# New Statistical Approaches to Estimating Mixture Models with Application in Anti-Cancer Drug Studies

A THESIS SUBMITTED TO

THE UNIVERSITY OF KENT AT CANTERBURY

IN THE SUBJECT OF STATISTICS

FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY BY RESEARCH

By

Tong Wang

September 2022

University of Kent

DOCTORAL THESIS

# New Statistical Approaches to Estimating Mixture Models with Application in Anti-Cancer Drug Studies

*Author:*

Tong WANG

*Supervisor:*

Prof. Jian ZHANG

School of Mathematics, Statistics & Actuarial Science

September, 2022

# *Acknowledgements*

I would like to give the deepest gratitude to my supervisor Professor Jian Zhang, for his great patience and constructive guidance throughout my PhD studies. Unselfishly imparting his knowledge and rigorous academic attitude make me benefit quite much not only in my research career but also in my life, without which this thesis could not have been accomplished. I would also like to express great thanks to my second supervisor Dr James Bentham, who gave me a lot of encouragement and helpful suggestions during my PhD life.

I also want to express sincere gratitude to the whole department and staff in the School of Mathematics, Statistics and Actuarial Science for their help and support during my study life at the University of Kent. I would like to give special thanks to my dear friends Alex Diana, Aniketh Pittea, Jose A. Perusquia, Zhenhao He and Zhiyi Xia who made my PhD life memorable and colourful.

Last but not least, I would like to express the greatest respect and love to my mother, to whom this PhD thesis is dedicated. Her unconditional love, understanding and support give me the power to rely on.

# Abstract

When confronted with applications to real data problems, it is always challenging to simultaneously deal with potential group structures, high-dimensional features and the relationship between predictors and response variables. Most of the time missing data exist across the whole dataset, which makes the problems even more tricky. Meanwhile, with the advent of big data and high-throughput technology, the dimension of the given data could easily exceed the sample size, which places the ordinary linear regression into a difficult position where the normal equation is degenerate and traditional statistical techniques cannot be used properly. Notwithstanding, generally speaking, there is only a small part of variables being informative to the needs of researchers by significantly affecting the dependent variables. To address these issues, we develop a model to realise classification, variable selection and parameter estimation simultaneously in this thesis. This model also shows flexibility and inclusiveness to datasets with missingness. Moreover, by introducing the $l_q-$norm penalty to tune the sparsity level to the specific needs of researchers, our methodology has been improved further.

With the help of Bayesian Information Criterion, we can specify the number of components and degree of penalty for this modelling. After that, the uses of marginal analysis and the k-means clustering method facilitate the following application to whole datasets by realising a dimension reduction purpose. In the application to the anti-cancer drug and screened gene expression data, our methodology shows good abilities for clustering drugs into a finite number of groups and screening out the related genes which play significant roles in configuring the corresponding groups. With our specific enhancements to the model, including missingness indication and adjustable sparsity level, our methodology has the potential to be applied to a wide range of datasets in the scientific area, including but not limited to economics, finance, biology, and physics. Based on the above applications, we also propose another method to determine the number of components in a mixture model, which provides an alternative view on the clustering problem.

Afterwards, we examine the inherent skewness of given data by resorting to skew normal distributions. After adaptations to the traditional skew normal density function, we success-

fully estimate the parameters in a skew normal distribution under different skewness scenarios. The asymptotic distributions for the MLE estimates of our skew normal distribution are also obtained with detailed proofs attached in the Appendix. Meanwhile, some intriguing asymptotic properties behind our skew normal function are discussed later in this chapter. Lastly, we propose the four-piece distribution family for skew normal mixture models to consider the group structure, which shows a good estimation accuracy in the following simulation studies. From these simulations, the above models have been verified as a complement to the existing R package *mclust* which is popular for handling model-based clustering, classification, and density estimation problems.

