 0)}, \tag{2.7}$$

where $\hat{p}_{ij}$ is the estimation we get from Equation 2.6, $\tilde{\mu}_{ij}$ and $\tilde{\omega_{ij}}$ are the initial estimators of the parameters from estimating the univariate zero-inflated count distribution respectively for Y_1 and Y_2. The "missing" value, i.e., the heaping value imputation is done using this conditional probability.

For y_{ij}s that do not fall on the heaping values, we use the observed values during the estimation process. For the ones that fall on the heaping values, the imputation probability is decided by Equation 2.7. If $H_{ij} = 1$, we randomly sample a number from $\left[y_{ij} - \lfloor \frac{h}{2} \rfloor, y_{ij} + \lfloor \frac{h}{2} \rfloor\right]$, and use it as the "observed" data. The steps of our algorithm are shown below,

Step 1 Estimate the two responses separately using zero-inflated count models, ignoring the heaping patterns. Obtain the estimated parameters $\tilde{\omega}_1, \tilde{\mu}_1, \tilde{\omega}_2, \tilde{\mu}_2, \tilde{\theta}$.

In the simplest case, all Y_{i1} has the same mean μ_1, zero-inflated probability ω_1, and all Y_{i2} has the same mean μ_2, zero-inflated probability ω_2.

Step 2 Given the estimated parameters in *Step 1*, one can estimate the marginal heaping probability for Y_1 and Y_2, $\hat{p}_1$ and $\hat{p}_2$.

Step 3 For each y_{ij} that falls on the heaping values, randomly assign it with a new value from $\left[y_{ij} - \lfloor \frac{h}{2} \rfloor, y_{ij} + \lfloor \frac{h}{2} \rfloor\right]$, with probability $\hat{P}(H_{ij} = 1 | y_{ij} \ mod \ h = 0)$, from Equation 2.7.

Step 4 Create 5 imputed data sets and analyze each one as it is from a zero-inflated distribution.

Step 5 Average over the 5 estimated parameters to get the point estimates.

It has been shown that there is no significant difference between 5 imputation datasets and 100 imputation datasets (Heitjan and Rubin, 1990). Hence, we decide to use 5 here to reduce the imputation time.

To obtain estimates of the standard errors, we apply bootstrapping on the original data for 500 times. Multiple imputation is applied on each bootstrapped data set. Previous simulation studies have demonstrated that when using multiple imputation and bootstrap simultaneously, bootstrapping the original data, and then applying multiple imputation to the bootstrapped datasets give better results than vice versa (Schomaker and Heumann, 2016). For the IFM method, we sample directly from the original data set to obtain the standard errors estimation under scaled heaping assumption. The performance of IFM under the interval-censored heaping assumption will not be discussed.

2.3 Simulation Studies

Simulations were used to examine the performance of our method. For interval-censored data, we simulated data under three heaping scenarios, with heaping probability on multiples of $h = 5$ for Y_1 and Y_2 respectively as (1)$p_1 = 0.3, p_2 = 0.25$, (2) $p_1 = 0.4, p_2 = 0.2$, and (3) $p_1 = 0.35, p_2 = 0.35$. Other parameters set-up is, $n = 2000, \omega_1 = 0.15, \omega_2 = 0.2, Y_1 \sim NegBin(\mu_1 = 14, \alpha_1 = 3), Y_2 \sim NegBin(\mu_2 = 17, \alpha_2 = 4)$ and initial Pearson's correlation of Y_1 and Y_2 is $\rho = 0.5$. We simulate 30 data sets for each scenario.

For scaled heaping, we assume,

- Y_1 and Y_2 both follow a zero-inflated Poisson distribution with $\mu_1 = 19, \omega_1 = 0.15, \mu_2 = 14, \omega_2 = 0.15$, initial Pearson correlation $\rho = 0.6$, and $p(b_{i1} = 1) = 0.25, p(b_{i2} = 1) = 0.25$. The scaled heaping happens on multiples of $h = 4$.

- Y_1 and Y_2 both follow a zero-inflated Poisson distribution with $\mu_1 = 22, \omega_1 = 0.2, \mu_2 = 14, \omega_2 = 0.15$, initial Pearson correlation $\rho = 0.6$, and $p(b_{i1} = 1) = 0.3, p(b_{i2} = 1) = 0.4$. The scaled heaping happens on multiples of $h = 4$.

- Y_1 and Y_2 both follow a zero-inflated Poisson distribution with $\mu_1 = 15, \omega_1 = 0.3, \mu_2 = 31, \omega_2 = 0.3$, initial Pearson correlation $\rho = 0.6$, and $p(b_{i1} = 1) = 0.35, p(b_{i2} = 1) = 0.25$. The scaled heaping happens on multiples of $h = 4$.

We simulate 100 data sets for each scenario. Clayton copula function is used for the bivariate distribution, since the dependency parameter of it has the simplest Kendall's τ_K expression, $\tau_K = \frac{\theta}{\theta+2}$ (Naifar, 2011).

The 95% coverage probabilities of the nominal 95% confidence intervals for the interval-censored heaping are 92.08%, 91.67%, and 90.00% respectively, excluding the dependency parameter. The 95% coverage probabilities are 92.11%, 94.86%, and

Table 2.2 Simulation results with 30 simulations, interval-censored heaping.

True value	Bias	SD	True value	Bias	SD	True value	Bias	SD
$\omega_1 = 0.15$	0.009	0.006	$\omega_1 = 0.15$	0.009	0.006	$\omega_1 = 0.15$	0.007	0.005
$\omega_2 = 0.2$	0.007	0.006	$\omega_2 = 0.2$	0.007	0.006	$\omega_2 = 0.2$	0.008	0.006
$p_1 = 0.3$	0.037	0.094	$p_1 = 0.4$	0.044	0.022	$p_1 = 0.35$	0.040	0.020
$p_2 = 0.25$	0.021	0.129	$p_2 = 0.2$	0.015	0.011	$p_2 = 0.35$	0.036	0.018
$\mu_1 = 14$	0.133	0.138	$\mu_1 = 14$	0.174	0.133	$\mu_1 = 14$	0.168	0.110
$\alpha_1 = 3$	0.135	0.102	$\alpha_1 = 3$	0.140	0.153	$\alpha_1 = 3$	0.121	0.091
$\mu_2 = 18$	0.181	0.018	$\mu_2 = 18$	0.139	0.127	$\mu_2 = 18$	0.173	0.120
$\alpha_2 = 3$	0.142	0.011	$\alpha_2 = 3$	0.161	0.144	$\alpha_2 = 3$	0.165	0.108

Table 2.3 Simulation results with 100 simulations, scaled heaping.

True value	Bias	SD	True value	Bias	SD	True value	Bias	SD
$\omega_1 = 0.15$	0.014	0.011	$\omega_1 = 0.2$	0.015	0.011	$\omega_1 = 0.3$	0.017	0.013
$\omega_2 = 0.15$	0.014	0.011	$\omega_2 = 0.15$	0.014	0.009	$\omega_2 = 0.3$	0.016	0.013
$p_1 = 0.25$	0.027	0.021	$p_1 = 0.3$	0.023	0.018	$p_1 = 0.35$	0.031	0.025
$p_2 = 0.25$	0.028	0.021	$p_2 = 0.4$	0.027	0.022	$p_2 = 0.25$	0.027	0.022
$\mu_1 = 19$	0.219	0.174	$\mu_1 = 22$	0.209	0.166	$\mu_1 = 15$	0.221	0.171
$\mu_2 = 14$	0.180	0.143	$\mu_2 = 14$	0.182	0.134	$\mu_2 = 31$	0.288	0.252
$\theta = 0.226$	0.083	0.052	$\theta = 0.129$	0.066	0.049	$\theta = 0.088$	0.087	0.065

90.00% respectively for the scaled heaping simulations. Table 2.2 and Table 2.3 show the overall estimation bias of the parameters, with their standard deviations.

Table 2.4 Simulation results under scaled heaping setting 1, comparing IFM with MLE.

True value	Difference	SD
$\omega_1 = 0.15$	0.009	0.009
$\omega_2 = 0.15$	0.007	0.008
$p_1 = 0.25$	0.016	0.023
$p_2 = 0.25$	0.009	0.012
$\mu_1 = 19$	0.080	0.071
$\mu_2 = 14$	0.074	0.064
$\theta = 0.226$	0.058	0.083

As shown in Table 2.4, the absolute difference between the IFM method the MLE method is small, demonstrating the effectiveness of the IFM.

2.4 Real Data Analysis

We are going to use Global Adult Tabacco Survey (GATS). GATS is a nationally representative household survey that was launched in February 2007 as a new component of the ongoing Global Tobacco Surveillance System. It consists of demographic information, smoking behavior, as well as other information. Bangladesh 2009 data will be used in this section.

Two questions are used for the development of our bivariate response, "*On average, how many Bidis do you currently smoke each day?*" (Y_1) and "*On average, how many manufactured cigarettes do you currently smoke each day?*" (Y_2). The Pearson correlation coefficient is -0.46. For simple demonstration, we use *age*, *sex*, and *residence* as the predictors for the count part. One predictor, *age*, is used for the zero-inflation part. Frank copula function is chosen for modeling the dependency between the two responses. For both responses, we assume a negative binomial distribution. There were $n = 423$ observations in the final analysis data set. Table 2.5 shows the frequencies of Y_1 and Y_2 respectively. Clearly, the frequencies are high at the multiples of 5. The results of our model, as well as the results from modeling them independently, are shown in Table 2.6.

Comparing the result of our bivariate model with the two univariate model results, the age effects in both of the zero-inflation parts are very similar. For the mean of the response Y_1, the residence does not have a significant effect in the bivariate model, while it's significant in the univariate model. All other significance levels are the same.

Eban

The Eban study is a multisite serodiscordant couples' intervention targeting to reduce unprotected vaginal sex within African American HIV-serodiscordant couples.

Table 2.5 Frequencies of the two responses.

Y_1	Frequency	Y_2	Frequency
0	**1126**	0	**703**
1	34	1	38
2	37	2	100
3	27	3	98
4	45	4	90
5	**53**	5	149
6	34	6	95
7	16	7	46
8	57	8	57
9	9	9	1
10	**92**	10	**304**
11	0	11	2
12	81	12	67
13	16	13	2
14	9	14	5
15	**66**	15	**73**
16	10	16	5
17	0	17	0
18	4	18	4
19	0	19	0
20	**63**	20	**128**
21	1	21	0
22	2	22	2
23	0	23	1
24	2	24	0
25	**210**	25	7
...	...	...	...

Couples were randomly assigned to the treatment and the placebo groups. Sexual behaviour questions were asked at three follow-up times. We are going to use responses 3 months after the intervention. In this application, Y_1 is the answer from males to the question of number of unprotected vaginal sex episodes in the last 90 days, and Y_2 is the answer from females to the same question. Note that in a given observation, Y_1 and Y_2 are the male/female in a specific serodiscordant couple. Thus, responses are expected to be similar.

Both responses are assumed to follow a negative binomial distribution. Three

Table 2.6 Bivariate Negative Binomial model with Frank copula and corresponding independent models, Global Adult Tabacco Survey Bangladesh 2009.

