A General Method to Couple Prior Distributions

by

Yiding Liu

Bachelor's of Applied Mathematics, China University of Geosciences, 2018

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This thesis was presented

by

Yiding Liu

It was defended on

April 6th, 2021

and approved by

Dr. Chris McKennan, Department of Statistics

Dr. Satish Iyengar, Department of Statistics

Dr. Zhao Ren, Department of Statistics

A General Method to Couple Prior Distributions

Yiding Liu, M.S.

University of Pittsburgh, 2021

There is a lot of statistic models based on marginal distribution and joint distribution relationships. Such statistical models are widely used in medicine, biology, finance, etc. Many modern medical datasets contain observations from multiple time points and treatment conditions. Adaptive shrinkage, a general method to estimate marginal prior distributions, has been developed to analyze such data for a single time point or condition, few method has been developed to analyze joint distribution for different time points or different conditions. The reason is mainly because the difficulty of constructing multi-dimensional prior distributions with dependent variables. A few Bayes' methods can be applied to these type data. Although, it is non-trivial and difficult to estimate joint distribution directly, we can easily estimate marginal prior distributions separately. In this thesis, I develop a simple, general and straightforward method to couple prior distributions for multi-dimensional genetic effects. The main goal is to research the relationship between the sign of effect of a phenotype at different time points. I couple prior distributions to model joint distribution and estimate parameters at multiple time points. Copula Estimation described from Copula Theory and Its Applications provides a parametric copula inference method for my estimation. I construct a model and develop a method to couple prior distributions to estimate my parameters at multiple time points by deriving useful expressions, applying R language for data simulations, and using maximum likelihood estimation. I simulate data from both the real copula model and multivariate normal distribution. The true model performs better in estimating the parameters. This copula model successfully bridge the gap between joint distribution and dependent marginal distributions. There is still more room to improve my copula model.

Keywords: Copula, Couple Prior Distributions, Dependent Variables, Bayes' Methods, Maximum Likelihood Estimation, Marginal Distribution.

Table of Contents

List of Figures

Preface

I am inspired by the data about the effect of a phenotype on DNA methylation. In my thesis, my goal is to research the relationship between the sign of effect of a phenotype at different time points. I develop a method to couple prior distributions to construct a joint distribution and estimate parameters in my model. I use a couple model to estimate marginal distributions first and then construct joint distribution with dependent marginal distributions. I simulate data to estimate parameters using maximum likelihood estimation, compare two set of simulations to evaluate my model, and propose a method to optimize limitations to improve my model. Meanwhile, I'd like to show my respect and thanks to my advisor Dr.Chris McKennan who helps me a lot and gives me such a golden change to finish my thesis. I also appreciate all the researchers who provided me with existing works about my thesis.

1.0 Introduction

Marginal distributions and joint distribution are important in statistical model because many modern real datasets are consist of observations from multiple time points and variable conditions. Many current methods are developed to analyze such data for a single marginal distribution or condition, very few methods were invented to analyze the joint distribution, especially for dependent variables. Previous studies showed it was non-trivial to estimate joint distribution directly. I tried to find a general method to copula prior distributions and built a copula model to estimate parameters in different simulations to bridge the gap between joint distribution and dependent marginal distributions. Inspired from the effect of a phenotype on DNA methylation at birth and age 7. My goal is to research the relationship between the sign of effect of a phenotype at different time points. I couple prior distributions to construct joint distribution and estimate parameters at multiple time points. There are several related existing work, such as Adaptive shrinkage ([9]) and Copula Theory and Its Applications ([7]). Adaptive Shrinkage is a general method to estimate marginal prior distributions but not flexible enough to jointly model the effect of a phenotype on DNA methylation at birth and age 7. It is mainly because Adaptive Shrinkage ([9]) assumes the effect of interest β_g as a scalar vector to model marginal distributions, not a vector. But I take the effect of interest β_g as a vector to model joint distribution. The Copula Estimation ([7]) provides a parametric copula inference method for i.i.d random sample and uses maximum likelihood methods.

A copula is a multivariate cumulative distribution function for which to describe the dependence between random variables. Copulas are popular in high-dimensional statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating marginals and copulas separately. I assume independent unimodal distributions for the effect of interest, propose a general method to couple prior distributions, and use maximum likelihood estimation to estimate my parameters in two sets of simulating data.

The thesis is organized as follows. Section 2 discusses several related existing work, like

Adaptive shrinkage, Copula Estimation and Copula model application. These existing works lay a solid starting point for my thesis and inspire me to figure out a general method to couple prior distributions. My copula model is inspired from the effect of a phenotype on DNA methylation at multiple time points. Section 3 lists some assumptions and describes how to couple prior distributions and construct the copula model. Section 4 discusses applying my estimator in two different cases after we build the copula model. Section 5.1 to 5.4 consider two sets simulating data to estimate parameters, one is simulating data from the true model, another is simulating data from multivariate normal distribution directly. These two different simulations lead to different results judging from R outcomes. Section 5.5 compares two simulations to understand the advantages of the general method to couple prior distributions. Section 6 discusses limitations of my copula model that need to be improved and further study, I propose to use MCMC to extend my model to more general studying.

2.0 Existing Work

There are several existing works about copula model. Adaptive Shrinkage described ([9]) is an empirical bayesian approach for marginal prior distributions. In this paper, the authors provide a general method to estimate marginal prior distributions, assume the distribution of effects is unimodal, and use effect sizes and standard errors to summarize each measurement. This method assumes the effect of interest β_g as a scalar to model marginal distributions, thus It has limitation on modeling joint distribution. Another important existing work that I utilized in my study is the Copula Estimation ([7]). In this book, they provide several different methods about estimation procedures for copula model, especially parametric copula inference method for i.i.d random sample with dependent marginal distributions. It also uses maximum likelihood method to estimate parameters.”DNA Methylation” ([5]) has a good introduction about DNA methylation. It is this paper which gives me a strong interest about the motivating data, intriguing me to research the effect of a phenotypes on DNA methylation at different time points. The copula model is widely used in different fields, see publication ”On Modeling Insurance Claims Using Copulas” ([4]). It is a very useful tool to couple marginal distributions for getting an estimator of the joint distribution.

3.0 Problem Setup

Suppose we measure the expression of p genes to understand the relationship between gene expression and d phenotypes. Let $g \in \{1, \ldots, p\}$ index gene and $j \in \{1, \ldots, d\}$ index phenotype, For gene expression: I expect p is 10^4 and d is small (≤ 5) while p represent gene and d represent multiple time points, there will be not too much time points for the effect of a phenotype on DNA methylation. I also define β_g to be the effect of a phenotype on DNA methylation and $\hat{\boldsymbol{\beta}}_g = \left(\hat{\beta}_{g1}, \ldots, \hat{\beta}_{gd}\right)^T$ is the Ordinary least squares(OLS) estimator for $\boldsymbol{\beta}_g$, the effect of a phenotype on the expression of gene g at different times. Assuming that $\boldsymbol{V}_g \in \mathbb{R}^{d \times d}$ is the variance of $\hat{\boldsymbol{\beta}}_g$. Because in linear regression model $\hat{\boldsymbol{\beta}}_g = (X^T X)^{-1} X^T Y_g$ and $\boldsymbol{V}_g = \sigma_g^2 (X^T X)^{-1}$, X is designed matrix. I assume $\hat{\boldsymbol{V}}_g$ is the estimator for $\boldsymbol{V}_g$ because Central limit theorem and Slutsky's theorem, and $\hat{\boldsymbol{\beta}}_g$ is the Ordinary least squares estimator for $\boldsymbol{\beta}_g$. Both are observed in my model.

$$\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{\beta}_g, \boldsymbol{V}_g \sim N_d\left(\boldsymbol{\beta}_g, \boldsymbol{V}_g\right), \quad g \in \{1, \ldots, p\} \tag{1a}$$

$$\boldsymbol{\beta}_g \mid h(\cdot) \sim h(\boldsymbol{\beta}_g), \quad g \in \{1, \ldots, p\} \tag{1b}$$

$$\beta_{gj} \mid h_j(\cdot) \sim h_j(\beta_{gj}), \quad g \in \{1, \ldots, p\}; j \in \{1, \ldots, d\} \tag{1c}$$

$$h_j(\cdot) \mid \boldsymbol{\pi}_j \sim \sum_{k=-K}^{K} \pi_{jk} f_{jk}(\cdot), \tag{1d}$$

where $h_j(\cdot)$ is a general prior for β_{gj} and f_{jk} is a simple and known density function. If $k > 0$,the f_{jk} is a distribution greater than 0; if $k < 0$,the f_{jk} is a distribution smaller than 0; if $k = 0$,the f_{j0} is the point mass at 0. In my thesis, since f_{jk} is a simple and known density function with such properties, I suppose to use the half normal distribution with known variance.