# Contents

**Acknowledgments** i

**Abstract** ii

**List of Figures** viii

**List of Tables** xii

**1 Introduction** 1

1.1 Cancer and Genes . . . 1

1.2 Cancer Drug Data . . . 2

1.3 Challenges and Contributions . . . 3

1.4 Organisation of The Thesis . . . 4

**2 Methodological Background** 5

2.1 Curse of Dimensionality . . . 5

2.2 LASSO and Extensions . . . 5

2.2.1 Original LASSO . . . 6

2.2.2 Ridge Regression . . . 7

2.2.3 Elastic Net . . . 8

2.2.4 Adaptive LASSO . . . 9

2.2.5 LASSO-OLS Hybrid, Thresholded LASSO and Relaxed LASSO . . . 12

2.2.6 Package *glmnet* . . . 13

2.3 Model-based Clustering and Mixture Model . . . 14

2.3.1 Mixture Regression Model . . . 16

2.4 Expectation–Maximisation Algorithm . . . 16

2.4.1 Introduction . . . 16

2.4.2 E-step . . . 17

2.4.3 M-step . . . 18

2.4.4 Monotonicity . . . 18
2.5 k-means Clustering . . . 20
2.5.1 Introduction . . . 20
2.5.1.1 Assign Data Point . . . 20
2.5.1.2 Update Centroid . . . 20
2.5.1.3 Stopping Criteria . . . 20
2.5.1.4 The Choice of $k$ . . . 21
2.6 BIC and K-fold Cross-Validation . . . 22
2.6.1 Introduction to BIC . . . 22
2.6.2 Introduction to K-fold Cross-Validation . . . 22
2.7 Skew Normal Distribution . . . 23
2.8 Fisher Information and Asymptotic Distribution of MLE . . . 25
2.9 Big-O and Small-O Notations in Probability . . . 25
2.10 Some Other Knowledge . . . 27
2.10.1 Markov's Inequality . . . 27
2.10.2 Chebyshev's Inequality . . . 27
2.10.3 Slutsky's Theorem . . . 28

**3 Multivariate Mixture Regression Models (with and without Missingness) 29**
3.1 Complete Case Analysis by Penalised Likelihood Fusion . . . 29
3.1.1 Introduction . . . 29
3.1.2 Preprocess Data . . . 29
3.1.3 Preliminary Analysis of Group Structure . . . 30
3.1.4 Penalised Likelihood Fusion . . . 32
3.1.5 The Choices of $\lambda$ and $K$ . . . 35
3.1.6 Simulation . . . 35
3.1.7 Real Data Performance . . . 37
3.1.7.1 Individual LASSO . . . 37
3.1.7.2 Marginal Likelihood Fusion . . . 37
3.1.7.3 Applying to the Full Model . . . 40
3.1.7.4 Cross-Validation and Comparison with *mclust* . . . 40
3.1.8 Conclusions . . . 44
3.2 Incomplete Case with Missingness . . . 44
3.2.1 Algorithm and Methodology . . . 47
3.2.1.1 Thresholding the Regression Coefficients . . . 53
3.2.1.2 The Choices of $\lambda_q$ and $K$ . . . 54

3.2.1.3 Combine $q = 1$ Case and $q = 0.5$ Case . . . 54
3.2.2 Initialisation . . . 55
3.2.3 Simulation Studies . . . 55
3.2.3.1 Selection Accuracy Evaluation . . . 55
3.2.3.2 Simulations on $n$, $p$, $J$ for $q = 1$ Case . . . 56
3.2.3.3 Compare the Performances of $q = 0.5$ Case and $q = 1$ Case . . . 60
3.2.3.4 Evaluate the Efficacy of the Warm Start . . . 60
3.2.4 Numerical Results on Anti-Cancer Drug Data . . . 61
3.2.5 Discussion and Conclusion . . . 64
3.3 Discussions about Imputation and Component Number Selection . . . 67
3.3.1 Introduction . . . 67
3.3.2 Algorithm . . . 68
3.3.2.1 Regression and Imputation . . . 68
3.3.2.2 Estimate $K$ . . . 69
3.3.3 Simulation . . . 70
3.3.4 Complete-Case Problem . . . 72

**4 Estimation of Skew Normal Model** **75**
4.1 Introduction . . . 75
4.2 Methodology . . . 76
4.2.1 Location-Scale Skew Normal Model . . . 76
4.2.2 Estimation with EM Algorithm . . . 77
4.3 $\alpha$-Penalised Estimation . . . 86
4.3.1 Cross-Validation for Selecting $\lambda$ . . . 87
4.3.2 Asymptotic Properties of Cross-Validated $\lambda_{CV}$ . . . 88
4.3.3 Asymptotic Distributions of $\alpha-$Penalised MLE . . . 94
4.4 Simulation Studies . . . 97
4.4.1 The Simulation Study for Location-Scale Skew Normal . . . 97
4.4.2 The Simulation Studies for $\alpha$-Penalised Skew Normal . . . 99
4.5 Anti-Cancer Drug Data Application . . . 100

**5 Estimating Skew Normal Mixtures** **105**
5.1 Four-Piece Distribution Family . . . 105
5.1.1 Density and Likelihood . . . 105
5.1.2 E-Step and M-Step in EM Algorithm . . . 108
5.1.3 Model Selection . . . 113