	Bivariate Model				Univariate Model			
	Y_1		Y_2		Y_1		Y_2	
Predictors	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	2.70	$(2.47, 2.94)$	2.11	(1.93, 2.28)	2.62	$(2.43, 2.81)$	2.17	$(2.03, 2.31)$
Age	0.00	$(-0.01, 0.00)$	0.00	(-0.00, 0.01)	0.00	$(-0.01, 0.02)$	0.00	$(-0.00, 0.01)$
Sex	-0.71	$(-1.03, -0.34)$	-1.36	(-2.34, -0.72)	-0.75	$(-0.97, -0.53)$	-1.80	$(-2.31, -1.28)$
Residence	0.08	$(-0.06, 0.21)$	-0.12	(-0.22, -0.03)	0.11	$(0.01, 0.22)$	-0.22	$(-0.30, -0.13)$
α	1.95		1.75		2.34		2.36	
Zero Inflation								
Intercept	1.32	$(0.98, 1.62)$	-2.52	(-2.93, -2.01)	1.37	$(1.07, 1.67)$	-2.29	$(-2.64, -1.94)$
Age	-0.03	$(-0.03, -0.02)$	0.04	(0.03, 0.04)	-0.03	$(-0.03, -0.02)$	0.04	$(0.03, 0.04)$

variables are used for modeling the mean of the count. For Y_1, male age (continuous), treatment status (binary), and marital status (binary) are the predictors. For Y_2, the age is substitute with female age. The zero-inflation part has only one predictor, treatment status. The Clayton copula is used for modeling the positive dependency between the responses. The result is shown in Table 2.7.

Again, most of the results are similar. Using a bivariate model, we detect a significant age effect on the number of unprotected sex among women. Moreover, we detect a significant effect of treatment on increasing the probability of not having unprotected sex at all. Note that the wide confidence intervals in the zero-inflation part might be due to the bootstrap random sampling.

2.5 Conclusion and Discussion

In this chapter, we proposed a way to handle bivariate heaping data. We discussed both interval-censored heaping and scaled heaping scenarios. Copula functions are used for establishing the joint distribution of two responses.

When the data are interval heaped, directly applying copula functions might result in unreliable estimates. This is due to the fact that the use of copula functions is based on marginal cumulative functions. For each value that falls on the multiples of h, the likelihood of that observation becomes larger than what it should have been.

Hence, we developed an algorithm of multiple imputation to improve the estimation accuracy. Bootstrapping was used to obtain the standard error estimates. In our simulation studies, we showed the coverage probability under 30 simulation data sets for three scenarios, which have not been shown in some of the current literature. The less than 95% coverage probability might be due to the small number of simulations. However, a model for the heaping probability might improve the performance of our method.

Scaled heaping data are easier to address. It is a special case of mixture of

Table 2.7 Bivariate Negative Binomial model with Clayton copula and corresponding independent models, Eban Study.

	Bivariate Model				Univariate Model			
	Y_1		Y_2		Y_1		Y_2	
Predictors	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Intercept	0.71	$(-0.65, 1.83)$	2.51	(1.04, 4.05)	0.60	$(-1.06, 2.26)$	2.52	$(1.03, 4.02)$
Age	-0.01	$(-0.03, 0.02)$	-0.04	(-0.07, -0.02)	-0.01	$(-0.04, 0.02)$	-0.04	$(-0.07, 0.01)$
Married	-0.66	$(-1.13, -0.22)$	-0.12	(-0.63, 0.36)	-0.70	$(-1.17, -0.23)$	-0.02	$(-0.52, 0.48)$
Treatment	-0.99	$(-1.48, -0.54)$	-0.61	(-1.20, -0.10)	-1.02	$(-1.52, -0.51)$	-0.6	$(-1.12, -0.08)$
α	0.34		0.31		0.29		0.26	
Zero Inflation								
Intercept	1.89	$(-1.06, 11.39)$	1.06	(-2.04, 9.69)	0.60	$(-0.80, 2.00)$	0.29	$(-1.14, 1.73)$
Treatment	2.48	$(0.07, 12.12)$	1.96	(-0.17, 10.94)	1.03	$(-0.28, 2.34)$	0.83	$(-0.43, 2.10)$

distributions. The direct modeling of the mixture probabilities is feasible. The usual logistic regression was used for modeling the heaping probability and the zero-inflation probability. Based on the simulations results, we note that the coverage probabilities are acceptable but not satisfying. This raises the question of how well the copula functions can perform when the marginals are from mixture distributions. Future studies are needed to answer this question.

Chapter 3

Bayesian variable selection for correlated data: a factor analysis approach

In this chapter we present a method for the selection of grouped correlated predictors. With sparse Bayesian infinite factor modeling as the cornerstone, we combine a classic Bayesian variable selection prior. More specifically, a normal mixture distribution is used for the factor loading elements corresponding to the response variable. The latent factor selection probabilities are modeled using Bernoulli distributions. These selection probabilities admit sparse Bayesian infinite factor models, and the number of latent groups is determined automatically while the selection is being completed. Gibbs sampling is used to generate the posterior distribution (Geman and Geman, 1984). Our method provides a way of selecting correlated variables as a group. Moreover, our method accommodates binary variables, which is applied to the previously studied colon cancer data.

3.1 Introduction

In studies involving large number of predictors p, it is always crucial to only select the relevant explanatory variables in the final model. Especially in the gene selection problems, we usually observe large amount of genes p, compared to the total number of subjects n in the study. This is known as the "Large p, small n" problem (West, 2003). When the number of observations is smaller than the number of predictors, ordinary least squares would fail to provide valid results, since $X^T X$ is degenerated. Moreover,

the genes sharing the same pathway are known to be highly correlated. Therefore, in order to select the most significant genes, we have to overcome the high correlations among predictors, since the classic variable selection assumes independence.

There are two issues that need to be addressed when one wants to select the important predictors with high correlations. One is grouping correlated predictors, such that, fewer lower-dimensional independent factors can represent the original high-dimensional correlated explanatory variables. Moreover, the independence among factors make the usual variable selection technique applicable in this originally correlated setting. The second issue is how to select the factors, which represent the original predictors. Given that the number of predictors is large, for example, $p = 3000$, even after grouping, the number of groups can still be relatively large. The selection of the important groups is necessary. Here, our main focus is to select correlated variables together, instead of focusing on individual variables.

The ideal correlated variable selection method should be able to eliminate non-significant variables and simultaneously select whole groups of correlated variables. One of the most classic methods in dealing with variable selection for correlated data is elastic net (Zou and Hastie, 2005), which is an extension of the Lasso estimate (Tibshirani, 1996; Zou, 2006). This method provides a state-of-art solution from the frequentist point of view. It has the built-in ability for variable selection, meanwhile, allows the correlated variables to have similar coefficients. The corresponding Bayesian Lasso and Bayesian elastic net have been proposed, which have shown that the marginal posterior mode of the regression coefficients is equivalent to estimates given by the Lasso and elastic net (Park and Casella, 2008; Li and Lin, 2010).

While there is a rich literature on grouping and clustering similar observations (Tadesse et al., 2005; Kim et al., 2006), especially in the unsupervised machine learning field, the grouping of similar variables have been studied less often. One popular method for addressing this topic is factor analysis in both frequentist and Bayesian

settings. The correlations between variables are represented by the factor loading matrix. In this chapter, we are going to use Bayesian factor models to group the correlated variables.

Nonparametric Bayesian methods for grouping correlated predictors have also been used with Indian buffet process and Dirichlet process (Knowles and Ghahramani, 2007; P. et al., 2012). In both papers, they elegantly group observed correlated variables together with the number of latent factors unspecified. The difference is that when using the India buffet process, predictors can load on multiple latent factors, while Dirichlet process only allows one group for each predictor. These methods focused on the flexibility of the number latent factors, as well as the grouping of correlated variables. The sparse Bayesian infinite factor models have been proposed for allowing flexible latent factor number as well (Bhattacharya and Dunson, 2011). In their paper, they used a shrinkage prior that can automatically discard unimportant latent factors, while simultaneously grouping the correlated variables.

Strategies have been discussed and compared for Bayesian spike and slab variable selection with classic methods (Ishwaran and Rao, 2005). They pointed out with increasing number of variables, the variable selection task becomes difficult when using spike and slab approaches. This again confirms the necessity of reducing the number of variables in order to have decent performance of variable selection. The original "spike and slab"is referred to a mixture distribution of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike) (Lempers, 1971; Mitchell and Beauchamp, 1988). In this chapter we are going to apply a well-known version of the spike and slab prior (George and McCulloch, 1993) on the latent factors, which assumes two mixture of the scaled normal distributions. Some other analogous constructions of the spike and slab prior, such as the mixture of point mass function with a normal distribution have also been commonly used in the literature (Kim et al., 2006; Bobb et al., 2015).

The discussion of “Large p, Small n ” problems has been growing quickly in recent years, including the elastic net mentioned before (Zou and Hastie, 2005). Using frequentist methods, some of the literature discussed how to control the type I error rate α in high-dimensional data (Wasserman and Roeder, 2009). Others using penalized methods studied the oracle properties, and consistency in group identification (Hoerl and Kennard, 1970; Tibshirani, 1996; Zou, 2006; Sharma et al., 2013).

Bayesin factor regression models have also been used in addressing "Large p, Small n" problems (West, 2003). The spike and slab prior has been assigned on every element of the factor loading matrix in the Bayesian factor model. Later, another study applied a similar method with different priors on gene expression data sets (Carvalho et al., 2012). One study has suggested to use a correlation-based stochastic search method on this type of data, which is an extension of the popular stochastic search variable selection (SSVS) (Kwon et al., 2011). Although the method that we are going to present is also covariance-based, the fundamental philosophy of our method is that we assume the correlations between predictors are induced by the latent sources, which is different from solely correlation-based method.

Despite the popularity of Bayesian factor regression for handling "Large p, Small n" problems, few of them allowing infinite number of latent factors. Moreover, the latent factor selection under this indefinite latent factor number assumption has not been discussed to our knowledge. Therefore, the sparse Bayesian infinite factor model will be the cornerstone of this chapter (Bhattacharya and Dunson, 2011). This method can group correlated variables using factor loading matrix, and automatically make the adaption to shrink the number of latent groups. The authors have also proved the order independence of their factor loading matrix and the posterior distributions have explicit forms. Therefore, Gibbs sampling can be directly used and the posterior sampling is stable (Geman and Geman, 1984).

In this chapter, we are going to extend the sparse Bayesian infinite factor model

to incorporate variable selection probabilities. The method is presented in the next section. We will introduce the sparse Bayesian infinite factor model first, and then combine it with Bayesian selection priors. Prior and Posterior distributions will be presented, as well as one way of selecting the original predictors based on the selection of latent factors. Simulation studies are done in Chapter 3.3, and a real data analysis is presented in Chapter 3.4. We will conclude this chapter in Chapter 3.5.