Similar to ([9]), I assume $f_{jk}(x)$:

$$f_{jk}(x) = HN_{\geq 0}(x; 0, \sigma_{jk}^2) \quad j \in \{1, \ldots, d\}; k \in \{1, \ldots, K\}$$

where $HN_{\geq 0}(x; 0, \sigma_{jk}^2) = \frac{2}{\sqrt{2\pi\sigma_{jk}^2}} \exp(-\frac{1}{2\sigma_{jk}^2} x^2) 1\{x \geq 0\}$ is the density of the positive half-normal distribution (the positive half of $N(0, \sigma_{jk}^2)$). We will assume the hyperparameters σ_{jk}^2 is known and if $\pi_{j(-k)} = \pi_{jk}$, then this implies the prior $h_j(\cdot) \mid \boldsymbol{\pi}_j \sim \pi_0\delta_0 + \sum_{k=1}^{K} 2\pi_{jk}N_1\left(\cdot; 0, \sigma_{jk}^2\right)$ for known σ_{jk}^2. For now f_{jk} is known, I only need to derive π_{jk} and it's obvious that we can easily estimate $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_d$ with ([9]), therefore, I assume that I can observe these in my model:

- $\hat{\boldsymbol{\beta}}_g$: the estimate for $\boldsymbol{\beta}_g$, effect of a phenotype on DNA methylation
- $\boldsymbol{V}_g$: the variance of $\hat{\boldsymbol{\beta}}_g$ is known.
- f_{jk} for all $j \in \{1, \dots, d\}$ and $k \in \{-K, \dots, K\}$: The non-negative and non-positive density functions used to parametrize the prior for β_{gj}.
- $\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_d$: The mixture weights that define the marginal priors for different phenotypes.

My primary goal is to determine the relationship between the sign of the effects $\boldsymbol{\beta}_g$ at different time j, that is same to estimate the correlation coefficients. Firstly, I need to estimate the prior $\text{pr}\left(\boldsymbol{\beta}_g\right)$ and posterior $\text{pr}\left(\boldsymbol{\beta}_g \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g\right)$. The latter will allow us to perform inference on the joint distribution of $\boldsymbol{\beta}_g$. We are particularly interested in being able to do inference on the signs of $\beta_{g1}, \dots, \beta_{gd}$. The article leads a way for us to think about this question ([2, pages 15-17]), Some statements might be:

$$\mathbb{P}\left(\{\beta_{gj}, \beta_{gj'} > 0\} \cup \{\beta_{gj}, \beta_{gj'} < 0\} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, h\right), \quad g \in \{1, \dots, p\}; j, j' \in \{1, \dots, d\}$$

The probability that the effect at two time points j and j' on DNA methylation are both positive or negative.

$$\mathbb{P}\left(\{\beta_{g1}, \dots, \beta_{gd} > 0\} \cup \{\beta_{g1}, \dots, \beta_{gd} < 0\} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, h\right), \quad g \in \{1, \dots, p\}$$

The probability that the effect at multiple time points from 1 to d on DNA methylation are all positive or negative.

$$\mathbb{P}\left(\beta_{gj} > 0, \beta_{gj'} \leq 0 \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, h\right), \quad g \in \{1, \ldots, p\}; j, j' \in \{1, \ldots, d\}$$

The probability that the effect of a single gene at two time points j and $j^{'}$ on DNA methylation are different. The effect is positive at time j and negative or 0 at time $j^{'}$.

3.1 Modelling the Relationship between h and $h_1, \ldots, h_d$

Before I construct a copula model, I need to estimate the prior $\text{pr}\left(\boldsymbol{\beta}_g\right)$. $h_j(\cdot)$ specify the marginal prior distribution for β_{gj}. Estimating these amounts to estimating the prior weights $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_d$ and can be easily done using ([9]). The function $h : \mathbb{R}^d \to \mathbb{R}$ is the multivariate prior density for $\boldsymbol{\beta}_g$. While we have not specified the functional form for h, we know that its jth marginal must be h_j. That is

$$h_j(x_j) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h\left(x_1, \ldots, x_d\right) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_d, \quad j \in \{1, \ldots, d\}$$

Therefore, I propose constructing h using a copula model.

Recall that the primary inferential goal is to understand the relationship between the signs of $\beta_{g1}, \ldots, \beta_{gd}$. Define the random variable z_{gj} to be:

$$\begin{aligned} z_{gj} \mid \boldsymbol{\pi}_j &\sim \text{Categorical}\left(\{-K, -K+1, \ldots, K\}; \boldsymbol{\pi}_j\right), \quad j \in \{1, \ldots, d\} \\ \mathbb{P}\left(z_{gj} = k \mid \boldsymbol{\pi}_j\right) &= \pi_{jk}, \quad k \in \{-K, -K+1, \ldots, K\}; j \in \{1, \ldots, d\}. \end{aligned}$$

Then the prior $\text{pr}(\beta_{gj} \mid h_j)$ can be expressed as

$$\begin{aligned} \beta_{gj} \mid z_{gj} &\sim f_{jz_{gj}}(\beta_{gj}) \\ z_{gj} \mid \boldsymbol{\pi}_j &\sim \text{Categorical}\left(\{-K, -K+1, \ldots, K\}; \boldsymbol{\pi}_j\right). \end{aligned}$$

Note that z_{gj} encodes the sign of β_{gj}. Assuming the magnitude and sign of β_{gj} are independent, if $z_{gj} < 0$, then $\beta_{gj} < 0$ (and vice-versa), since our goal is to perform inference on the signs of

$\beta_{g1}, \ldots, \beta_{gd}$. I make following assumption:

$$\operatorname{pr}\left(\beta_{g1}, \ldots, \beta_{gd} \mid z_{g1}, \ldots, z_{gd}\right) = \prod_{j=1}^{d} \operatorname{pr}\left(\beta_{gj} \mid z_{gj}\right). \tag{2}$$

That is, once we know the signs and expected magnitudes of $\beta_{g1}, \ldots, \beta_{gd}, \beta_{g1}, \ldots, \beta_{gd}$ are independent. The prior for $\boldsymbol{\beta}_g$ can therefore be expressed as

$$\operatorname{pr}(\boldsymbol{\beta}_g) = \int \prod_{j=1}^{d} \operatorname{pr}\left(\beta_{gj} \mid z_{gj}\right) \operatorname{pr}\left(z_{g1}, \ldots, z_{gd}\right) dz_{g1} \cdots dz_{gd}. \tag{3}$$

Therefore, I only need to model the dependence between the categorical (i.e. multinomial) random variables $z_{g1}, \ldots, z_{gd}$ to specify the prior for $\boldsymbol{\beta}_g$. One way to do this would be to construct a copula model to specify the prior for $\boldsymbol{\beta}_g$.