5.2 Simulation Studies . . . . . 114

**6 Conclusions and Future Works** **117**

6.1 Conclusions . . . . . 117

6.2 Future Works . . . . . 118

**Appendix** **120**

**Bibliography** **203**

# List of Figures

2.2.1 The coefficient-path plot against the $l_1-$norm of the coefficient vector. Y axis indicates the coefficient value of the variable, and X axis is the value of bound $t$. 7

2.2.2 The contours of residual sum of squares $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ are plotted with red ellipses. $\hat{\boldsymbol{\beta}}$ is the OLS estimate $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$. The LASSO case is on the left with the constraint area $|\beta_1| + |\beta_2| \leq t$. The ridge case is on the right with the constraint area $\beta_1^2 + \beta_2^2 \leq t^2$. (Tibshirani, 1996; James et al., 2013) . . . . . 8

2.2.3 The number of remaining genes for different drugs after applying the Elastic Net. We illustrate the variable selection effects (by counting how many genes are screened out for each drug) under different $\alpha$s. Y axis indicates the number of non-zero-coefficient variables (genes) after the Elastic Net screening. . . . . 10

2.2.4 The number of genes for different anti-cancer drugs after screenings with the original LASSO (blue line) and Adaptive LASSO (orange line). Y axis indicates the number of non-zero variables after the screening. . . . . 11

2.5.1 For different number of groups $k$s(from 1 to 10 groups), this line chart shows the average distance to the within-group centroid for each $k$. Here the within-group centroid is the imaginary or real data point locating at the centre of a cluster. . 21

3.1.1 The regression coefficient plot from LASSO regression for 10 drugs. All 17737 genes (unlabelled) run from top to bottom in each vertical long line. Small line segments pointing to the right indicate positive weights, and to the left, negative weights. . . . . 30

3.1.2 79 anti-cancer drugs are grouped into 20 clusters. Darker green the edge between two nodes is, the greater weight they have. For the name tags, different colours are assigned to different clusters. . . . . 31

3.1.3 The coefficient plots for all three drug groups to show the performance of penalised likelihood fusion model. Red hollow triangular markers are the preset true values of $\boldsymbol{\beta}$. Black hollow circular markers are our estimates $\hat{\boldsymbol{\beta}}$. . . . . 36

3.1.4 The rBIC values for different genes from high to low with red vertical line classifying them into two groups based on k-means clustering method. . . . . 40
3.1.5 The clustering result for anti-cancer drugs. There are 9 groups in total with each group displayed in a different colour. . . . . 41
3.1.6 The coefficient plot for each anti-cancer drug group with 5% significance level bounds (red lines). . . . . 42
3.1.7 The exhibition of significant genes related to specific anti-cancer drug groups from the screening results in Figure 3.1.6. . . . . 43
3.1.8 From test 1 to test 5, the likelihood values and numbers of groups for the results of our penalised likelihood fusion modelling and *mclust* respectively. The line graphs are the likelihood values $l_v\left(\hat{\Phi} \mid \mathbf{Y}_v\right)$ and $l_v^{\star}\left(\hat{\Phi}^{\star} \mid \mathbf{Y}_v\right)$. Blue one is from our modelling, yellow one is from *mclust*. The triangle (our modelling) and circle (*mclust*) markers show the comparison of group numbers. . . . . 44
3.2.1 The sensitivity table, line chart and box plot for 100 trials under the numbers of drugs $J = 50$ (blue) and $J = 70$ (orange). . . . . 57
3.2.2 The sensitivity table, line chart and box plot for 100 trials under the sample sizes $n = 200$ (blue) and $n = 300$ (orange). . . . . 58
3.2.3 The sensitivity table, line chart and box plot for 100 trials under the numbers of genes $p = 1000$ (blue) and $p = 1300$ (orange). . . . . 59
3.2.4 The Relative Error comparison for $q = 0.5$ case (with the warm start) and $q = 1$ case against the increase of underlying coefficient sparsity (from 5 to 50 active coefficients). . . . . 61
3.2.5 Coefficient bar-plot for each group. Selected genes are labelled and run from top to bottom in each vertical long line. Small hollow bars pointing to the right indicate positive weights, and those to the left indicate negative weights. . . . . 63
3.3.1 Box-and-whisker plot for chosen $\hat{K}$ from 50 sample trials. MAE made use of the underlying true missing entries. . . . . 71
3.3.2 Box-and-whisker plot for chosen $\hat{K}$ from 100 sample trials. MAE made use of the entries adjacent to missing ones. . . . . 72

4.3.1 The box plots for $\lambda_{CV}/n$ and trend chart for average (variance) of $\lambda_{CV}/n$ under **Case 1**. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/n$ as $n$ increases. . . . . 90
4.3.2 The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 2. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases. . . . . 91

4.3.3 The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 3. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases. . . . . 92

4.3.4 The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 4. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases. . . . . 93