3.2 Methods

A typical latent factor model has the form

$$y_i = \Lambda\eta_i + \epsilon_i, \epsilon_i \sim N_{p+1}(0, \Sigma), \tag{3.1}$$

where $i = 1, ..., n$ is the indicator for observations, and y_i is the $(p+1)$-dimensional continuous variable, with the first element y_{i1} being the response, $y_{i2}, ..., y_{i(p+1)}$ being the explanatory variables. All y_is consist of the data matrix $Y = (y_1, ..., y_n)^T$, which has dimension $n \times p$. Without loss of generality, we assume all columns of Y have been centered and scaled before analysis. η_i is the $k \times 1$ latent factors, that are independently distributed between observations with $N_k(0, I_k)$. Λ is a $(p+1) \times k$ factor loading matrix, and ϵ_i is an idiosyncratic error with variance-covariance matrix $\Sigma = diag(\sigma_1^2, ..., \sigma_{p+1}^2)$. In the next chapter, we would relax the assumption for the error term of the response, ϵ_{i1}. By marginalizing the latent factors, the distribution of y_i becomes $N_p(0, \Omega), \Omega = \Lambda\Lambda^T + \Sigma$.

The number of latent factors k is usually unknown. One advantage of the sparse Bayesian infinite factor model is that it can group correlated variables into latent groups, at the same time, automatically update the number of latent factors k. The idea is to use a shrinkage prior for the latent factor loading element λ_{jk}, such that the number of non-zero elements in Λ is decreasing as k increases. We assign the factor loading element λ_{jk} a normal prior with mean 0 and precision $\phi_{jh}\tau_h$, where

$j = 2, ..., p+1, h = 1, ..., k$. Both ϕ_{jh} and τ_h have their own prior distributions, and the prior distribution of τ_h depends on h. Hence, as h increases, τ_h can increase stochastically. The elements of the factor loading matrix are then generated from a small neighborhood of zero as the factor number increases.

Under their working frame, we modify the prior distribution for the first row of Λ. The first row of Λ, $\lambda_{1\cdot}$, represents the relationship between the response and the k latent factors. For each element of $\lambda_{1h}, h = 1, ..., k$, we assume the prior distribution is,

$$\lambda_{1h}|\phi_{1h}, \gamma_h \sim (1-\gamma_h)N(0, c_{1h}\phi_{1h}^{-1}) + \gamma_h N(0, c_{2h}\phi_{1h}^{-1}), \tag{3.2}$$

where γ_h is an indicator variable with $P(\gamma_h = 1) = 1 - P(\gamma_h = 0) = p_h$, and p_h is the selection probability of latent factor η_h. This is the selection prior suggested in previous studies (George and McCulloch, 1993). The constants c_{1h} and c_{2h} are chosen such that c_{1h} is small c_{2h} is relatively large. Therefore, if η_h is selected, i.e. $\gamma_h = 1$, we would sample from the prior $N(0, c_{2h}\phi_{1h}^{-1})$, which has a large variance. Hence, it covers a wide range of the real line. If η_h is not selected, i.e. $r_h = 0$, the factor loading element λ_{1h} is then sampled from $N(0, c_{1h}\phi_{1h}^{-1})$, which has a small variance when c_{1h} is small.

To automatically adapt the number of latent factors, the suggested way is to choose the probability of adapting at $p = \frac{1}{exp(1+0.0005i)}$, where i here is the iteration number (Bhattacharya and Dunson, 2011). They have stated this probability has been chosen to satisfy the diminishing adaption condition theorem in a previous study (Roberts and Rosenthal, 2007). We will follow their suggestion, and make the adaption based on the magnitude of the factor loadings. When every element in that column is smaller than a predefined small number, e.g. $|\lambda_{jk}| < 0.001$, this redundant column is dropped and the latent factor number is reduced by 1. On the hand, the number of latent factors would be increased by one if any element of the factor loading at iteration i is greater than 0.001.

When the response is binary, we use data augmentation and create a latent continuous variable Z_i that connects the observed binary response y_{i1} with the latent factors η_i, where $Z_i \sim N(\lambda_{1.}\eta_i, 1)$. The generation of the latent continuous variable Z_i is based on the value of observed Y_is (Albert and Chib, 1993). That is, generate $Z_i > 0$ from a truncated normal distribution from below if $Y_i = 1$, and generate $Z_i < 0$ from a truncated normal distribution with upper limit 0 if $Y_i = 0$. Thus, the posterior joint distribution of $\lambda_{1.}$ and $Z = (Z_1, ..., Z_n)$ is

$$\pi(\lambda_{1.}, Z|y) = C\pi(\lambda_{1.}) \prod_{i=1}^{n} \{I(Z_i > 0)I(y_i = 1) + I(Z_i \leq 0)I(y_i = 0)\} \times \pi(Z_i) \quad (3.3)$$

The posterior sampling of the truncated normal distribution is done using exponential rejection sampling (Geweke, 1991). After obtaining the continuous variable Z_is, the latent factor loading method can be applied as usual.

The prior specifications are shown below,

- $\eta_i \sim N_k(0, I_k)$
- $\sigma_j^{-2} \sim Gamma(a_\sigma, b_\sigma), j = 1, ..., p, p+1$, where σ_j^2 is the diagonal element of Σ in Equation 3.1.
- $\lambda_{jh}|\phi_{jh}, \tau_h \sim N(0, \phi_{jh}^{-1}\tau_h^{-1}), j = 2, ..., p, p+1,\ h = 1, ..., k$
- $\lambda_{1h}|\phi_{jh}, \gamma_h, c_{1h}, c_{2h} \sim (1 - \gamma_h)N(0, \phi_{jh}^{-1}c_{1h}) + \gamma_h N(0, \phi_{jh}^{-1}c_{2h})$
- $r_h \sim Bernoulli(\pi_h)$, where $\pi_h = 0.5$ is the uniform or "indifference" prior for selection probabilities.
- c_{1h} and c_{2h} are shrinkage parameters, e.g. $c_{1h} = 0.1, c_{2h} = 1$, or $c_{1h} = 0.1^{h-1}, c_{2h} = 0.9^{h-1}$
- $\phi_{jh} \sim Gamma(\nu/2, \nu/2)$
- $\tau_h = \prod_{l=1}^{h} \delta_l, \delta_1 \sim Gamma(a_1, 1), \delta_l \sim Gamma(a_2, 1)$, for $l \geq 2$

The parameter τ_h is the shrinkage parameters mentioned before for automatically decreasing the dimension of the factor loading matrix, as the number of latent factors k increases.

One can use the posterior mean of the selection probabilities to decide which latent factors to select in order to predict the response. The selection probabilities and the factor loading elements give a general idea of which original variables are of importance. Depending on the ideal size of the final model, we suggest to combine the posterior mean of selection probabilities with the quantiles of the posterior mean of the factor loading matrix. First, given the posterior mean of selection probabilities of the latent factors. A cut-off probability of 0.5 will be used in this chapter. That is, we first select the latent factor groups based on the posterior mean of the selection probabilities. Then, one can examine the factor loading columns that are corresponding to these selected latent factors. Given the posterior mean of the factor loading matrix $\hat{\Lambda}$, one can select the original predictors which have large factor loading in absolute values. The definition of "large" depends on different situations. One can decide based on the target final model size. Some ad hoc methods can be applied, such as trying different quantile cut-off point, and selecting the one with least mean square error.

Gibbs sampling is used to generate samples from posterior distributions. Suppose the number of latent factors is k^* at the current iteration. We use $\pi(x|-)$ to denote the conditional distribution on all other variables. The posterior distributions are,

- Sample $\eta_i, i = 1, ..., n$, from conditionally independent posteriors

$$\pi(\eta_i|-) \sim N_{k^*}\{(I_{k^*} + \Lambda_{k^*}^T \Sigma^{-1} \Lambda_{k^*})^{-1} \Lambda_{k^*}^T \Sigma^{-1} y_i, (I_{k^*} + \Lambda_{k^*}^T \Sigma^{-1} \Lambda_{k^*})^{-1}\}$$

- Sample $\sigma_j^{-2}, j = 1, ..., p, p+1$, from conditionally independent posteriors

$$\pi(\sigma_j^{-2}|-) \sim Gamma\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \lambda_j^T \eta_i)^2\right)$$

- Sample each row of Λ_{k^*}, $\lambda_j^T, j = 2, ..., p, p+1$ from conditionally independent posteriors (given the latent factors, the predictors are independent, hence each row of Λ_{k^*} are independent)

$$\pi(\lambda_j|-) \sim N_{k^*}\{(D_j^{-1} + \sigma_j^{-2}\eta^T\eta)^{-1}\eta^T\sigma_j^{-2}y^{(j)}, (D_j^{-1} + \sigma_j^{-2}\eta^T\eta)^{-1}\},$$

  where η is the $n \times k^*$ latent factor matrix, with each row corresponding to each observation, $D_j = diag(\phi_{j1}^{-1}\tau_1^{-1}, ..., \phi_{jk^*}^{-1}\tau_{k^*}^{-1})$, and $y^{(j)} = (y_{1j}, ..., y_{nj})^T$, for $j = 2, ..., p+1$.

- Sample the first row of Λ_{k^*}, λ_1^T in a similar fashion, just replace D_j with $D_y = diag\{\phi_{11}^{-1}((1-r_1)c_{11} + r_1c_{21}), ..., \phi_{1k^*}^{-1}((1-r_{k^*})c_{1k^*} + r_{k^*}c_{2k^*})\}$.

- Sample $\phi_{jh}, j = 2, ..., p, p+1,\ h = 1, ..., k^*$ from

$$\pi(\phi_{jh}|-) \sim Gamma\left(\frac{\nu+1}{2}, \frac{\nu + \tau_h\lambda_{jh}^2}{2}\right).$$

- Sample $\phi_{1h}, h = 1, ...k^*$ from

$$(1-r_h)Gamma\left(\frac{\nu+1}{2}, \frac{\nu + \lambda_{1h}^2/c_{1h}}{2}\right) + r_hGamma\left(\frac{\nu+1}{2}, \frac{\nu + \lambda_{1h}^2/c_{2h}}{2}\right).$$

- Sample $\pi_h, h = 1, ..., k^*$, which is the probability of $\pi(r_h = 1)$ from

$$\pi(r_h = 1|-) = \frac{\pi(\lambda_{1h}|r_h = 1, ...)}{\pi(\lambda_{1h}|r_h = 1, ...) + \pi(\lambda_{1h}|r_h = 0, ...)}.$$

- Sample δ_1 from

$$\pi(\delta_1|-) \sim Gamma\left\{a_1 + \frac{pk^*}{2}, 1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau_l^{(1)}\sum_{j=2}^{p+1}\phi_{jl}\lambda_{jl}^2\right\},$$

  and sample $\delta_h, h \geq 2$ from

$$\pi(\delta_h|-) \sim Gamma\left\{a_2 + \frac{p}{2}(k^* - h + 1), 1 + \frac{1}{2}\sum_{l=h}^{k^*}\tau_l^{(h)}\sum_{j=2}^{p+1}\phi_{jh}\lambda_{jl}^2\right\},$$

  where $\tau_l^{(h)} = \prod_{t=1, t\neq h}^{l}\delta_t$ for $h = 1, ..., k^*$.

- In the case when y_{i1} is binary, sample $Z_i, i = 1, ..., n$ from $Z_i \sim N(\lambda_1.\eta_i, 1)$ truncated at left by 0 if $y_{i1} = 1$. Sample $Z_i \sim N(\lambda_1.\eta_i, 1)$ truncated at right by 0 if $y_{i1} = 0$. For the sampling of Λ and η, replace the corresponding elements in y_i respectively.