3.2 Constructing the Copula Model

A copula is a way of modelling the dependence between random variables given the marginal distributions of each random variable. I suggest looking at ([6, pages 1-12]) for an introduction to copula modelling, and ([8]) for an in-depth look at copulas applied to categorical (i.e. multinomial) data. Given the modelling assumption in (2) and the resulting prior in (3), I need to use a copula to model the dependence between $z_{g1}, \ldots, z_{gd}$, specifically, let $\boldsymbol{z}_g = \left(z_{g1}, \ldots, z_{gd}\right)^T$. Let $N_d(\mathbf{0}, \boldsymbol{R})$ be the multivariate normal distribution with correlation matrix $\boldsymbol{R} \in \mathbb{R}^{d \times d}$, and define

$$x_{j(-K)} < \cdots < x_{j(K-1)}$$

to be a partition of $\mathbb{R}$ that act as cut points in a latent variable. Note that $x_{j(-K)}, \dots, x_{j(K-1)}$ are uniquely determined by $\boldsymbol{\pi}_j$. A normal copula model assumes for the latent vector $\boldsymbol{w}_g$,

$$\boldsymbol{w}_g = \left(w_{g1}, \dots, w_{gd}\right)^T \mid \boldsymbol{R} \sim N_d(\boldsymbol{0}, \boldsymbol{R}), \quad g \in \{1, \dots, p\}$$

$$z_{gj} \mid \boldsymbol{\pi}_j = -K1\{w_{gj} \leq x_{j(-K)}\} + K1\{w_{gj} > x_{jK}\} + \sum_{k=-K+1}^{K-1} k1\{x_{jk} < w_g \leq x_{j(k+1)}\}$$

$$\boldsymbol{R}_{12}, \dots, \boldsymbol{R}_{(d-1),d} \overset{i.i.d}{\sim} U[-1, 1]1\{\boldsymbol{R} \geq \boldsymbol{0}\}.$$

I can either estimate $\boldsymbol{R}$ using MLE (this will be time consuming given the constraints on the problem) or sample from the posterior with MCMC. I assume Model (1) is true and Assumption (2) holds. I assume the following, which are standard when modelling genetic data:

- $f_{j0} = \delta_0$, which is the point mass at 0.
- For $k \in \{1, \dots, K\}$, $f_{jk}(-x) = f_{j(-k)}(x) = 0$ for all $x \geq 0$. That is, f_{jk} is the density of a positive random variable and $f_{j(-k)}$ is the density of a negative random variable.
- The marginal $h_j(\cdot)$ is symmetric around 0. That is, $f_{jk}(x) = f_{j(-k)}(-x)$ for $x \geq 0$ and $\pi_{jk} = \pi_{j(-k)}$ for all $j \in \{1, \dots, d\}$, $k \in \{1, \dots, K\}$.

Now, the first step is to derive expressions the following assuming $f_{jk}(x) = HN_{\geq 0}(x; 0, \sigma_{jk}^2)$. That will be used in my simulations part:

$$\mathrm{pr}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, z_{g1}, \dots, z_{gd}\right) \tag{4}$$

$$\mathrm{pr}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, w_{g2}, \dots, w_{gd}\right) \tag{5}$$

When computing (4) and (5), I have to integrate out $\boldsymbol{\beta}_g$. The specific process will be shown in Appendices.

4.0 Deriving Estimator

4.1 Starting With the Simple Case

I define $d = 2$ index time points and $K = 1$ index the upper bound for k as the simple case. My primary goal is to determine the relationship between the sign of effects of $\beta_{g1}, \ldots, \beta_{gd}$, and we know β_{gj} is decided by z_{gj} and z_{gj} is decided by w_{gj} in my copula model. I assume the correlation coefficient for w_{gj} and $w_{gj'}$ is ρ. I need to derive likelihood function first and then use maximum likelihood estimation to estimate my parameter ρ to maximize log-likelihood function. $d = 2$ means $\boldsymbol{w}_g, \boldsymbol{z}_g, \boldsymbol{\beta}_g$ are vectors contain two elements while R is a 2 by 2 correlation matrix for w_g.

$$w_g = \begin{bmatrix} w_{g1} \\ w_{g2} \end{bmatrix} z_g = \begin{bmatrix} z_{g1} \\ z_{g2} \end{bmatrix} \boldsymbol{\beta}_g = \begin{bmatrix} \beta_{g1} \\ \beta_{g2} \end{bmatrix} R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{6}$$

$k = 1$ means $z_{gj} \mid \boldsymbol{\pi}_j \sim \text{Categorical}\left(\{-1, 0, 1\}; \boldsymbol{\pi}_j\right), \quad j \in \{1, 2\}$ and π_j is a length-3 vector and represents probabilities for $z_{gj} = \{-1, 0, 1\}$ for$j \in \{1, 2\}$.

My goal is to estimate parameter ρ to maximize the likelihood function:

$$L(\rho) = \prod_{g=1}^{p} \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right)$$

Transforming $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right)$ with Bayes' Theorem to be the following form:

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right) = \sum_{\boldsymbol{z}_g \in I} \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g\right) \mathbb{P}\left(\boldsymbol{z}_g \mid \rho\right)$$

Where $I = \{(0, 0), (0, 1), (0, -1), (1, 0), (1, 1), (1, -1), (-1, 0), (-1, 1), (-1, -1)\}$

When $\boldsymbol{z}_g = (0, 0)$, it means $\beta_{g1} = 0$ and $\beta_{g2} = 0$

When $\boldsymbol{z}_g = (0, 1)$, it means $\beta_{g1} = 0$ and $\beta_{g2} > 0$

When $\boldsymbol{z}_g = (0, -1)$, it means $\beta_{g1} = 0$ and $\beta_{g2} < 0$

When $\boldsymbol{z}_g = (1, 0)$, it means $\beta_{g1} > 0$ and $\beta_{g2} = 0$

When $\boldsymbol{z}_g = (1, 1)$, it means $\beta_{g1} > 0$ and $\beta_{g2} > 0$

When $\boldsymbol{z}_g = (1, -1)$, it means $\beta_{g1} > 0$ and $\beta_{g2} < 0$

When $\boldsymbol{z}_g = (-1, 0)$, it means $\beta_{g1} < 0$ and $\beta_{g2} = 0$

When $\boldsymbol{z}_g = (-1, 1)$, it means $\beta_{g1} < 0$ and $\beta_{g2} > 0$

When $\boldsymbol{z}_g = (-1, -1)$, it means $\beta_{g1} < 0$ and $\beta_{g2} < 0$

In order to obtain the likelihood function. My first goal is to derive the probabilities $\mathbb{P}\left(z_{gj} \mid \rho\right)$ which is a length-9 vector because each z_{gj} can be chosen from $\{-1, 0, 1\}$. In my thesis. I set bounds for each truncated normal distribution and integrate $\boldsymbol{w}_g$ to get 9 different probabilities for $\mathbb{P}\left(z_{gj} \mid \rho\right)$. Next, I derive the likelihood $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g\right)$. I combine bivariate normal distribution, half normal distribution, complete square transformation, and Expression(4) to obtain corresponding likelihoods for 9 different scenarios because $\boldsymbol{z}_g$ can be chosen from set I in which contain 9 different vectors. Then I iterate the same processes for each gene from 1 to p to get the final likelihood function and apply one dimensional optimization function in R to estimate the parameter ρ.

4.2 More General Case

I define $d > 2$ and $K > 1$ as the general case. d will be a little bit larger(≤ 5), but d can not be to much large because it represents multiple time points. d means $\boldsymbol{w}_g, \boldsymbol{z}_g, \boldsymbol{\beta}_g$ are vectors contain d elements, R is a d by d correlation matrix for w_g:

$$w_g = \begin{bmatrix} w_{g1} \\ w_{g2} \\ \vdots \\ w_{gd} \end{bmatrix} z_g = \begin{bmatrix} z_{g1} \\ z_{g2} \\ \vdots \\ z_{gd} \end{bmatrix} \beta_g = \begin{bmatrix} \beta_{g1} \\ \beta_{g2} \\ \vdots \\ \beta_{gd} \end{bmatrix} R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \vdots & \vdots & 1 \end{bmatrix} \tag{7}$$

$K > 1$ means $z_{gj} \mid \boldsymbol{\pi}_j \sim \text{Categorical}\left(\{-K, -K+1, \ldots, K\}; \boldsymbol{\pi}_j\right), \quad j \in \{1, \ldots, d\}$ has larger upper bound and lower bound. The π_j is a length-$(2K + 1)$ vector and represents probabilities for $z_{gj} = \{-K, -K+1, \ldots, K\}$ for $j \in \{1, \ldots, d\}$

In order to obtain the likelihood function, my first goal is to derive the probabilities of $\mathbb{P}\left(z_{gj} \mid \rho\right)$ which is a length-$(2K+1)^d$ vector because each z_{gj} can be chosen from $\{-K, \ldots, K\}$. In the general case, I set more bounds for each truncated normal distribution and integrate $\boldsymbol{w}_g$ to get $(2K+1)^d$ different probabilities in $\mathbb{P}\left(z_{gj} \mid \rho\right)$. Next, I should derive the likelihood $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g\right)$ when z_g has $(2K+1)^d$ different chosen vectors. I use multivariate normal distribution, truncated normal and Expression(4) to obtain corresponding likelihoods for different scenarios. It is really hard to estimate correlation matrix R with MLE because it's hard to guarantee the correlation matrix R is a a positive semidefinite matrix when I assume all parameters from a uniform distribution. For more general case, I propose to use MCMC to estimate correlation matrix R.