4.4.1 The box plots for estimated location parameter $\mu$, scale parameter $\sigma$ and shape parameter $\alpha$ after 100 times of resampling and estimation with our location-scale skew normal modelling under true parameter values $\tilde{\mu} = 2,\ \tilde{\sigma} = 0.5,\ \tilde{\alpha} = 1.5$ with sample size $n = 300$. The averages of these parameter estimates (marked with crosses in the boxes) are $\bar{\mu} = 2.003,\ \bar{\sigma} = 0.495,\ \bar{\alpha} = 1.485$. . . . . 98

4.4.2 Box plots for the simulation estimation result of skew normal data sampled from $SN(-2, 1, 0)$, based on sample size $n = 400$, re-sample and replicate 50 times. The left one did not use penalty ($\lambda = 0$). The right one used $\lambda\alpha^2$ penalty and each time $\lambda_{CV}$ is selected from a grid of candidates by 10−folded cross validation. 100

4.4.3 The box plots correspond to our $\lambda\alpha^2$ penalty result area in Table 4.1. There are three different settings for skew normal distribution skewness parameter $\alpha$ (they are $1, 3, 5$), and three different sample size $n$ (they are $50, 100, 200$) for each $\alpha$. . . 102

4.5.1 The scatter plot and histogram of $\hat{\alpha}$s for 187 anti-cancer drugs. The scatter plot shows the value of $\hat{\alpha}$ for each drug from 1st drug to 186th drug. For the convenience of the layout, the drug 'Sepantronium.bromide' is removed from the scatter plot, as it has an oversized $\hat{\alpha} = 27.8853$. Refer to **Appendix C** Figure 6.2.4 for details. . . . . 103

4.5.2 The box plots for bootstrapping of drug 'Pelitinib'. Resample $n = 397$ samples from $SN(0.00742,\ 2.56255,\ -1.64613)$ and perform estimations for 200 times to produce these box plots. . . . . 104

6.2.1 The estimation result for skew normal regression model under $n = 300,\ p = 20,\ \sigma = 0.5,\ \alpha = 1.5$. The averages of these estimated parameters after the 100 trials are $\overline{RE(\boldsymbol{\beta})} = 0.019,\ \bar{\sigma} = 0.482,\ \bar{\alpha} = 1.556,\ \overline{\text{SNR}} = 0.593$. RE means relative error. . . . . 190

6.2.2 The fixed sparse underlying real $\beta$ for penalised skew normal regression model simulation trials. . . . . 194

6.2.3 The penalised skew normal regression estimation results under $\alpha = 0.5, 1.0$and$1.5$ with underlying real parameters ($n = 300,\ p = 200,\ \sigma = 0.5$). RE means relative error. . . . . 194

6.2.4 The detailed estimation results of parameters $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ for 187 anti-cancer drugs after applied with $\alpha$-penalised skew normal model and corresponding EM algorithm computations. . . . . 195

# List of Tables


3.1 The summary table for the clustering and variable selection results using our penalised likelihood fusion. . . . . 41

3.2 The summary table about SEN, SPE and Relative Error for different penalty and initialisation settings. . . . . 61

3.3 BIC grid for $q = 1$ case on real data to show that $(\lambda_1 = 0.025, K = 5)$ has the minimum value 1220.526 which is regarded as the optimal. . . . . 62

3.4 BIC values show that $\left(\lambda_{0.5} = 0.005, \hat{K} = 5\right)$ has the minimum value 1202.692, it is an optimal point. Generally we just need to compute the $\hat{K} = 5$ column to locate the optimal point, which can save much time. . . . . 64

3.5 Drugs in each group and some group characters. (Part 1) . . . . . 65

3.6 Drugs in each group and some group characters. (Part 2) . . . . . 66

3.7 Drugs in each group and some group characters. (Part 3) . . . . . 67

3.8 Statistics table as complementary information for 50 sample trials related to Figure 3.3.1. . . . . 72

3.9 Statistics table as complementary information for 100 sample trials related to Figure 3.3.2. . . . . 72

3.10 Statistics for different missing percentage levels with respect to indicator matrix $\mathbf{Q}$. . . . . 73

4.1 The summary for skew normal models with our $\lambda\alpha^2$ penalty (left block) and $c_1 \log\left(1 + c_2\alpha^2\right)$ penalty (right block) respectively when they are estimated with our EM algorithm under distribution $SN(0, 1, \alpha)$ when true parameter $\tilde{\alpha} = 1, 3, 5$, based on sample size $n = 50, 100, 200$. Each entry is based on 50 replicative trials. . . . . 101