Other than the response, we assume continuous distributions for all the predictors. However, the predictors can be binary as well. The extension to allow multinomial distribution is straightforward (Albert and Chib, 1993). In the next sections, we are going to present a simulation study assuming all variables are continuous. The performance of our methods when the response is binary is examined in the Chapter 3.4 real data application.

3.3 Simulation Studies

To examine the performance of our methods, simulation studies were done under three scenarios. For each simulation setting, we generate $n = 200$ subjects, with the first 100 subjects as the training set and the rest of the 100 subjects as the testing set. Thus, the modeling is based on 100 observations. We compare the mean squared prediction error from our methods with the results from Lasso and elastic net methods, which are readily available in R package *glmnet*. The three simulation settings are,

- Setting 1, $p = 100, k = 5$, 2 out of 5 latent factors predict the response.
- Setting 2, $p = 300, k = 7$, 4 out of 7 latent factors predict the response.
- Setting 3, $p = 500, k = 10$, 4 out of 10 latent factors predict the response.

We generate non-zero elements of the factor loading matrix Λ from a normal distribution with mean 0 and variance 9 for all three settings. And the location of these non-zero elements are randomly selected out of the p rows for each column.

For setting 1, the numbers of non-zero columns from left to right are $10, 9, 8, 7, 6$, respectively. For setting 2, the numbers are $14, 11, 8, 9, 10, 13, 12$. For setting 3, they are $20, 15, 11, 12, 16, 13, 17, 18, 18$. The number of non-zero elements for each column is between $2k$ to $k+1$. The diagonal elements of Σ, $\sigma_1^2, ..., \sigma_{p+1}^2$ are 0.01 in all three settings. Note that, since the location of the non-zero factor loading is randomly selected. It is possible that a predictor has loading on more than one latent factor.

The parameters of the prior distributions are, $\sigma_j^{-2} \sim Gamma(a_\sigma = 1, b_\sigma = 0.3)$, where a_σ is the shape parameter and b_σ is the rate parameter. For the parameter for ϕ_{jh}, i.e. ν, we randomly sample a number from $unif(2, 4)$. The a_1 and a_2 for the shrinkage parameters δ_1 and $\delta_h, h = 1, ..., k$ are randomly sampled from $unif(1.1, 3.1)$ and $unif(2.1, 4.1)$, respectively. We assign $c_{1h} = 0.1, c_{2h} = 1$. We generate 25000 posterior samples with the first 5000 as the burn-in, and collect every 5th sample.

In Table 3.1 and Table 3.2, we illustrate the prediction and estimation performances of this methods. We report the mean square prediction error (MSPE), absolute average prediction error (AAPE), and maximum absolute prediction error (MAPE) in Table 3.1. The MAPE can be seen as a measure for the worst performance of the methods. In Table 3.2, we report the mean square error (MSE), absolute average bias (AAB), and maximum absolute bias (MAB) for the estimated β, that is, the original coefficients. Overall, the performance is quite well.

The variable selections using latent factor selection probabilities and the quantiles of the factor loadings are shown in Tables 3.3. The results from different cut-off quantiles are for comparison. Under the simulation setting 1, the 92.5th quantile cut-off point has the best false positive and true positive, i.e. power performance. The advantage of elastic net is not obvious in this case. Under the second setting, the 96.5th quantile cut-off point and the elastic net has similar performance. For the last simulation setting, $p = 500, k = 10$, the 97.5th quantile cut-off point has acceptable performance, though the false positive rate is relatively high. Note that, the Lasso

Table 3.1 Simulation study results based on $N = 100$ simulation data sets, mean squared prediction error (MSPE), average absolute prediction error (AAPE), and maximum absolute prediction error (MAPE) of Y.

Setting 1	Y	Proposed Method	Lasso	Elastic Net
MSPE	mean	0.0116	0.0144	0.0147
	min	0.0080	0.0087	0.0093
	max	0.0151	0.0219	0.0235
AAPE	mean	0.0860	0.0953	0.0967
	min	0.0723	0.0748	0.0771
	max	0.1010	0.1210	0.1188
MAPE	mean	0.2929	0.3251	0.3306
	min	0.2047	0.1989	0.2315
	max	0.4412	0.4780	0.5260
Setting 2	Y	Proposed Method	Lasso	Elastic Net
MSPE	mean	0.0305	0.0179	0.0179
	min	0.0088	0.0105	0.0108
	max	0.9477	0.0320	0.0317
AAPE	mean	0.1022	0.1063	0.1065
	min	0.0769	0.0810	0.0826
	max	0.7708	0.1372	0.1382
MAPE	mean	0.3471	0.3663	0.3736
	min	0.2283	0.2469	0.2417
	max	2.7545	0.5017	0.6131
Setting 3	Y	Proposed Method	Lasso	Elastic Net
MSPE	mean	0.0987	0.0194	0.0185
	min	0.0078	0.0107	0.0094
	max	2.0050	0.0317	0.0317
AAPE	mean	0.1439	0.1111	0.1081
	min	0.0724	0.0799	0.0755
	max	1.1605	0.1466	0.1424
MAPE	mean	0.4845	0.37703	0.3684
	min	0.2149	0.2554	0.2596
	max	3.3049	0.5330	0.5280

performs the worst under all three settings.

3.4 Real Data Analysis

We apply our method on the Alon colon cancer data set that was used in some previous studies (Alon et al., 1999; Yang and Song, 2010). This data set contains $p = 2000$ genes and $n = 62$ observations in total. We use the first 40 observations as

Table 3.2 Simulation study results based on $N = 100$ simulations, mean square error (MSE), average absolute bias (AAB), and maximum absolute bias (MAB) of the original β.

Setting 1	β	Proposed Method	Lasso	Elastic Net
MSE	mean	0.0007	0.0013	0.0015
	min	0.0006	0.0004	0.0008
	max	0.0008	0.0033	0.0034
AAB	mean	0.0104	0.0122	0.0143
	min	0.0091	0.0070	0.0104
	max	0.0140	0.0248	0.0257
MAB	mean	0.1175	0.2055	0.1896
	min	0.0990	0.1192	0.1179
	max	0.1496	0.4479	0.3376
Setting 2	β	Proposed Method	Lasso	Elastic Net
MSE	mean	0.0001	0.0003	0.0001
	min	0.0001	0.0003	0.0004
	max	0.0006	0.0014	0.0010
AAB	mean	0.0042	0.0073	0.0051
	min	0.0036	0.0049	0.0034
	max	0.0055	0.0103	0.0107
MAB	mean	0.0635	0.2158	0.1699
	min	0.0539	0.0997	0.0796
	max	0.3424	0.4110	0.3718
Setting 3	β	Proposed Method	Lasso	Elastic Net
MSE	mean	0.0001	0.0006	0.0003
	min	0.0001	0.0002	0.0001
	max	0.0050	0.0016	0.0008
AAB	mean	0.0030	0.0052	0.0043
	min	0.0024	0.0032	0.0028
	max	0.0103	0.0091	0.0072
MAB	mean	0.0721	0.3013	0.2085
	min	0.0479	0.1041	0.0888
	max	1.4060	0.6868	0.4603

the training set and the rest of the 32 observations as the testing set. The response is binary with 0 being healthy patients, and 1 being diagnosed with colon cancer. The training set has 27 patients with colon cancer, and 13 healthy patients. The testing set has 13 patients with colon cancer, and 9 healthy patients.

We used the same priors as shown in the simulation studies, and we generated 25000 posterior samples with the first 5000 as the burn-in, and collected every 5th

Table 3.3 Percentages of false positives and true positives from Variable selection under three settings, with selected quantiles reported.

		Setting 1:	$p = 100$,	$k = 5$	
		Quantiles			
	90th	92.5th	95th	Lasso	Elastic net
False positives (%)					
mean	21.62	0.71	1.00	21.50	20.09
min	11.11	0.00	0.00	0.00	0.00
max	36.36	38.10	53.33	68.00	61.29
True positives (%)					
mean	96.88	93.56	62.19	53.25	80.50
min	87.50	81.25	43.75	25.00	68.75
max	100.00	93.75	62.50	81.25	100.00
		Setting 2:	$p = 300$,	$k = 7$	
		Quantiles			
	90th	92.5th	95th	Lasso	Elastic net
False positives (%)					
mean	19.95	5.35	5.31	13.85	9.79
min	6.25	0.00	0.00	0.00	0.00
max	34.48	13.79	12.07	64.10	50.91
True positives (%)					
mean	96.41	90.54	68.93	29.09	75.17
min	71.74	69.57	50.00	13.04	54.35
max	100.00	95.65	78.26	47.83	89.13
		Setting 3:	$p = 500$,	$k = 10$	
		Quantiles			
	90th	92.5th	95th	Lasso	Elastic net
False positives (%)					
mean	31.03	16.75	16.62	17.48	10.82
min	0.00	0.00	0.00	0.00	0.00
max	46.32	31.96	25.77	72.88	44.58
True positives (%)					
mean	96.23	87.52	69.82	27.51	76.52
min	52.31	41.54	32.31	12.31	61.54
max	100.00	95.38	78.46	41.54	84.62

sample. Our proposed method did not perform as well as Lasso and elastic net in classifying the 40 observations in the training set. It performed as well as elastic net in the prediction of Ys in the testing set. Figure 3.1 shows four selected elements in the factor loading matrix Λ, $\lambda_{1,1}, \lambda_{1,5}, \lambda_{1,8}$, and $\lambda_{1903,5}$. The chains have acceptable convergence.

Table 3.4 Summary of the colon cancer probability of successfully classification results for training set and testing set.

Method	Training	Testing
Proposed Method	33/40	17/22
Lasso	38/40	14/22
Elastic Net	39/40	17/22

Table 3.5 Top 32 selected genes for classification, based on 99.8th quantile cut-off point.

Gene number	Gene name	Gene number	Gene name
28	T63484	111	R78934
305	Z11584	455	L02426
482	D00761	538	R37428
612	R52000	629	T60318
709	T67921	712	T90036
739	X12369	744	X53004
758	T78104	785	D13627
806	X15882	834	U29092
840	X66975	1052	U02493
1053	T72599	1071	H40108
1170	X17644	1239	L37112
1296	X82166	1306	D17400
1386	L40992	1401	D13243
1434	R85479	1566	X07384
1601	U21914	1631	X63469
1642	L05485	1997	H18490

The 2000 genes have been grouped into 10 latent factor groups. Figures 3.4 to 3.4 show the posterior mean of the factor loading matrix $\Lambda_{(-1)}$, which does not include the first row. Figure 3.4 shows that most of the genes from gene number 1 to 100 have obvious loadings on the first 6 latent factors. Some of the genes between gene number 501 to 600 (Figure 3.4), 1001 to 1100 (Figure 3.4), and 1501 to 1600 (Figure 3.4) have negative loading on the latent factor 10. There are many genes from 1 to 100, and 1501 to 1600 have large loading on latent factor 1 (Figures 3.4 and 3.4).