5.0 Simulations

5.1 First Set of Simulations in the Simple Case

I simulate data from the true model to estimate the parameter in a simple case when $d = 2$ index two different times, $K = 1$. The true model here refers to the copula model When I couple prior distributions .

(1) I set a correlation coefficient ρ^* in variance of $\hat{\boldsymbol{\beta}}_g$: $\boldsymbol{V}_g = \frac{1}{n} \begin{pmatrix} 1 & \rho^* \\ \rho^* & 1 \end{pmatrix}$ $n = 100$ is the sample size, $p = 10000$ is the number of genes.

(2) Then I can assume $\rho^* \sim \mathbf{U}[0, 0.5]$ because I have proved $\boldsymbol{V}_g$ is known. I also assume standard deviation for β_{g1} and β_{g2} are 0.2.

(3) I assume the correlation for w_{g1} and w_{g1} is ρ and simulate data as follows:

(i) I can draw w_{g1} and w_{g2} from a bivariate normal distribution with mean $\vec{0_2}$ and variance $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$

(ii) I will draw z_{g1} and z_{g2} based on the relationship between z_{g1} and w_{g1}, and z_{g2} and w_{g2}.

(iii) I will draw β_{g1} and β_{g2} from truncated normal distribution because z_{g1} and z_{g2} encode the sign of β_{g1} and β_{g2}.

(iv) I enable to draw $\hat{\boldsymbol{\beta}}_g$ from a bivariate normal distribution with mean $\boldsymbol{\beta}_g$ and variance V_g.

(4) I iterate 100 times to obtain 100 data sets and obtain $\boldsymbol{\beta}_g$ and $\boldsymbol{V}_g$ for $\hat{\boldsymbol{\beta}}_g$ in each data set. My goal is to estimate the parameter ρ in the copula model.

5.2 Plots and Interpretation

I choose three different $\rho \in \{0, 0.5, 0.75, 1\}$ to estimate the parameter ρ. For $\rho = 0, 0.5, 0.75$, I obtain the histograms of estimators for ρ as follows:

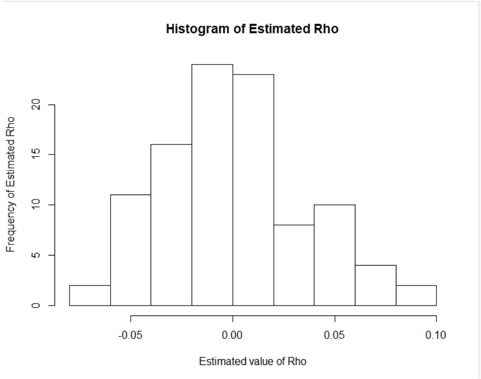

Figure 1: Histogram for $\rho(0)$

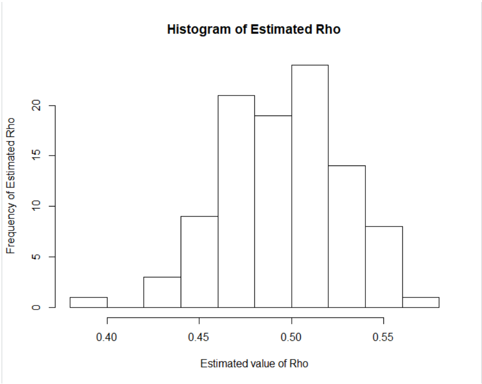

Figure 2: Histogram for $\rho(0.5)$

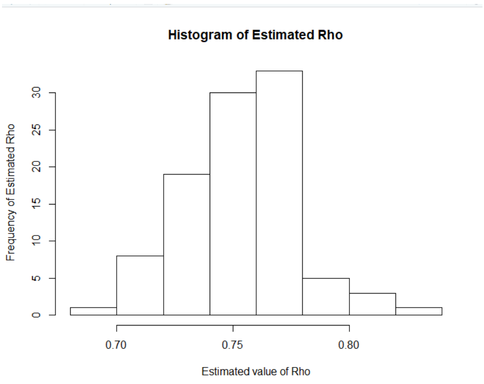

Figure 3: Histogram for $\rho(0.75)$

It's obvious that when $\rho = 0$, estimators for ρ mass around 0 from above plots. When $\rho = 0.5, 0.75$, estimators for ρ mass around 0.5 and 0.77. That is good to show estimators for ρ is very close to the initialized ρ. But when $\rho = 1$, all estimators for ρ is equal to 1. That means when latent variable $w_{g1} = w_{g2}$, the sign of the effect of β_{g1} and β_{g2} are the same. This data simulation in my copula model is flexible enough for all scenarios when initialized ρ is equal from -1 to 1. The copula model works well in simple case.

I find the estimators for ρ is very close to our expectation when we set a initialized correlation ρ.

5.3 Second Set of Simulations in the Simple Case

For the second set of simulations in the simple case. I would change the simulation procedures to test the generality of my estimator.

(1) I also set a correlation coefficient ρ^* in variance of $\hat{\boldsymbol{\beta}}_g$: $\boldsymbol{V}_g = \frac{1}{n} \begin{pmatrix} 1 & \rho^* \\ \rho^* & 1 \end{pmatrix}$ $n = 100$ is the sample size, $p = 10000$ is the number of genes.

(2) Then I can assume $\rho^* \sim \mathbf{U}[0, 0.5]$ because I have proved $\boldsymbol{V}_g$ is known. I also assume standard deviation for β_{g1} and β_{g2} are 0.2.

The difference begin with here

(3) I draw $\boldsymbol{\beta}_g$ from the following model: $\boldsymbol{\beta}_g \sim \pi_0\delta(0,0) + (1-\pi_0)\,N_2\left(\vec{0_2}, (\frac{2}{\sqrt{100}})^2\begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}\right)$.

Where $\tau \in \{0, 0.5, 1\}$ is the correlation coefficient for β_{g1} and β_{g2}, $\pi_0 = 0.7$.

(4) I simulate 100 data sets and observe 10000 $\hat{\boldsymbol{\beta}}_g$ and 10000 $\boldsymbol{V}_g$ for 10000 $\boldsymbol{\beta}_g$ in each data set. My goal is to estimate the parameter ρ in the second set of simulating data.

5.4 Plots and Interpretation

I choose three different $\tau \in \{0, 0.5, 1\}$ to estimate the parameters. For $\tau = 0, 0.5$, I obtain the histograms of estimators for ρ as follows:

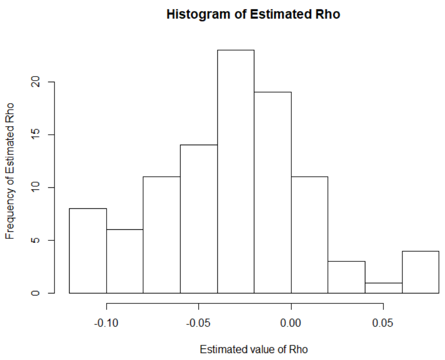

Figure 4: Histogram for $\tau(0)$

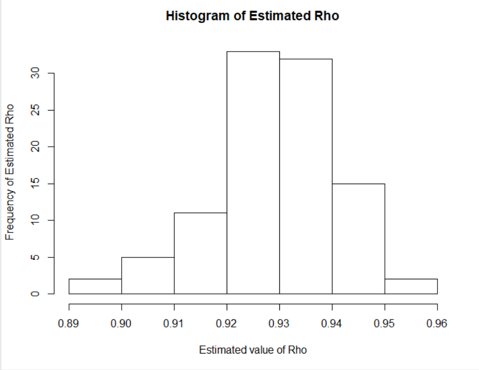

Figure 5: Histogram for $\tau(0.5)$

It's obvious that when $\tau = 0$, estimators for ρ mass around 0. When $\tau = 0.5$, estimators for ρ mass around 1. That is too much high and not consistent with my true model. But when $\tau = 1$, all estimators for ρ is equal to 1 again. That means when the effect of β_{g1}, β_{g2} are the same, the estimators for ρ also supports the effect sign of β_{g1}, β_{g2} are the same.