5.1 The Improvement Rates (the percentage where at least one of our four-piece distribution family's BICs was superior to *mclust*'s BIC) under 21 different settings where 4 possible mixture model types and different skewness degrees were simulated. Repeated each scenario 100 times by resampling and replicating the estimation process to produce each Im_R value. . . . . 115

5.2 Comparison of the estimation performances between our four-piece distribution family and *mclust*. RI is Rand Index. $\text{MAE}_{\boldsymbol{\mu}}$, $\text{MAE}_{\boldsymbol{\sigma}}$ and $\text{MAE}_{\boldsymbol{\alpha}}$ are the Mean Absolute Errors for the estimated parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$. The left-side four statistics (RI, $\text{MAE}_{\boldsymbol{\mu}}$, $\text{MAE}_{\boldsymbol{\sigma}}$, $\text{MAE}_{\boldsymbol{\alpha}}$) are the results from our four-piece distribution family. The right-side four dashed statistics (RI', $\text{MAE}'_{\boldsymbol{\mu}}$, $\text{MAE}'_{\boldsymbol{\sigma}}$, $\text{MAE}'_{\boldsymbol{\alpha}}$) are the results from R package *mclust*. Repeated each scenario 50 times by resampling and replicating the estimation process, then the average values were recorded in this table. . . . . 116

6.1 From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 1) . . . . . 126

6.2 From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 2) . . . . . 127

6.3 From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 3) . . . . . 128

# Chapter 1

# Introduction

This thesis is motivated by the development of anti-cancer drug studies. They are close to our human being life as a result of the wide use of targeted anti-cancer drugs to help the related treatment. First of all, we will introduce the concept of cancers in connection with human genes to explain the importance of personalised treatment. Then the anti-cancer drug dataset used in this thesis for our statistical analysis is introduced to provide a real-data application for our statistical algorithms and methodologies. After the description of existing challenges, our contributions are also mentioned. The organisation of this thesis is given at the end of this chapter to help readers have a glimpse of the whole thesis.

## 1.1 Cancer and Genes

Cancer comes from the changes or mutations in DNA which can alter essential cellular processes that govern cell behaviours, resulting in the invasion and move of cells to places they do not belong to (Elk and Landrine, 2012). These changes or mutations affect cell processes through gene expression which is the process of transferring coded information in a gene to a functional product such as a protein. In practice, these kinds of changes in genes are genetically unique, which indicates discrepancies for the same type of cancer among different individuals. These discrepancies can lead to different responses to one treatment for patients with the same type of cancer.

Almeida and Barry (2011) asserted that there is no single treatment for cancer. After that, the personalised treatment of cancer started to bloom in modern oncology. The personalised treatment uses the information about the patient and his tumour to figure out what is driving the disease, and then create a unique treatment plan with anti-cancer drugs for the patient. The most important task for the personalised treatment is to match up the right drugs with

the right patient (Yang et al., 2013).

The aim of fighting cancer more precisely drives the development of targeted therapy (targeted drug) which is a kind of drug designed to react with specific molecules to interfere with the progression of tumour cells. Targeted therapy has become the most effective way of treatment over the past years. Owing to the precision of targeted therapy, the amount of damage to normal cells is considerably minimised (Kidd et al., 2015).

Most of the time genes function in groups instead of alone. This is called genetic interactions, where two or more genes show a combined effect which does not manifest itself through only one of them alone (Bebek, 2012). Thus, the gene expression profile becomes a helpful tool for discovering the functional cooperativity between different genes (Garnett et al., 2012). And the research on gene interaction networks becomes necessary to understand the pathway and regulation behind the working principle of genes (Bebek, 2012), then an insight into cancer can be provided to power the research of disease-specific cure and personalised treatment. This thesis focuses on genes and anti-cancer drugs to reveal statistical features and hidden connections between them.

## 1.2 Cancer Drug Data

Doctors and scientists are always exploring better ways to treat patients with cancer. One way to do this is by creating new drugs. They also try to find new ways of using drugs which are already available. Before any new drug becomes available to be prescribed to patients, there is a long process taking many years and lots of resources. In general, there are 3 steps of developing a new anti-cancer drug: preclinical research, clinical research, and post-clinical research. Our research can contribute to the first step 'preclinical research' where the drug is discovered and designed.