In these figures, we can obtain the information of which variables "come from" the same "source". The top selected 32 genes based on the posterior means of selection probability and factor loading matrix are shown in Table 3.5. Most of the top

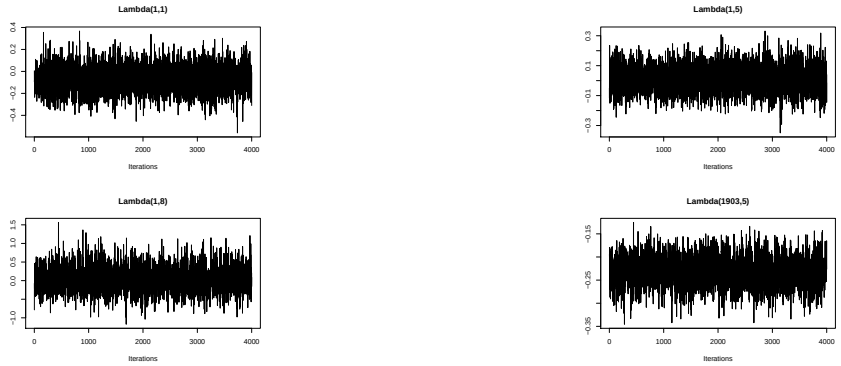

Figure 3.1 Selected factor loading elements, from left to right, first row to second row, $\Lambda_{1,1}, \Lambda_{1,5}, \Lambda_{1,8}$, and $\Lambda_{1903,5}$.

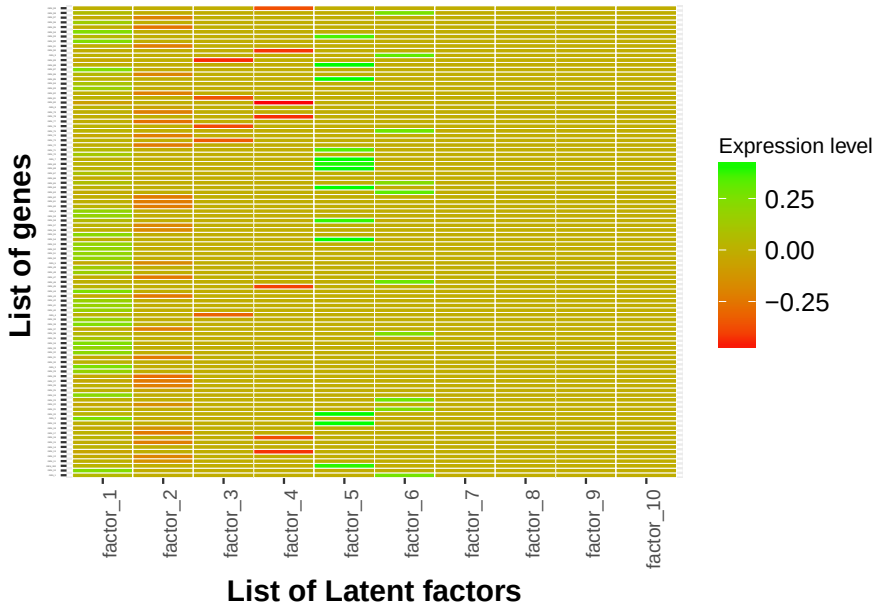

Figure 3.2 Posterior means of factor loading for genes 1 to 100 on the latent factors.

32 genes selected have been reported in previous studies. Genes T78104, D00761, X15882, X12369, X66975, R37428, X17644, D13627, U29092, U02493, and X63469 have been identified in a study for systems-level molecular mechanisms of tumorigenesis (Hernández et al., 2007). Other genes, for example, H18490, T90036, D17400, D13243, T60318, H40108, R78934, X12369, and X15882 have also been identified in other studies (Au et al., 2005; Li and Li, 2008; Rao and Dey, 2005; Shaik and M.,

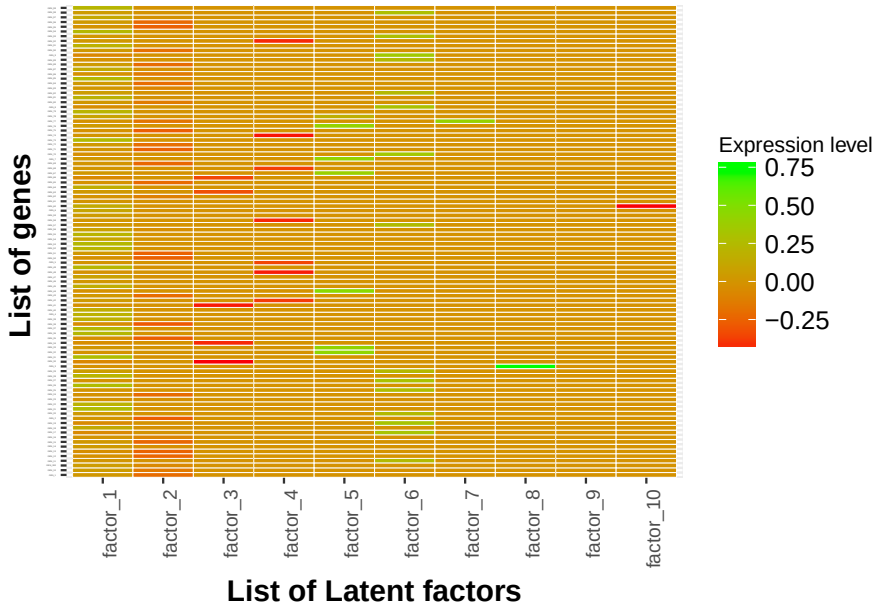

Figure 3.3 Posterior means of factor loading for genes 501 to 600 on the latent factors.

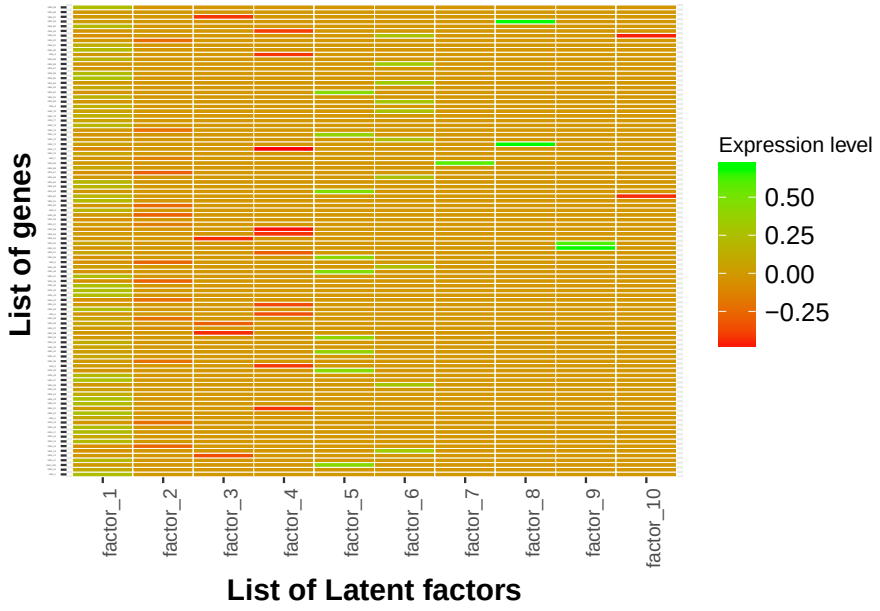

Figure 3.4 Posterior means of factor loading for genes 1001 to 1100 on the latent factors.

2007; Li and M., 2002).

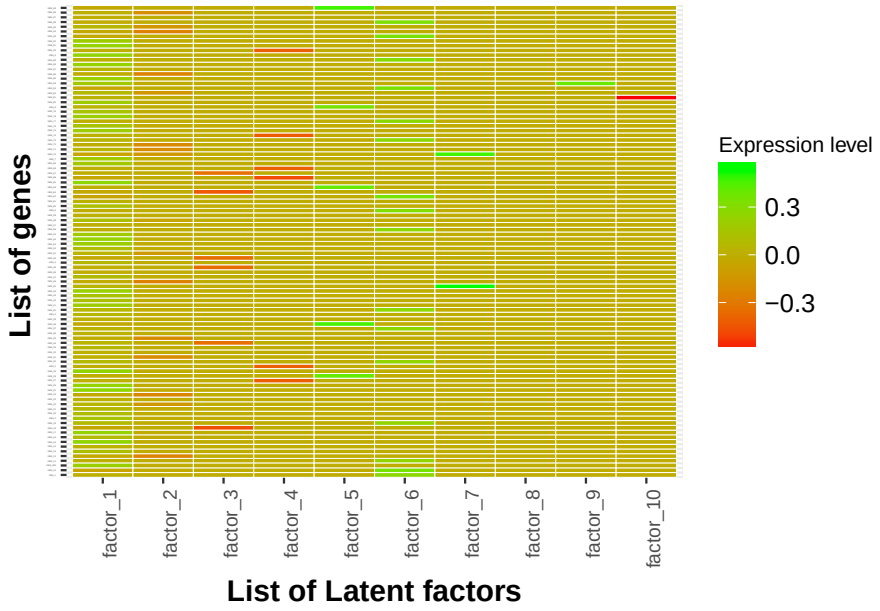

Figure 3.5 Posterior means of factor loading for genes 1501 to 1600 on the latent factors.

3.5 Conclusion and Discussion

In this chapter we proposed a method for selection of latent factors. More specifically, we innovatively combined the classic spike and slab priors with one of the newly developed sparse Bayesian infinite factor models. Our method is straightforward and easy to apply based on existing methods.

The proposed method has good performance when all the variables are from continuous normal distribution. More specifically, under the scenario when $p > n$, our method can select the predictors based on their latent groups. For now, the selection of the original predictors are based on a cut-off quantile that one sees to fit. In general, we suggest to examine the posterior means of selection probability and the factor loading matrix. By only evaluating the columns with larger than 0.5 selection probability, the number of the original predictors can be significantly reduced. Further, the predictors that have large loadings in the same column are in the same latent factor group. This can usually be observed directly.

The quantile cut-off point method is suggested for a direct method of variable selection. The choice of a specific cut-off point depends on the number of original explanatory variables, and the target number of variables to keep in the model. As shown in the real data analysis section, we reported the top 32 genes. Some studies on the same data set report top $16, 18, 20, 50$ genes. One can adjust the cut-off point according to the study goal.

Our proposed method has not outperformed existing method elastic net when dealing with binary response, as seen in the real data analysis. Since the updating process takes a long time, future study can focus on improving the efficiency when modeling non-normal distributions within the latent factor framework. Moreover, one might relax the distribution assumptions of the variables and allow Poisson, multinomial, Gamma distributions, as well as others. For example, when the response follows a multinomial distribution with g categories, one can define $g-1$ cut-off points for the latent continuous variable Z (Albert and Chib, 1993).

The important difference between our proposed method and existing methods is that we assume observed correlated variables come from some latent groups. The correlations between variables are induced by the same group membership. Therefore, comparing with the methods with shrinkage parameters or priors directly applied on the observed variables, our method provides additional information of the latent relationship between variables, while simultaneously select the important latent groups.

Chapter 4

Nonparametric Bayesian latent factor modeling

In this chapter, we explore the nonparametric Bayesian latent factor modeling. We built upon the sparse Bayesian factor models mentioned in Chapter 3, and utilized the Dirichlet process for the response variable in order to allow a non-normal distribution. To circumvent the non-closed form of the posterior distributions, Bayesian hierarchical models are applied in this chapter. Our objective is to group correlated variables, while simultaneously clustering similar observations into clusters.