The estimators for ρ when $\tau = 0$ and $\tau = 1$ performs well in the second set of simulating data, but not make sense when $\tau = 0.5$. Now I will explore why the second set of simulations model does not work well when $\tau = 0.5$. We know the contour of bivariate normal distribution when correlation coefficient is 0.5 looks like below:

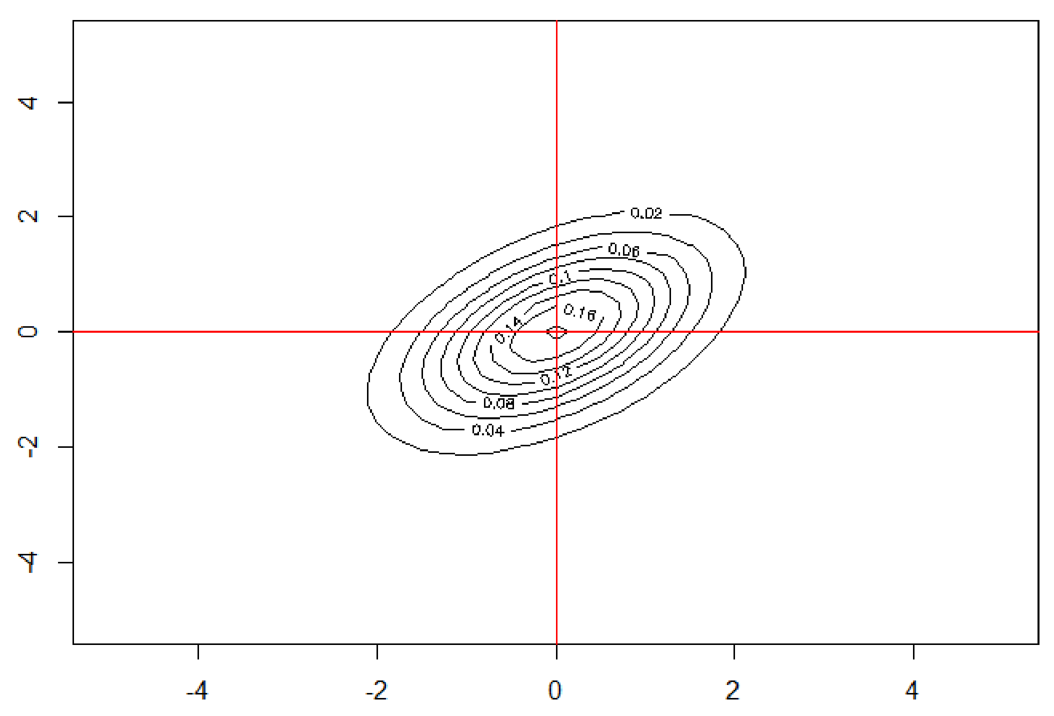

Figure 6: Contour for bivariate normal distribution with correlation(0.5)

When $\tau = 0.5$, estimators for ρ is too much large in the second set of simulating data. It means the second set of simulations model is less complicated and not flexible for all scenarios. That is because when $\tau = 0.5$, the contour of bivariate normal distribution is a rotate ellipse in level sets and the region in the first and third quadrants is larger than the region in the second and fourth quadrant. That means the effect size of the first and third quadrants respect to the noise is much larger than the effect size of the second and fourth quadrants. We will have more power to observe β_g from the first and third quadrants with same sign since these regions have a larger effect size than other two quadrants. β_{g1} and β_{g2} will be both positive or negative. The correlation between them will be positive all the time. That is why the estimators for ρ in second set of simulating data is larger than 0.5 or even close to 1 When $\tau = 0.5$. We will need a larger sample size If we want to observe the effects $\boldsymbol{\beta}_g$ from the second and fourth quadrants. It will violate the reality If we set a larger sample size.

5.5 Comparison of the Two Simulations

I find estimators for ρ are nearly the same in the true model and the second set of simulating data when $\rho = 0$ and $\rho = 1$ for the first set of simulations and $\tau = 0$ and $\tau = 1$ for the second set of simulations. When $\tau = 0.5$ or some values between -1 and 1, the estimators for ρ in the true model is more accurate than the second set of simulating data. That means the second set of simulating data only works for some special τ values while the true model is more flexible to work for all possible ρ values. That is mainly because when the τ chosen from -1 to 1 and is not equal to 0. Take $\tau = 0.5$ as an example, the contour of bivariate normal distribution is a 45° rotate ellipse in level sets and the effect size of the first and third quadrants respect to the noise is much larger than the effect size of the second and fourth quadrants.We will have more power to observe $\boldsymbol{\beta}_g$ from the first and third quadrants with same sign of effect since these regions have a larger effect size than other two quadrants. The final estimators for ρ will be larger and close to 1 in the second set of simulating data. We will not encounter such problem if we draw $\boldsymbol{\beta}_g$ from the true model.

6.0 Discussion and Future Direction

My copula Model works well and conclusion make sense in the simple case. There is still a lot of limitations I can optimize to improve my model. Firstly, my two set of simulations only focus on the simple case where $K = 1$ and $d = 2$, but I don't consider the general scenario where $d > 2$ and $K > 1$. I need to figure out a more general estimator for the more general case. Secondly, although I use MLE to estimate parameters in the simple case, it is really difficult to derive likelihood function when $d > 2$ and $K > 1$ and use maximum likelihood estimation to estimate all parameters in correlation matrix. I need to use another method to estimate correlation R for more general studying.

My model becomes more complicated and latent variable $\boldsymbol{w}_g$ will transfer from bivariate normal distribution to multivariate normal distribution when $d > 2$ and $K > 1$. It is difficult to estimate R with MLE because it's hard to guarantee the correlation matrix R is a a positive semidefinite matrix when I assume all parameters from a uniform distribution. My goal is to determine the re-lationship between sign of the effect $\boldsymbol{\beta}_g$ at different time. I propose to use MCMC to estimate R.

Step 1: I initialize a correlation matrix R.

Step 2: I generate $\boldsymbol{z}_g$ based on the same assumption we made in copula model.

That is $z_{gj} \mid \boldsymbol{\pi}_j \sim \text{Categorical}\Big(\{-K, -K+1, \ldots, K\}; \boldsymbol{\pi}_j\Big), \quad j \in \{1, \ldots, d\},\ g \in \{1, \ldots, p\},\ p = 10000$ is the number of genes.

Step 3: I use truncated normal distribution to draw $\boldsymbol{w}_g$ given $\boldsymbol{z}_g$.

That is $w_{gj} \mid z_{gj} \sim TN(0, R, \{x_{jk}, x_{j(k+1)}\})$, $g \in \{1, \ldots, p\}$, $k \in \{-K, \ldots, K\}$. Where x_{jk} and $x_{j(k+1)}$ are bounds in truncated normal distribution.

Step 4: I use Wishart distribution to draw the inverse of a new covariance matrix.

That is $R_0^{(-1)} \sim W_d\left(\frac{p}{2} + \delta, \frac{1}{2p}\sum_{g=1}^{p} \boldsymbol{w}_g \boldsymbol{w}_g^T + \delta I_d\right)$. Where R_0 is the new covariance and δ is a small number equal to 0.01, $p = 10000$ is the number of genes.

Step 5: I derive a new correlation matrix R^* given covariance matrix R_0.

That is R_n^*=diag($R_0^{(-\frac{1}{2})}$)R_0diag($R_0^{(-\frac{1}{2})}$).

Step 6: I take R_n^* as the new correlation matrix and plug it into the step 1 and iterate m times from the step 2 to step 6.

Step 7: I generated m correlation matrices $R_1^*, \ldots, R_m^*$ and tossed away the first b ($b < m$) as burnin, then I calculate the mean of the last $(m - b)$ correlation matrices as the estimators for R.