The data of our research comes from Iorio et al. (2016), who integrated heterogeneous molecular data of 11289 tumours and 1001 cell lines, and they also measured the responses of these 1001 cancer cell lines to 265 anti-cancer drugs. Thus, our data consist of two parts. The first one is the sensitivity of different cell lines (COSMIC_ID) to various anti-cancer drugs, which is quantified by the logarithm of the half maximal inhibitory concentration (Ln_IC50). IC50 is a quantitative measure of how much (concentration) one drug is needed to inhibit the multiplication of that cell line by 50% (Schwab, 2008). Generally, IC50 is estimated by two steps. Firstly, measure the inhibition activity of one cell line against a range of dilutions of the drug (as well as without the drug). Secondly, perform curve-fitting (a sigmoid curve) of these inhibition results against concentrations, then IC50 is calculated as half of the upper asymptote

value. The other data part is the gene expression data for the correspondingly aforementioned cell lines. In this article (Iorio et al., 2016), the authors take advantage of these data by using ANOVA to identify gene-drug associations, where many relevant associations have been known clinically. While some associations do not contain known cancer genes, this may give clues to the likely driver cancer genes through the patterns of drug sensitivity. Following this research, we used other methods to analyse the associations between genes and drugs to hope for more insights into the gene-cancer research to power the development of anti-cancer drugs.

## 1.3 Challenges and Contributions

When multiple genes are connected with different drugs, the modelling for drugs and genes becomes a multivariate multiple linear regression problem. When using the gene expression data, high dimensionality imposes a hurdle against the modelling of these data as a result of the large number of genes in human being. Each gene will add one more dimension to our analysis, and there are thousands of genes. However, the number of cell lines is much fewer than that quantity. To overcome the problem of overfitting, some constraints are needed to filter parameters. Pan and Shen (2007) used $l_1-$norm penalty to help multivariate clustering but not under a regression basis. Different from existing models, we build a penalised multivariate multiple regression model and then simultaneously solved variable selection (for genes) and clustering (for drugs) problems. After that, missingness is considered in our modelling to include more data into our analysis and avoid any improper deletion, which is very close to real-world data situations since missingness exists everywhere in research applications.

When building these models, the group structure for anti-cancer drugs is always accommodated because drug combination therapy has been proved more effective than monotherapy when killing tumour cells (Janku et al., 2014; Musgrove et al., 2011), where these drugs fulfil their functions by binding to one or more protein targets. Moreover, drug resistance is largely reduced when drugs are used simultaneously on patients (Carter et al., 2016).

In the field of anti-cancer drug statistical researches, when building models, skewness has rarely been considered. While the inherent skew property of the drug data could distort the shape of normal modelling. The introduction of the skew normal distribution could better depict our skewed data, but the inference for this model is not easy. To reach a more accurate estimation result, starting from the location-scale skew normal model, we introduced a different penalised skew normal model with a new way of EM algorithm to estimate it in Chapter 4. Using this skew normal estimation, we revealed the inherent skewness within the given drug-and-cell-line data for some drugs. In the next asymptotic analysis part, the asymptotic distributions

for the MLE of skew normal distribution under the regular condition (skewness is not zero) (Azzalini, 2014) are calculated to the end, which has not been found in the literature to our best knowledge. For the skewed data with a group structure, we also proposed a new way of estimating the skew normal mixture models, which is different from existing methods within the literature.

## 1.4 Organisation of The Thesis

There are six chapters in total except for four more chapters in the appendix. The following chapters form the rest of this thesis's main body.

In Chapter 2, we briefly introduce some concepts used in this thesis as a preliminary for the next analysis, and we also review some backgrounds for these concepts.

In Chapter 3, a new penalised likelihood fusion model is introduced. To estimate the parameters in this model, the EM algorithm is laid out step by step in detail. Then we extend this model by adapting it for missing entries, and the corresponding EM algorithm is also shown here. The cancer-drug and gene-expression data analysis using this model is also carried out, and a group-related gene map is plotted with drug clustering tables.

In Chapter 4, we introduce a skew normal location-scale model with a different EM algorithm from the conventional ones in the skew normal estimation. A new penalised likelihood is also put forward to address the odd behaviour of estimation when true skewness is near 0. Then asymptotic distributions for MLE are computed afterwards following some newfound asymptotic properties.

In Chapter 5, we introduce a new modelling called 'the four-piece distribution family' for skew normal mixtures, it provides a new view of skew normal mixture problems. The following simulation studies demonstrate the validity and good performances of our method for skew normal mixture modelling.

Finally, Chapter 6 contains the conclusions and future works of this thesis. Not only does it show the constraints or limits of our modellings, but also suggests some potential extensions for our modellings and algorithms in the future. The technical details are postponed to the Appendix.

# Chapter 2

# Methodological Background

## 2.1 Curse of Dimensionality

Curse of Dimensionality refers to various phenomena arising as a result of dimension increasing. The impact can be multiple and includes different ways. Giraud (2021) mentioned four aspects: firstly, data points are isolated in high-dimensional immensity; secondly, the accumulation of small fluctuations results in a large global fluctuation; thirdly, rare events accumulate to non-rarity; fourthly, numerical computations and optimisations become excessively intensive in high-dimensional spaces.