4.1 Introduction

In this chapter, we are going to explore nonparametric Bayesian latent factor modeling. There are quite a few studies on using nonparametric assumptions for the number of latent factors, as mentioned in the last chapter. Their focus is on the factor loading matrix, or the number latent factors. However, none of them relaxes the normality assumption for the response variable.

There are two main motivations for assuming a nonparametric distribution for the response variable. First, using Gaussian process and Dirichlet process, one can achieve the classification of observations (Kim and Lee, 2007; Teh et al., 2005). Both of these processes are a distribution over distributions and have wide application in the machine learning area. Therefore, by imposing the nonparametric assumption, we can cluster similar observations while grouping correlated variables. Second, there are

circumstances when the normal assumption becomes too strict for certain variables. When the response exhibits heavy tails or multimodality, methods have been proposed to flexibly estimate the mean function meanwhile allowing the residual density to change nonparametrically with predictors (Pati and Dunson, 2014) . In this chapter, we are going to use the same framework, with the residual density being independent of the predictors.

For Bayesian latent factor regression, the common assumption of the distribution of all continuous variables is the normal distribution. Based on the methods from last chapter, we are going to relax the Gaussian assumption for the response, and to assume a nonparametric distribution for the response. More specifically, Dirichlet process is used for building the distribution of the response to allow a heavy tail distribution, as well as a multimodal feature.

The Dirichlet process is a distribution over distributions, which is commonly used in Bayesian nonparametric modeling. A Dirichlet process G which has base distribution G_0 and concentration parametrer α_0, is a probability measure over a measurable space (Θ, B). For any finite measureble partition $(A_1, ..., A_r)$ of Θ, the random vector $(G(A_1), ..., G(A_r))$ is distributed as a finite-dimentional Dirichlet distribution with parameters $(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r))$, i.e.,

$$(G(A_1), ..., G(A_r)) \sim Dir(\alpha_0 G_0(A_1), ..., \alpha_0 G_0(A_r)).$$

A detailed introduction of the Dirichlet process can be referred elsewhere (Teh, 2010; Teh et al., 2006).

The original construction of the Dirichlet process is cumbersome when used for generating posterior distributions. In this chapter, we are going to use the stick-breaking construction of Dirichlet process to accelerate posterior sampling (Ishwaran and James, 2001, 2003). By using a blocked Gibbs approach, one can avoid marginalizing over the prior (Ishwaran and James, 2001). Hence, it makes directly sampling of the nonparametric posterior possible. More specifically, the stick-breaking construc-

tion of Dirichlet process is defined as,

$$G = \sum_{q=1}^{\infty} \pi_q \delta_{\theta_q},$$

where $\pi_q = V_k \prod_{l=1}^{q-1}(1 - V_q), V_k \sim Beta(1, b_q), b_q = \alpha_0$, δ is the Dirac delta function, and $\theta_q \sim G_0$. The G_0 is the base function and $\{\theta_q\}_{q=1}^{\infty} \sim G_0$ independently. $\{\pi_q\}_{q=1}^{\infty}$ are the weights that sum up to 1.

In the next sections, we are going to introduce on how to apply the Dirichlet process assumption under the sparse Bayesian latent factor model framework. Then we will present a small simulation study, and one real data analysis. We conclude this chapter in the Chapter 4.5 with possible future work.

4.2 Methods

With similar notations as in chapter 3, we use Bayesian hierarchical model to separate the original model into two parts

$$\begin{aligned} y_i &= f_i + \epsilon_i \\ f_i &= \Lambda \eta_i + e_i \end{aligned} \tag{4.1}$$

where $i = 1, ..., n$ is the indicator for observation, and y_i is the $(p + 1)$-dimensional continuous variable vector, including the first term for the response. The difference between this model and the model in Chapter 3 is that we define an intermediate latent variable $f_i \sim N(\Lambda\eta_i, \Sigma_e)$, so that the latent factor model has two layers, and we separate the distribution of the error ϵ_i and the distribution of the factor loading matrix Λ apart.

The Gibbs sampler can be implemented independently given f_i. We assume e_i has the usual distribution, which is $e_i \sim N_p(0, \Sigma_e)$, where $\Sigma_e = diag(\sigma_1^2, ..., \sigma_{p+1}^2)$. The nonparametric distribution of the response y_{i1} is assumed through the first element of ϵ_i, that is ϵ_{i1}. For the rest of ϵ_i, $\epsilon_{i(-1)}$, that is ϵ_i without the first element, follows $N_p(0, \Sigma_{\epsilon_{i(-1)}})$, where $\Sigma_{\epsilon_{i(-1)}}$ is a diagonal matrix with positive diagonal elements.

Recall that if y_{i1} follows a normal distribution, we have

$$\pi(y_{i1}|-) = N(f_{i1}, \theta_q^{-1}).$$

Now we are going to show how to modify this normality assumption into a Dirichlet process.

The error term corresponding to the response ϵ_{i1} follows a *location-scale symmetrized stick-breaking mixtures of Gaussians* (Pati and Dunson, 2014). That is,

$$\begin{aligned} \epsilon_{i1} \sim f(\cdot) = \int N(\cdot; \mu, \theta^{-1}) dP^s(\mu, \theta),\ dP^s(\mu, \theta) = \frac{1}{2} dP(-\mu, \theta) + \frac{1}{2} dP(\mu, \theta), \\ P = \sum_{q=1}^{\infty} \pi_q \delta_{(\mu_q, \theta_q)},\ (\mu_q, \theta_q) \sim P_0 \end{aligned} \tag{4.2}$$

where $\{\theta_q\}$ is the weights, and $\mu_q \sim N(0, \sigma_\mu^2)$, $\theta_q \sim Gamma(\alpha_\theta, \beta_\theta)$. For simplicity, we assume that every "cluster" of the Dirichlet process has the same prior distribution, as denoted by $N(0, \sigma_\mu^2)$, and $Gamma(\alpha_\theta, \beta_\theta)$.

We are going to assume the number of weights is finite N. The validity has been proved and we can replace π_N with $1 - \pi_1 - ... - \pi_{N-1}$, which is equivalent to replacing V_N with $V_N = 1$ (Ishwaran and James, 2001). The definition of $\pi_1, ..., \pi_N$ becomes,

$$\pi_1 = V_1,\ \text{and}\ \pi_q = (1 - V_1)(1 - V_2) \cdots (1 - V_{q-1}) V_q,\ q = 2, ..., N, \tag{4.3}$$

where $V_1, ..., V_{N-1}$ follow independent distribution $Beta(a_q, b_q)$, and $V_N = 1$. Again, for simplisity, we assume $a_1 = ... = a_q = ... = a_N = a$, and $b_1 = ... = b_q = ... = b_N = b$, i.e. $V_1, ..., V_N$ have the same distribution. The resulting distribution for ϵ_{i1} is,

$$f(\cdot) = \sum_{q=1}^{N} \frac{\pi_q}{2} \{N(\cdot; -\mu_q, \theta_q^{-1}) + N(\cdot; \mu_q, \theta_q^{-1})\}. \tag{4.4}$$

The general idea of using this as the new error distribution is to allow multimodality for the response, at the same time constraining the error term to be symmetric about zero. The distribution for y_{i1} becomes,

$$\pi(y_{i1}|-) = \sum_{q=1}^{N} \frac{\pi_q}{2} \{N(f_{i1} - \mu_q, \theta_q^{-1}) + N(f_{i1} + \mu_q, \theta_q^{-1})\}. \tag{4.5}$$

To facilitate the smapling from posterior distributions, We are going to apply blocked gibbs sample (Ishwaran and James, 2001).

As shown in Equation 4.1, given f_i, y_i follows a $(p+1)$-dimensional multivariate normal distribution with mean f_i, and a diagonal variance covariance matrix Σ_ϵ. The sampling of the second part of Equation 4.1 can be completed using the same methodology in chapter 3. Instead of sampling based on the observed data Y, we sample based on the intermediate variable f. Hence, replacing all the ys in the previous chapter with fs.

The main focus of this chapter is to sample the posterior distribution for the first part of Equation 4.1. Note that, from Equation 4.5, we can treat the Dirichlet process modeling as an clustering procedure that cluster similar responses into one group, which means that they come from the same mixture distribution $0.5N(f_{i1} + \mu_q, \theta_q^{-1}) + 0.5N(f_{i1} - \mu_q, \theta_q^{-1})$.

In order to sample from the posterior distributions of the cluster information, we apply the blocked gibbs algorithm (Ishwaran and James, 2001). This algorithm provide a direct way of sampling the cluster information for each observation, so that we would know which q that each observation y_{i1} belongs. Let $Z_{q_i} = (\mu_q, \theta_q)$, which stands for the cluster information for observation i, $i = 1, ..., n$, and $q_i = 1, ..., N$. To summarize the prior procedures,

$$
\begin{aligned}
(y_{i1}|Z, q, f...) &\sim^{ind} \pi(y_{i1}|Z_{q_i}, f...), i = 1..., n, \\
(q_i|\mu_q, \theta_q) &\sim^{ind} \sum_{q=1}^{\infty} \pi_q \delta_{(\mu_q, \theta_q)} \\
(\mu_q, \theta_q) &\sim P_0 = N(0, \sigma_\mu^2) \times Gamma(\alpha_\theta, \beta_\theta)
\end{aligned}
$$

The direct sampling order of the posterior distributions are,

$$(Z|q, Y, ...)$$

$$(q|Z, \pi, Y, ...)$$

$$(\pi|q),$$

and the sampling of other parameters follows.

The new latent factor model is based on the intermediate latent variable f_i, which has the same dimension as y_i. It follows

$$f_i \sim N(\Lambda\eta_i, \Sigma_f),$$

where Σ_f is a diagonal matrix. For the part of grouping correlated variables, we use the same scheme as shown in chapter 3.

Suppose at current iteration there are m clusters and they are $\{q_1^*, ..., q_m^*\}$. Given all initial values and prior distributions, we first update the allocation variable q, as well as the the stick-breaking random variables $\pi, Z = (\mu_q, \theta_q)$, which shows as following:

1. Draw $(\mu_{q_j^*}|q, y_1, ...)$ from the density

$$\pi(\mu_{q_j^*}|-) \propto N(0, \sigma_\mu^2) \prod_{\{i:q_i=q_j^*\}} \pi(y_{i1}|\mu_{q_j^*}, ...),\ j = 1..., m. \tag{4.6}$$

2. Draw $(\theta_{q_j^*}|q, y_1, ...)$ from the density

$$\pi(\theta_{q_j^*}|-) \propto Gamma\Big(\frac{M_{q_j^*}}{2} + \alpha_\theta, \beta_\theta + \sum_{i:q_i=q_j^*} (y_{i1} - f_{i1})^2\Big),\ j = 1..., m, \tag{4.7}$$

   where $M_{q_j^*}$ is the number of q_i that equals q_j^*.

3. For each observation i, draw q_i

$$(q_i|Z, \pi, y_1, ...) \sim \sum_{q=1}^{N} \pi_{q,i}\delta_q(\cdot), i = 1, ..., n, \tag{4.8}$$

   where $(\pi_{1,i}, ..., \pi_{N,i}) \propto (\pi_1\pi(y_{i1}|Z_1), ..., \pi_N\pi(y_{i1}|Z_N))$.