Appendix A Deriving Expression(4) with Bayes' Theorem

Before starting our assumption, I highly recommend to look through the article about Truncated normal ([1, page 20-26]), according to my assumption, we know that:

$$\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{\beta}_g \sim N_d\left(\boldsymbol{\beta}, \boldsymbol{V}_g\right), \quad g \in \{1, \dots, p\}$$

$$\boldsymbol{\beta}_g \mid z_{g1}, \dots, z_{gd} \sim TN\left(\mathbf{0}, \Sigma, [A_{(k)}, A_{(k+1)}]\right)$$

Where $[A_{(k)}, A_{(k+1)}]$ are the bound for Truncated Normal distribution and Σ is diagonal matrix as follows:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_d^2 \end{bmatrix}$$

While we know that $z_{gj} \mid \boldsymbol{\pi}_j \sim$ Categorical $\{(-K, -K+1, \dots, K\}; \boldsymbol{\pi}_j,) j \in \{1, \dots, d\}$

Σ will depend on Z Since z_{gj} encodes the sign and the expected magnitude of β_{gj}. Meanwhile, The sign and magnitude of β_{gj} depends on z_{gj}

We define $\boldsymbol{V}_g^{-1} = \boldsymbol{\Omega}_g$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid V_g, Z_{g1}, \dots Z_{gd}\right) = \int_{-\infty}^{0} \cdots \int_{0}^{\infty} \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{\beta}_g, \boldsymbol{V}_g\right) \mathbb{P}\left(\boldsymbol{\beta}_g \mid Z_{g1}, \dots, Z_{gd}\right) d\beta_g$$

$$= \frac{1}{\sqrt{\pi^{2d}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} \exp\left(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g\right) \int_{-\infty}^{\mathbf{0}} \cdots \int_{\mathbf{0}}^{\infty} \exp(\frac{-1}{2}\{\boldsymbol{\beta}_g^T\left(\boldsymbol{\Omega}_g + \Sigma^{-1}\right)\boldsymbol{\beta}_g - 2\boldsymbol{\beta}_g^T \Omega_g \hat{\boldsymbol{\beta}}_g\}) d\boldsymbol{\beta}_g$$

If we assume $\boldsymbol{A}_g = \boldsymbol{\Omega}_g + \Sigma^{-1}$

Then we have $\mu_g = \boldsymbol{A}_g^{-1}\Omega_g\hat{\boldsymbol{\beta}}_g$

Using complete square to transform above equation and we will obtain following expression:

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid V_g, Z_{g1}, \dots Z_{gd}\right) = \frac{1}{\sqrt{\pi^{2d}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g\mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T\boldsymbol{\Omega}_g\hat{\boldsymbol{\beta}}_g)$$

$$\int_{-\infty}^{\boldsymbol{0}} \cdots \int_{\boldsymbol{0}}^{\infty} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g\left(\boldsymbol{\beta}_g - \mu_g\right))d\boldsymbol{\beta}_g$$

For here we have recognized this is a generally form of a normal function with mean $\mu = \mu_g$ and variance $\sigma^2 = \boldsymbol{A}_g^{-1}$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, z_{g1}, \dots Z_{gd}\right) = \frac{(2\pi)^{(d/2)}}{\sqrt{|A_g|\pi^{2d}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g\mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T\boldsymbol{\Omega}_g\hat{\boldsymbol{\beta}}_g)$$

$$\int_{-\infty}^{\boldsymbol{0}} \cdots \int_{\boldsymbol{0}}^{\infty} \frac{\sqrt{|A_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g\left(\boldsymbol{\beta}_g - \mu_g\right))d\boldsymbol{\beta}_g$$

Appendix B Deriving Expression(5) with Bayes' Theorem

Our goal is to find Expression for $\mathbb{P}\left(W_{g1} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, w_{g2}, \ldots, w_{gd}\right)$ and according to Bayes' Theorem.

$$\begin{aligned}\mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, w_{g2}, \ldots, w_{gd}\right) &\propto \mathbb{P}\left(w_{g1} \mid w_{g2}, \ldots w_{gd}\right) \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, w_{g1} \ldots w_{gd}\right) \\ &= \sum_{k=-K}^{K} \mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, z_{g1} = k\right) \mathbb{P}\left(z_{g1} = k \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, w_{g1}\right)\end{aligned}$$

Next, I define $\mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, z_{g1} = k\right)$ as (1*), and $\mathbb{P}\left(z_{g1} = k \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, w_{g1}\right)$ as (2*)
For (1*)

$$\mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, z_{g1} = k\right) \propto \quad \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid w_{g(-1)}, z_{g1} = k, w_{g1}\right) \mathbb{P}\left(w_{g1} \mid W_{g(-1)}, z_{g1} = k\right)$$

Since $\hat{\boldsymbol{\beta}}$ is not a function with w_{g1}. Therefore, $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid w_{g(-1)}, z_{g1} = k, w_{g1}\right)$ can be regarded as $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid w_{g(-1)}, z_{g1} = k\right)$ which should be a constant.

In this way, $\mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}, z_{g1} = k\right) \propto \mathbb{P}\left(w_{g1} \mid w_{g(-1)}, z_{g1} = k\right)$.

We know that for conditional distribution $\left(w_{g1} \mid w_{g2} \ldots w_{gd}\right) \sim N\left(\boldsymbol{\mu}_{-1}, \boldsymbol{\sigma}^2_{-1}\right)$. I also define the correlation matrix as R.

$$R_{d,d} = \begin{bmatrix} 1 & R_{12} & \cdots & R_{1d} \\ R_{21} & 1 & \cdots & R_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ R_{d1} & R_{d2} & \cdots & 1 \end{bmatrix} \tag{8}$$

And we divide the matrix into 4 different parts.

$$\Sigma_{11} = 1$$

$$\Sigma_{12} = \begin{bmatrix} R_{12} & \cdots & R_{1d} \end{bmatrix} \tag{9}$$

$$\Sigma_{21} = \begin{bmatrix} R_{21} \\ \vdots \\ R_{d1} \end{bmatrix} \tag{10}$$

$$\Sigma_{22} = \begin{bmatrix} 1 & R_{23} & \cdots & R_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ R_{d2} & \cdots & 1 & \end{bmatrix} \tag{11}$$

We obtain that $\boldsymbol{\mu}_{-1} = \Sigma_{12}\Sigma_{22}^{-1}(w_{-1})$ and $\boldsymbol{\sigma}^2_{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{(-1)}\Sigma_{21} = 1 - \Sigma_{12}\Sigma_{22}^{(-1)}\Sigma_{21}$ with the conditional distribution properties.

$$\mathbb{P}\left(w_{g1} \mid w_{g(-1)}, z_{g1} = k\right) \propto \exp(\frac{-(w_{g1} - \Sigma_{12}\Sigma_{22}^{-1}w_{(g-1)})^2}{2(1 - \Sigma_{12}\Sigma_{22}^{(-1)}\Sigma_{21})} 1\{w_{g1} \in A_k\}.$$

Therefore, we obtain that $\mathbb{P}\left(w_{g1} \mid \hat{\boldsymbol{\beta}}_g, \boldsymbol{V}_g, w_{g2}, \ldots, w_{gd}\right) \propto \sum_{k=-K}^{K} \pi_k TN\left(\mu_{-1}, \sigma^2_{-1}, [A_k, A_{k+1}]\right)$ For (2*)

$$\mathbb{P}\left(z_{g1} = k \mid \hat{\boldsymbol{\beta}}_g, w_{g(-1)}\right) \propto \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid z_{g1} = k, w_{g(-1)}\right)\mathbb{P}\left(z_{g1} = k \mid w_{g(-1)}\right)$$

$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid z_{g1} = k, w_{g(-1)}\right)$ is what I have completed in the expression(4) and same to the probability $\mathbb{P}\left(\boldsymbol{\beta}_g \mid w_{g1}, \ldots, w_{gd}\right)$, since $z_{g1} = K$ will determine which bound does w_{g1} belongs to.

Appendix C Processes of Estimating R for Simple Case

My goal is deriving the likelihood function $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right) = \sum \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g\right)\mathbb{P}\left(\boldsymbol{z}_g \mid \rho\right)$. Where ρ is in correlation matrix R.

$$R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{12}$$

It's easy for us to calculate $\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g\right)$ with Expression(4), all we need to do is find $\mathbb{P}\left(\boldsymbol{z}_g \mid \rho\right)$, however, it's really hard to derive this when k is large, therefore, we can begin with a simple case.