## 2.2 LASSO and Extensions

In high-dimensional problems, we need variable selection methods to reduce the model complexity effectively but at the same time balance model bias and variance. Variable selection is supported by the idea that only a small fraction of predictors are informative and significantly connected with the response variables (Bühlmann and Van De Geer, 2011). Least Absolute Selection and Shrinkage Operator (LASSO) and its variants were developed for variable selection. For a high-dimensional dataset, we consider the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where $\mathbf{x}_i$ is a $p-$dimensional covariate independent of each other coming from $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T$, and $y_i$ is a univariate response variable from $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. To perform regression, we assume a linear model $y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \, (i = 1, \ldots, n)$ where $x_{ij}$ is the explanatory variable from $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $\beta_j$ is the coefficient from vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$, and error term $\varepsilon_i$ is from $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$, $\beta_0$ is the intercept. When $p < n$, this regression problem is solved by minimising the objective function $\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$. When $p > n$, generally there is $\boldsymbol{\beta}$ to make the objective function 0, but $\boldsymbol{\beta}$ is not unique, and the data are also over-fitted as a

result of parameters outnumbering observations.

### 2.2.1 Original LASSO

LASSO regression (Tibshirani,1996) is an improvement to the ordinary least square (OLS), it adds a $l_1-$norm constraint onto the coefficients:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Generally the intercept $\beta_0$ can be omitted by standardising the dataset:

$$\frac{1}{n}\sum_i y_i = 0, \qquad \frac{1}{n}\sum_i x_{ij} = 0, \qquad \frac{1}{n}\sum_i x_{ij}^2 = 1.$$

For the convenience of notations, we can also put it in the Lagrangian form:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n + \lambda \|\boldsymbol{\beta}\|_1 \right),$$

where $\lambda \geq 0$ is a tuning parameter to control the severity of the penalty term $\lambda \|\boldsymbol{\beta}\|_1$. Through this, the shrinkage degree of the estimated coefficients can be controlled. And there is a one-to-one correspondence between $\lambda$ and $t$ depending on given data.

The solution for LASSO regression can be seen as the classic linear least-squares method with a bound constraint $\sum_{j=1}^{p} |\beta_j| \leq t$, which can help to yield a sparse solution and also reach coefficient shrinkage. In Figure 2.2.1, each curve shows the gradual change of each coefficient from zero to non-zero with the increase of the constraint $t$ on the $l_1-$norm $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$. The bound $t$ varies from 0 on the leftmost side, all the way to a large value 14 on the rightmost side in this coefficient-path plot. The constraint $t$ has no effect anymore when it becomes large enough. More coefficients are released to non-zero as the severity of the penalty becomes looser with the extension of the horizontal axis.

LASSO becomes very useful when the data is provided with $p \gg n$ feature, but actually, $k\,(< n)$ variables have non-zero coefficients in the underlying true model. When $p > n$, with the ordinary least square (OLS) objective function $\min \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$, the estimates which set the objective function as zero are not unique. From this property, it also suggests that the model is overfitted. Moreover, all of the coefficients are usually estimated as non-zero, which cannot realise variable selection purpose. To explore the variable selection property of LASSO, we first introduce ridge regression as a comparison.

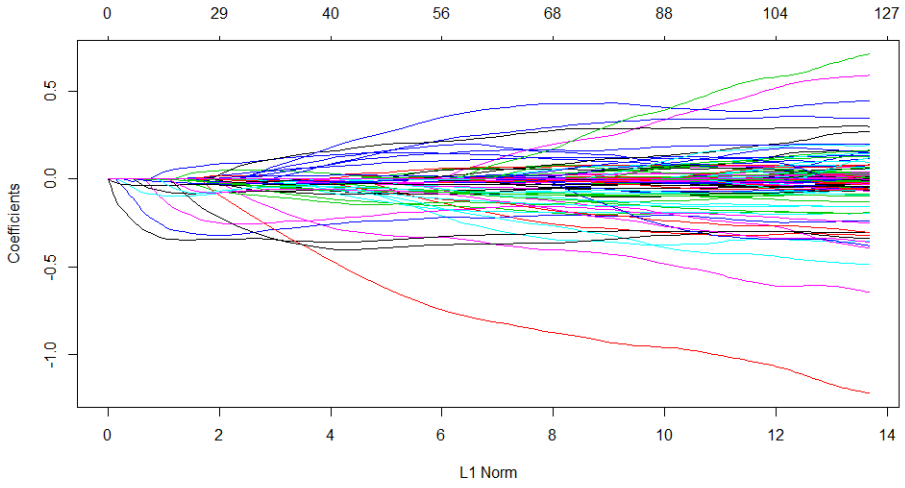



Figure 2.2.1: The coefficient-path plot against the $l_1-$norm of the coefficient vector. Y axis indicates the coefficient value of the variable, and X axis is the value of bound $t$.