4. From the conjugacy of the Dirichlet distribution to multinomial sampling, we have

$$\pi_1 = V_1^*, \pi_q = (1 - V_1^*)(1 - V_2^*) \cdots (1 - V_{q-1}^*)V_q^*, \ q = 2, ..., N - 1. \quad (4.9)$$

where $V_q^* \sim Beta(a + M_q, b + \sum_{l=k+1}^{N} M_l)$ for $q = 1, ..., N - 1$. M_q is the number of q_i that equals q.

5. Recall that the prior distribution for f_{i1} is $\pi(f_{i1}) = N(\lambda_{1.}\eta_i, \sigma_1^2)$. The posterior distribution of f_{i1} given all other variables is,

$$\pi(f_{i1}|-) = N((\sigma_1^{-2} + \theta_{q_i})^{-1}(\sigma_1^2\lambda_{1.}\eta_i + \theta_{q_i}y_{i1}), (\sigma_1^{-2} + \theta_{q_i})^{-1}), \ i = 1..., n. \quad (4.10)$$

For $f_{i(-1)} = (f_{i2}, ..., f_{iq})$, the posterior distribution is,

$$\begin{aligned}\pi(f_{i(-1)}|-) = N((\Sigma_{e(-1)}^{-1} + \Sigma_{\epsilon(-1)}^{-1})^{-1}(\Sigma_{e(-1)}^{-1}\Lambda_{(-1).}\eta_{(-1)i} \\ + \Sigma_{\epsilon(-1)}^{-1}y_{(-1)i}), (\Sigma_{e(-1)}^{-1} + \Sigma_{\epsilon(-1)}^{-1})^{-1}),\end{aligned} \quad (4.11)$$

where (-1) denotes with the first element removed.

The posterior distribution of the latent factor part $f_i = \Lambda\eta_i + e_i$ can be constructed the formulas in Chapter 3.

4.3 Simulation Study

We simulate 50 data sets each for two settings, both with sample size $n = 200$. We generate $p = 50$ and $p = 100$, respectively for the two settings, both including one response. For setting 1, when $p = 50$, the error term of the response $Y_{i1}, i = 1, ..., n$, follows,

$$\begin{aligned}\epsilon_{i1} \sim & \frac{2}{10}\{0.5N(1.5, 0.03) + 0.5N(-1.5, 0.03)\} \\ & + \frac{3}{10}\{0.5N(9, 0.02) + 0.5N(-9, 0.02)\} \\ & + \frac{2}{10}\{0.5N(30, 0.01) + 0.5 * N(-30, 0.01)\} \\ & + \frac{3}{10}\{0.5N(20, 0.01) + 0.5 * N(-20, 0.01)\}.\end{aligned}$$

The error term of the response in the setting 2 follows,

$$\begin{aligned}\epsilon_{i1} \sim & \frac{3}{10}\{0.5N(3, 0.03) + 0.5N(-3, 0.03)\} \\ & + \frac{3}{10}\{0.5N(15, 0.02) + 0.5N(-15, 0.02)\} \\ & + \frac{4}{10}\{0.5N(50, 0.01) + 0.5 * N(-50, 0.01)\}.\end{aligned}$$

For the rest of the set-up, the elements of the factor loading matrix Λ are generated from $N(0, 1)$, $\epsilon_{i2}, ..., \epsilon_{ip}, i = 1, ..., n$ are from $N(0, 0.01)$. The number of clusters is $N = 20$. The prior parameters settings are, $\sigma_\mu^2 = 4$, $\alpha_\theta = 10, \beta_\theta = 1$, which is chosen to make the different clusters distinct. The rest of the prior parameters are the same as in Chapter 3, $\sigma_j^{-2} \sim Gamma(a_\sigma = 1, b_\sigma = 0.3)$, $\nu \sim unif(2, 4)$, $a_1 \sim unif(1.1, 3.1)$, and $a_2 \sim unif(2.1, 4.1)$. We generate 25000 posterior samples with the first 5000 as the burn-in, and collect every 5th sample.

The density plots of the responses for both settings are shown in Figure 4.3 and Figure!4.3. We see the response variables are multimodal. We compare the mean square error (MSE), average absolute bias (AAB), and maximum average bias (MAB) from this nonparametric method with the method in chapter 3, as well as Lasso and elastic net.

Table 4.1 Simulation study results from setting 1 based on $N = 50$ simulations with $p = 50, k = 4$, mean square error (MSE), absolute average bias (AAB), and maximum average bias (MAB) of standardized Y.

	Y	Nonparametric	Parametric	Lasso	Elastic Net
MSE	mean	0.9621	0.9530	0.9937	1.0070
	min	0.9156	0.8771	0.9302	0.9418
	max	0.9927	0.9857	0.9950	1.0999
AAB	mean	0.8162	0.8127	0.8291	0.8362
	min	0.7826	0.7682	0.8023	0.7976
	max	0.8323	0.8314	0.8360	0.8731
MAB	mean	1.8689	1.8733	1.8226	1.8936
	min	1.7104	1.7378	1.7554	1.7709
	max	2.0819	2.0987	1.9512	2.0912

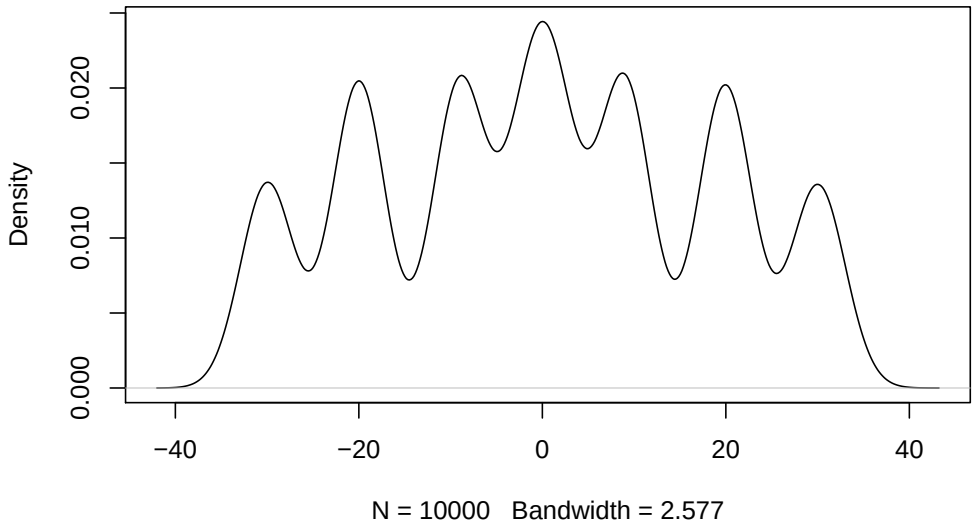

Figure 4.1 Density of the response Y_1 from the $N = 50$ simulation dataset under setting 1.

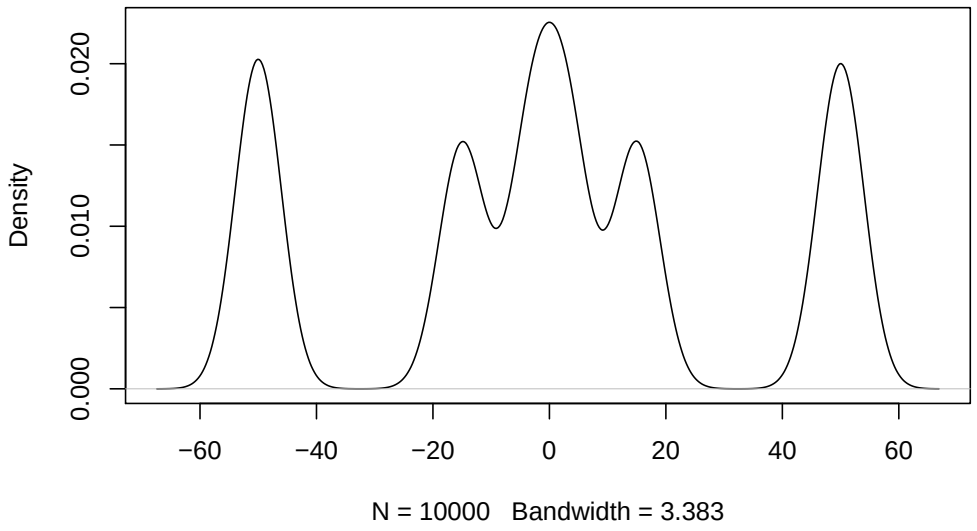

Figure 4.2 Density of the response Y_1 from the $N = 50$ simulation dataset under setting 2.

Table 4.2 Simulation study results from setting 2 based on $N = 50$ simulations with $p = 100, k = 10$, mean square error (MSE), absolute average bias (AAB), and maximum average bias (MAB) of standardized Y.

	Y	Nonparametric	Parametric	Lasso	Elastic Net
MSE	mean	0.9532	0.9453	0.9950	1.0184
	min	0.8827	0.8725	0.9950	0.9950
	max	0.9858	0.9831	0.9950	1.1114
AAB	mean	0.7625	0.7617	0.7746	0.7917
	min	0.7348	0.7329	0.7691	0.7715
	max	0.7778	0.7752	0.7791	0.8488
MAB	mean	1.7990	1.8426	1.6713	1.7852
	min	1.6469	1.6689	1.6335	1.6447
	max	1.9730	2.1020	1.7387	2.0405

We notice that both nonparametric and parametric methods perform better than Lasso and Elastic net, in terms of mean square error. Interestingly, the performance of the nonparametric method is not better than the parametric one, based on the mean square error. One might also notice that among all four methods, the nonparametric method has the lowest MAB, which means that under our simulated settings the nonparametric method has the best performance for the worst case scenario. The gain of the nonparametric assumption is that we safely model the data without worrying whether the normality assumption holds.

4.4 Real Data Analysis

We use data from Carneigie Mellon University on estimation for percentage of body fat. The predictors in this data set are age (years), weight (lbs), height (inches), neck circumference (centimeters), chest circumference (centimeters), abdomen 2 circumferences (centimeters), hip circumferences (centimeters), thigh circumference (centimeters), knee circumference (centimeters), ankle circumference (centimeters), biceps circumference (centimeters), forearm circumference (centimeters), and wrist circumference (centimeters). There are $n = 252$ subjects. Figure 4.4 shows the density plot of the percentage of body fat. It does not appear to be normally

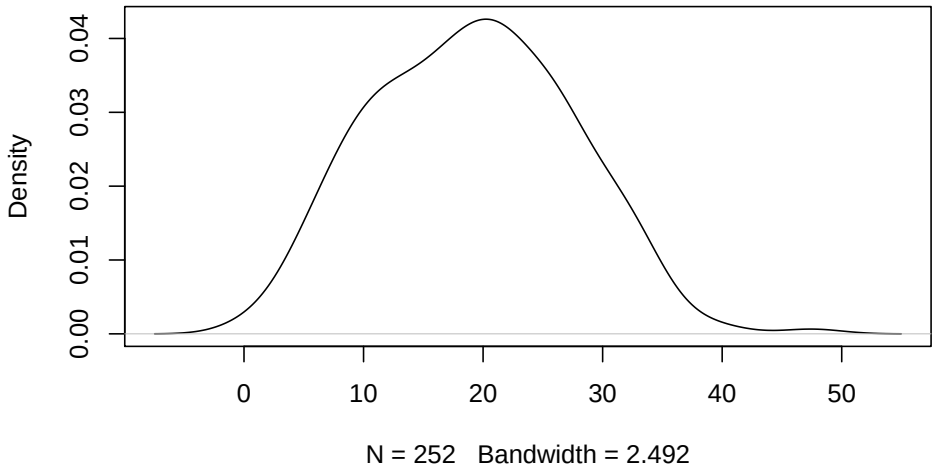

Figure 4.3 Density plot of the percentage of body fat.

distributed and there is a slight hint of bimodality. The mean square error (MSE), average absolute bias (AAB), and maximum absolute bias (MAB) are shown in Table 4.3.