Suppose d=2,k=1 and x_{11}, x_{12} are bounds in truncated normal distribution for w_{g1} while x_{21}, x_{22} are bounds in truncated normal distribution for w_{g2}, therefore, $\mathbb{P}\left(\boldsymbol{z}_g = (0,0) \mid \rho\right)$ can be calculated as follows:

$\mathbb{P}\left(\boldsymbol{z}_g = (0,0) \mid \rho\right) = \mathbb{P}\left(w_{g1} \in [x_{11}, x_{12}], w_{g2} \in [x_{21}, x_{22}] \mid \rho\right)$

For $\boldsymbol{z}_g = (0,0)$

$\mathbb{P}\left(\boldsymbol{z}_g = (0,0) \mid \rho\right) = sadmvn(c(x_{11}, x_{21}), c(x_{12}, x_{22}), mean = c(0,0), R)$

We can find other 8 different probabilities with same procedure.

For $\boldsymbol{z}_g = (0,1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (0,1) \mid \rho\right) = sadmvn(c(x_{11}, x_{22}), c(x_{12}, Inf), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (0,-1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (0,-1) \mid \rho\right) = sadmvn(c(x_{11}, -Inf), c(x_{12}, x_{21}), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (1,0)$

$\mathbb{P}\left(\boldsymbol{z}_g = (1,0) \mid \rho\right) = sadmvn(c(x_{12}, x_{21}), c(Inf, x_{22}), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (1,1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (1,1) \mid \rho\right) = sadmvn(c(x_{12}, x_{22}), c(Inf, Inf), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (1,-1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (1,-1) \mid \rho\right) = sadmvn(c(x_{12}, -Inf), c(Inf, x_{21}), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (-1,0)$

$\mathbb{P}\left(\boldsymbol{z}_g = (-1,0) \mid \rho\right) = sadmvn(c(-Inf, x_{21}), c(x_{11}, x_{22}), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (-1,1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (-1,1) \mid \rho\right) = sadmvn(c(-Inf, x_{22}), c(x_{11}, Inf), mean = c(0,0), R)$

For $\boldsymbol{z}_g = (-1,-1)$

$\mathbb{P}\left(\boldsymbol{z}_g = (-1,-1) \mid \rho\right) = sadmvn(c(-Inf, -Inf), c(x_{11}, x_{21}), mean = c(0,0), R)$

$\boldsymbol{L}_{(p*9)}$ is a p by 9 fixed matrix, where $g = 1, \ldots, p$,which is not function of ρ as follows:

$$\boldsymbol{L}_{(p*9)} = \begin{bmatrix} \mathbb{P}\left(\hat{\beta}_1 \mid z_1 = (0,0)\right), & \cdots, & \mathbb{P}\left(\hat{\beta}_1 \mid z_1 = (-1,-1)\right) \\ \mathbb{P}\left(\hat{\beta}_2 \mid z_2 = (0,0)\right), & \cdots, & \mathbb{P}\left(\hat{\beta}_2 \mid z_2 = (-1,-1)\right) \\ \vdots & \vdots & \vdots \\ \mathbb{P}\left(\hat{\beta}_p \mid z_p = (0,0)\right), & \cdots, & \mathbb{P}\left(\hat{\beta}_p \mid z_p = (-1,-1)\right) \end{bmatrix} \tag{13}$$

$$\mathbb{P}\left(\hat{\beta}_1 \mid z_1 = (0,0)\right) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{V}_g|}} \exp\left(-\frac{1}{2}\hat{\beta}_1{}^T \boldsymbol{V}_g^{(-1)} \hat{\beta}_1\right)$$

$\Omega_g = \boldsymbol{V}_g^{-1}$

$\boldsymbol{A}_g = \boldsymbol{\Omega}_g + \Sigma^{-1}$

$\mu_g = \boldsymbol{A}_g^{-1}\boldsymbol{\Omega}_g\hat{\boldsymbol{\beta}}_g.$

When one of $z_{gj} = 0$ in the simple case, we need to recalculate the expression for likelihood. For $z_{g2} = 0$, e_1 is a vector which first element is 1 while others are 0 and (σ^2_{1k}) is the variance for β_{g1}, for likelihood we have recognized this is a generally form of normal function and then transform this equation with complete square to find the mean and variance for normal distribution.

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, z_{g1}, z_{g2}\right) = \frac{1}{\sqrt{2\pi^3|\boldsymbol{V}_g|\sigma^2_{1k}}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_0^\infty \exp\left(-\frac{1}{2}\{\beta^2_{g1}\left(e_1{}^T\boldsymbol{\Omega}_g e_1 + \frac{1}{\sigma^2_{1k}}\right) - 2\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g e_1 \beta_{g1}\}\right) d\boldsymbol{\beta}_{g1}$$

The only difference between this and the former equation is that we only need to define $A_1 = e_1{}^T\Omega_g e_1 + \frac{1}{\sigma^2_{1k}}$ and $B_2 = \hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g e_1$ and we can easily find that the integration part is a normal distribution with mean=$\frac{B}{A}$ and Variance=$\frac{1}{A}$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, z_{g1}, z_{g2}\right) = \frac{1}{\sqrt{\pi^2|\boldsymbol{V}_g|\sigma_{1k}^2 A_1}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g + \frac{B_1^2}{2A_1})$$

$$\int_0^\infty \frac{\sqrt{A_1}}{\sqrt{2\pi}} \exp\left(-\frac{A_1}{2}(\beta_{g1} - \frac{B_1}{A_1})^2\right) d\beta_{g1}$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (1,0)\right) = \frac{1}{\sqrt{\pi^2|\boldsymbol{V}_g|\sigma_{1k}^2 A_1}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g + \frac{B_1^2}{2A_1})$$

$$(1 - \int_0^\infty \frac{\sqrt{A_1}}{\sqrt{2\pi}} \exp\left(-\frac{A_1}{2}(\beta_{g1} - \frac{B_1}{A_1})^2\right) d\beta_{g1})$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (-1,0)\right) = \frac{1}{\sqrt{\pi^2|\boldsymbol{V}_g|\sigma_{1k}^2 A_1}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g + \frac{B_1^2}{2A_1})$$

$$\int_0^\infty \frac{\sqrt{A_1}}{\sqrt{2\pi}} \exp\left(-\frac{A_1}{2}(\beta_{g1} - \frac{B_1}{A_1})^2\right) d\beta_{g1}.$$

For $z_{g1} = 0$, we only need to change e_1 to e_2, a vector second element is 1 while others are 0, and A and B in my last expression.

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (0,1)\right) = \frac{1}{\sqrt{\pi^2|\boldsymbol{V}_g|\sigma_{2k}^2 A_2}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g + \frac{B_2^2}{2A_2})$$

$$(1 - \int_0^\infty \frac{\sqrt{A_2}}{\sqrt{2\pi}} \exp\left(-\frac{A_2}{2}(\beta_{g1} - \frac{B_2}{A_2})^2\right) d\beta_{g1})$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (0,-1)\right) = \frac{1}{\sqrt{\pi^2|\boldsymbol{V}_g|\sigma_{2k}^2 A_2}} exp(-\frac{1}{2}\hat{\boldsymbol{\beta}}_g^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g + \frac{B_2^2}{2A_2})$$

$$\int_0^\infty \frac{\sqrt{A_2}}{\sqrt{2\pi}} \exp\left(-\frac{A_2}{2}(\beta_{g1} - \frac{B_2}{A_2})^2\right) d\beta_{g1}$$

Where $A_2 = e_2{}^T \Omega_g e_2 + \frac{1}{\sigma_{2k}^2}$ and $B_2 = \hat{\beta}_g^T \Omega_g e_2$.