Similar as the simulation results, the nonparametric method has not out-performed the original parametric method. Among all four methods, the parametric method has the lowest MSE. Comparing with the nonparametric method to the Lasso and Elastic net, all three measures are still higher. We argue that as we can see in Figure 4.4, the multimodality of the response is less obvious than we would prefer. The advantage of it does not appear in this case.

Table 4.3 Mean squared error results from four methods, using bodyfat data.

Y	Nonparametric	Parametric	Lasso	Elastic Net
MSE	0.3095	0.2582	0.2948	0.3001
AAB	0.4479	0.4147	0.4455	0.4522
MAB	2.0347	1.3448	1.8994	1.7586

4.5 Conclusion and Discussion

In this study we proposed a nonparametric Bayesian latent factor modeling, with the nonparametric assumption for the response variable. We started with the latent factor models in Chapter 3, which is focused on variable selection. This method has been extended to allow nonparametric assumption for the response variable. More specifically, we discuss the application of stick-breaking construction of Dirichlet process in the Bayesian latent factor modeling.

The results from the simulation study interestingly showed that the parametric model actually performed better than the nonparametric model, when the response is from a multimodal mixture of distributions. This presents some ideas for future study. Since we have only done a small simulation study, a more extensive one might reveal that the nonparametric method is better. This brings out the problem of computing time. The algorithm of sampling from the posterior distributions should be optimized, in order to reduce computing time. The computing time for the nonparametric method is significantly longer than the parametric counterpart. This is because we introduced the intermediate latent factors, as well as the Dirichlet process clustering variables. Currently the updates of some of the variables are done element wise. Future study can focus on block updates for the posterior distributions.

The most important question is whether combining factor analysis with clustering techniques is possible under the Bayesian setting. This certainly is very useful for real data analysis. One can imagine for the colon cancer data in Chapter 3, suppose we did not know the whether one has colorectal cancer or not, instead, we had some continuous measurement. Using the proposed method in this chapter, we can automatically group people into several groups (ideally 2, with and without cancer), meanwhile, find out which groups of genes are important.

One of the strengths when using Bayesian latent factor modeling is that it allows multivariate response. Since the method treating all variables, both responses and

predictors, as the observed variables, the dimension of the response vector can be arbitrary. Even though we only discuss the nonparametric assumption for univariate response, the extension to multivariate cases is straightforward.

Chapter 5

Conclusions and future work

5.1 Conclusions

The dissertation has addressed two statistical questions, and explored the extension of Bayesian sparse latent factor modeling into the nonparametric settings.

The first question being addressed is how to handle bivariate count heaping data. We has discussed two situations of heaping, interval censored heaping and scaled heaping. The bivariate probability distributions are expressed in terms of four parts of copula functions. For both heaping scenarios, our methods on the simulation data sets have presented adequate results. Note that the coverage probabilities are generally less than 95%. Additional simulation studies were done without the data being heaped. However, the resulting coverage probabilities were similar as before. Therefore, caution needs to be taken when using copula to modeling mixture distributions.

In chapter 3, we propose a method for group variables selection. Our method has good performance especially when the variables are all continuous. When the response is binary, we compared our method with the elastic net method and the prediction error is larger using our method. However, we argue that our methods, first, can be used without any preproccessing. Second, since predicting a binary variable requires a cut-off point, the randomness of this value might affect the performance of our method.

The Bayesian sparse latent factor modeling has been extended to the nonparametric setting in chapter 4. It simultaneously achieves the grouping of correlated

variables and clustering similar observations. By using a vector of intermediate variables, the bayesian latent factor model has been connected with Dirichlet process. The application of the nonparametric assumption on Bayesian latent factor modeling is relatively new and offers a lot of possibilities.

5.2 Future work

Both the advantages and disadvantages of the methods suggested in this dissertation can be further studied. For interval censored bivariate heaping, how to handle heaping in zero-inflated count data while modeling the probability of heaping can be further explored. At the same time, one should always keep in mind that the multivariate distribution is expressed using copula functions. Hence, the accuracy of the individual likelihood when heaping exists should be handled with caution, especially when the person has observed count 0. For the scaled heaping scenario, we applied copula functions for mixture modeling. More simulation studies should be carried out and the 95% coverage probability needs to be validated.

Group variable selection method suggested in chapter 3 can be developed for Poisson, multinomial, and other distributions. Improving the posterior sampling efficiency and stability should be one the objectives when extending the method to the aforementioned distributions.

The nonparametric Bayesian latent factor modeling in chapter 4 can be extend to other nonparametric settings, such as Gaussian process. The efficiency of posterior sampling can be improved. We used blocked Gibbs sampler approach for posterior sampling. Other sampling techniques should be considered and appplied. Again, more simulation studies should be done to examine the performance of this method.

Bibliography

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

U. Alon, N. Barkai, D.A. Notterman, K. Gish, Ybarra S., D. Mack, and et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissures probed by oligonucleotide arrays. *Proceedings National Academy of Sciences. USA*, 96:6745–6750, 1999.

W.H. Au, K.C.C. Chan, A.K.C. Wong, and Y. Wang. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(2):83–101, 2005.

H.Y. Bar and D.R. Lillard. Accounting for heaping in retrospectively reported event data - a mixture-model approach. *Statistics in Medicine*, 31(27):3347–3365, 2011.

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

J.F. Bobb, L. Valeri, B.C. Henn, D.C. Christianti, Wright R.O., and et al Mazumdar, M. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508, 2015.

C.M. Carvalho, J. Chang, J.E. Lucas, J.R. Nevins, Q. Wang, and West M. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2012.

T.H. Cummings, J.W. Hardin, A.C. Mclain, J.R. Hussey, K.J. Bennett, and G.M. Wingood. Modeling heaped count data. *The Stata Journal*, 15(2):457–479, 2015.

F. Famoye. On the bivariate negative binomial regression model. *Journal of Applied Statistics*, 37(6):969–981, 2010.

S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

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

J. Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities, 1991.

D.F. Heitjan and D.B. Rubin. Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85(410):304–314, 1990.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.

P. Hernández, J. Huerta-Cepas, D. Montaner, F. Al-Shahrour, . Valls, J, . Gómez, L, . Capellá, G, J. Dopazo, and M. Pujana. Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics*, 8(1):185, 2007.

A.E. Hoerl and R.W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

H. Ishwaran and L.F. James. Some further developments for stick-breaking priors: finite and infinite clustering and classification. *Sankhy: The Indian Journal of Statistics*, 65(3):577–592, 2003.

H. Ishwaran and J.S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773, 2005.

H. Joe. *Multivariate models and dependence concepts, monographs on statistics and applied probability*. London Chapman & Hall, 1997.

H.C. Kim and J. Lee. Clustering based on gaussian processes. *Neural Computation*, 19(11):3088–3107, 2007.

S. Kim, Tadesse M.G., and M. Vannucci. Variable selection in clustering via dirichlet process mixture models. *Biometrika*, 93(4):877–893, 2006.

D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. *Lecture Notes in Computer Science*, 4666:381–388, 2007.

D. Kwon, Landi M.T., M. Vannucci, J.I. Haleem, D. Prieto, and R.M. Pfeiffer. An efficient stochastic search for bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics and Data Analysis*, 55:2807–2818, 2011.

F.B. Lempers. Posterior probabilities of alternative linear models. 1971.

D. Li and S. Li. *DNA Microarray Technology and Data Analysis in Cancer Research*. World Scientific, 2008.

Q. Li and N. Lin. The bayesian elastic net. *Bayesian Analysis*, 5(1):151–170, 2010.

W. Li and Xiong M. Tclass: tumor classification system based on gene expression profile. *Bioinformatics*, (2):325–326, 2002.

A.W. Marshall and I. Olkin. A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association*, 80(390):332–338, 1985.

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

N. Naifar. Modelling dependence structure with archimedean copulas and applications to the itraxx cds index. *Journal of Computational and Applied Mathematics*, 235:2459–2466, 2011.

A.K. Nikoloulopoulos and D. Karlis. Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, 37(9):1555–1568, 2010.

NIMH Multislice HIV/STD Prevention Trial for African American Couples Group. Project eban: An hiv/std intervention for african american couples. *Journal of Acquired Immune Deficiency Syndromes*, 49(Suppl 1):S15–S27, 2008.

Konstantina P., Z. Ghahramani, and David A.K. A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems 25*, pages 2987–2995. Curran Associates, Inc., 2012.

A. Panagiotelis, C. Czado, and H. Joe. Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, 107(499):1063–1072, 2012.

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

D. Pati and David B. Dunson. Bayesian nonparametric regression with varying residual density. *Annals of the Institute of Statistical Mathematics*, 66(1):1–31, 2014.

C.R. Rao and D.K. Dey. *Handbook of Statistics: Bayesian Thinking, Modeling and Computation*. Elsevier, 2005.

G. Roberts and J. Rosenthal. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of Applied Probability*, 44:458–475, 2007.

M. Schomaker and C. Heumann. Bootstrap inference when using multiple imputation. *arXiv e-print*, 2016.

J.S. Shaik and Yeasin M. A unified framework for finding differentially expressed genes from microarray experiments. *BMC Bioinformatics*, 8:347, 2007.

D.B. Sharma, Bondell H.D., and Zhang H.H. Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics*, 22(2):319–340, 2013.

P.X.K. Song, M. Li, and Y. Yuan. Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68, 2009.

M.A. Suchard, R.E. Weiss, and F.W. Crawford. Sex, lies and self-reported counts: bayesian mixture models for heaping in longitudinal count data via birth-death processes. *Annals of Applied Statistics*, 9(2):572–596, 2015.

M.G. Tadesse, N. Sha, and M. Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470): 602–617, 2005.

Y.W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems 17*, pages 1385–1392. 2005.

Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierachrchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Ser. B(58):267–288, 1996.

L. Wasserman and K. Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, 2009.

M. West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics*, pages 723–732, 2003.

A.J. Yang and X.Y. Song. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2):723–732, 2010.

X. Zhao and X. Zhou. Copula models for insurance claim numbers with excess zeros and time-dependence. *Insurance: Mathematics and Economics*, 50(1):191–199, 2012.

S. Zinn and A. Würbach. A statistical approach to address the problem of heaping in self-reported income data. *Journal of Applied Statistics*, 43(4):682–703, 2016.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005.