When it comes to $\boldsymbol{z}_g \neq (0,0)$ we are able to use expression(4) to derive their likelihood:

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{V}_g, z_{g1}, \dots z_{gd}\right) = \frac{(2\pi)^{(d/2)}}{\sqrt{|\boldsymbol{A}_g|\pi^{2d}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g \mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_{-\infty}^{\mathbf{0}} \cdots \int_{\mathbf{0}}^{\infty} \frac{\sqrt{|A_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g \left(\boldsymbol{\beta}_g - \mu_g\right)) d\boldsymbol{\beta}_g.$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (1,1)\right) = \frac{(2\pi)^{(2/2)}}{\sqrt{|\boldsymbol{A}_g|\pi^{4}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g \mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_{0}^{\infty} \int_{0}^{\infty} \frac{\sqrt{|\boldsymbol{A}_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g \left(\boldsymbol{\beta}_g - \mu_g\right)) d\boldsymbol{\beta}_g.$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (-1,-1)\right) = \frac{(2\pi)^{(2/2)}}{\sqrt{|\boldsymbol{A}_g|\pi^{4}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g \mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_{-\infty}^{0} \int_{-\infty}^{0} \frac{\sqrt{|\boldsymbol{A}_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g \left(\boldsymbol{\beta}_g - \mu_g\right)) d\boldsymbol{\beta}_g.$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (1,-1)\right) = \frac{(2\pi)^{(2/2)}}{\sqrt{|\boldsymbol{A}_g|\pi^{4}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g \mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_{0}^{\infty} \int_{-\infty}^{0} \frac{\sqrt{|\boldsymbol{A}_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g \left(\boldsymbol{\beta}_g - \mu_g\right)) d\boldsymbol{\beta}_g$$

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \boldsymbol{z}_g = (-1,1)\right) = \frac{(2\pi)^{(2/2)}}{\sqrt{|\boldsymbol{A}_g|\pi^{4}|\boldsymbol{V}_g|\sigma_1^2\sigma_2^2}} exp(\frac{1}{2}\mu_g^T \boldsymbol{A}_g \mu_g - \frac{1}{2}\hat{\boldsymbol{\beta}}_g{}^T \boldsymbol{\Omega}_g \hat{\boldsymbol{\beta}}_g)$$

$$\int_{-\infty}^{0} \int_{0}^{\infty} \frac{\sqrt{|\boldsymbol{A}_g|}}{(2\pi)^{(d/2)}} \exp(\frac{-1}{2}\left(\boldsymbol{\beta}_g - \mu_g\right)^T \boldsymbol{A}_g \left(\boldsymbol{\beta}_g - \mu_g\right)) d\boldsymbol{\beta}_g$$

We know in simple case, $\pi(\rho)$ is a vector, including 9 probabilities with $0 \leq \pi(\rho) \leq 1$, and not a function of g.

$$\pi(\rho) = \left[\mathbb{P}\left(\boldsymbol{z}_g = (0,0) \mid \rho\right), \quad \cdots, \quad \mathbb{P}\left(\boldsymbol{z}_g = (-1,-1) \mid \rho\right)\right] \tag{14}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (0,0) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{11}}^{x_{12}} \int_{x_{21}}^{x_{22}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (0,1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{11}}^{x_{12}} \int_{x_{22}}^{\infty} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (0,-1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{11}}^{x_{12}} \int_{-\infty}^{x_{21}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (1,0) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{12}}^{\infty} \int_{x_{21}}^{x_{22}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (1,1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{12}}^{\infty} \int_{x_{22}}^{\infty} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (1,-1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{x_{12}}^{\infty} \int_{-\infty}^{x_{21}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (-1,0) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{-\infty}^{x_{11}} \int_{x_{21}}^{x_{22}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (-1,1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{-\infty}^{x_{11}} \int_{x_{22}}^{\infty} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

$$\mathbb{P}\left(\boldsymbol{z}_g = (-1,-1) \mid \rho\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)}) \int_{-\infty}^{x_{11}} \int_{-\infty}^{x_{21}} \left(w_{g1}^2 - 2\rho w_{g1} w_{g2} + w_{g2}^2\right) \boldsymbol{dw_{g2} dw_{g1}}$$

When we consider the maximum likelihood function:

$$L(\rho) = \prod_{g=1}^{p} \mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right)$$

$$Log\,(L(\rho)) = \sum_{g=1}^{p} Log\{\mathbb{P}\left(\hat{\boldsymbol{\beta}}_g \mid \rho\right)\} = \sum_{g=1}^{p} Log\left(L_{(g'srow)}^{T} * \pi(\rho)\right)$$

Therefore, we need to generate some functions to get the log-likelihohd function in R, I mainly use the instructions in ([3]).

(1) I write a function that generates $\boldsymbol{L}$ given $\hat{\beta}_1 \ldots, \hat{\beta}_p, V_1, \ldots, V_p, \sigma_1^2, \sigma_2^2$.

(2)I write a function that computes the log-likelihood given ρ and L.

(3) I write a function that simulates data given π_0 and ρ^*.

(4) I use One dimensional optimization function to test my estimators ρ.

Appendix D Related Theories

D.1 Gibbs Sampling

Gibbs sampling or a Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm for obtaining a sequence of observations which are approximated from a specified multivariate probability distribution, when direct sampling is difficult. This sequence can be used to approximate the joint distribution (e.g., to generate a histogram of the distribution); to approximate the marginal distribution of one of the variables, or some subset of the variables (for example, the unknown parameters or latent variables); or to compute an integral (such as the expected value of one of the variables). Typically, some of the variables correspond to observations whose values are known, and hence do not need to be sampled. Gibbs sampling, in its basic incarnation, is a special case of the Metropolis–Hastings algorithm. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution.

D.2 Truncated Normal Distribution

The truncated normal distribution is the probability distribution derived from that of a normally distributed random variable by bounding the random variable from either below or above (or both). The truncated normal distribution has wide applications in statistics and econometrics. For example, it is used to model the probabilities of the binary outcomes in the probit model and to model censored data in the to bit model.

Suppose X has a normal distribution with mean μ and variance σ^2 and lies within the interval (a,b) with $-\infty <= a < b <= \infty$. Then X conditional on $a < X < b$ has a truncated normal distribution. Its probability density function,f,fora¡=x¡+b,is given by:

$$f\,(x;\mu,\sigma,a,b) = \frac{1}{\sigma}\frac{\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma})-\Phi(\frac{a-\mu}{\sigma})}$$, and by f=0 for otherwise.

Here,$\phi(\xi) = \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}\xi^2)$ is the probability density function of the standard normal distribution and $\Phi(.)$ is its cumulative distribution function:

$$\Phi(x) = \frac{1}{2}\left(1 + erf(\frac{x}{\sqrt{2}})\right)$$

By definition,if b=∞,then $\Phi(\frac{b-\mu}{\sigma}) = 1$,and similarly,if a=$-\infty$,then $\Phi(\frac{a-\mu}{\sigma}) = 0$.

D.3 Wishart Distribution

The Wishart distribution is a generalization to multiple dimensions of the gamma distribution. It is a family of probability distributions defined over symmetric, nonnegative-definite matrix-valued random variables (“random matrices”). These distributions are of great importance in the estimation of covariance matrices in multivariate statistics. In Bayesian statistics, the Wishart distribution is the conjugate prior of the inverse covariance-matrix of a multivariate-normal random-vector.

Suppose G is a p by n matrix,each column of which is independently drawn from a p-variate normal distribution with zero mean:

$$G_i = \left(g_i^1, \ldots, g_i^p\right)^T \sim N_p\,(0, V).$$

Then the Wishart distribution is the probability distribution of the p by p random matrix. $S = GG^T = \sum_{i=1}^{n} G_i G_i^T$ known as the scatter matrix. One indicates that S has that probability distribution by writing.

$$S \sim W_p\,(V, n)$$

The positive integer n is the number of degrees of freedom.Sometimes this is written $w\,(V, p, n)$.For $n \geq p$ the matrix S is invertible with probability 1 if V is invertible.

Bibliography

[1] John Burkardt. The truncated normal distribution, October 2014.

[2] Catherine Stanhope Chris McKennan, Katherine Naughton. Longitudinal data reveal strong genetic and weak non-genetic components of ethnicity-dependent blood DNA methylation levels. *Epigenetics*, 24(1):1–30, 9 2020.

[3] Peter. Dalgaard. Introductory statistics with r, April 2008.

[4] FILIP ERNTELL. ON MODELING INSURANCE CLAIMS USING COPULAS, 2013.

[5] Tyler J. Gorrie-Stone. DNA Methylation: Methods and Analyses, September 2019.

[6] Martin Haugh. An introduction to copulas, April 2016.

[7] Barbara Choroś Rustam Ibragimov and Elena Permiakova. Copula Theory and Its Applications, May 2010.

[8] Aristidis K. Nikoloulopoulos. Copula-based models for multivariate discrete response data. In *Copulae in Mathematical and Quantitative Finance*, pages 231–249. Springer Berlin Heidelberg, 2013.

[9] Stephens and Matthew. False discovery rates: a new deal. *Biostatistics*, 18(2):275–294, 10 2016.