### 2.2.2 Ridge Regression

Ridge regression (Hoerl et al., 1970) is the OLS method with a constraint on the $l_2-$norm of coefficients:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

Similarly, omit intercept $\beta_0$ by standardisation and put it in the Lagrangian form:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n + \lambda \|\boldsymbol{\beta}\|_2^2 \right),$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$, and $\lambda \geq 0$ is a tuning parameter with a one-to-one correspondence to $t$. However, ridge regression can not yield sparse solutions because the contour (red ellipse curves on the graph) can not touch the $l_2-$norm constraint (blue circle on the right graph) with the coefficient of one dimension being zero. It is illustrated in Figure 2.2.2. As is shown, the RSS contour in the LASSO case reaches the minimum value by touching the square's corner where $\beta_1 = 0$ as marked on the left graph. That is why LASSO has the ability of assigning zero values to some coefficients to realise variable selection. Thus, the sparsity is achieved by using LASSO regression, where sparsity means the property of a model with few non-zero coefficients. This key property comes from the geometry of $l_1-$norm, and it can be applied in other statistical models.

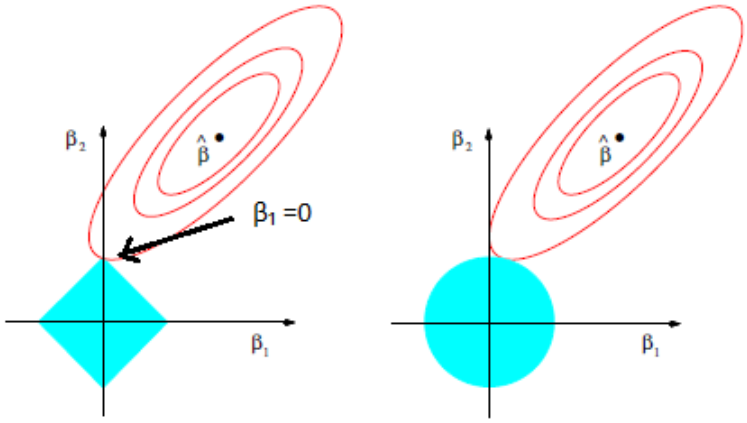



Figure 2.2.2: The contours of residual sum of squares $\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ are plotted with red ellipses. $\hat{\boldsymbol{\beta}}$ is the OLS estimate $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$. The LASSO case is on the left with the constraint area $|\beta_1|+|\beta_2| \leq t$. The ridge case is on the right with the constraint area $\beta_1^2+\beta_2^2 \leq t^2$. (Tibshirani, 1996; James et al., 2013)

### 2.2.3 Elastic Net

If we turn to variations of the basic LASSO $l_1-$norm penalty itself, we can find that the $l_1-$norm penalty makes it possible for the shrinkage and selection of variables because of the existence of the non-differentiable property from the $l_1-$norm. But with this character, we may set some coefficients as zero, which is sometimes overly absolute without any prior knowledge. To keep variable selection and shrinkage, but compromise with the accuracy from the ridge regression, the Elastic Net can be considered, which makes the coefficient estimates more accurate but less sparse.

Elastic Net (Zou and Hastie, 2005) is a compromise between LASSO and ridge penalties by solving the following convex problem:

$$\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2) = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 /n + \lambda_1 \|\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_1 \right),$$

where $\lambda_1, \lambda_2 \geq 0$ are regularisation parameters, and the problem reduces to LASSO or ridge regression when $\lambda_1$ or $\lambda_2$ is set as 0 respectively. This problem can also be expressed as

$$\hat{\boldsymbol{\beta}}(\alpha,\lambda) = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 /n + \lambda \left[ (1-\alpha) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right] \right)$$

which can show the compromise between LASSO and ridge, where $\alpha$ is a proportion parameter. But there are two tuning parameters $\alpha, \lambda$ to be determined. Here $\alpha$ can be seen as a higher-level parameter which can be set by users. Alternatively, we use a grid of values to select $\alpha$ and $\lambda$ from cross-validation. In the field of genomics, we often find that correlated features work in

groups, where some genes express together with the same biological pathway. So their data have strong correlations with each other. But LASSO does not perform well with variables which are highly correlated with each other. By using squared $l_2-$norm penalty and $l_1-$norm penalty together, it can work better with such correlated features and tends to screen out the correlated variables together.

The drawback of the Elastic Net is always obvious. It is computationally more expensive than LASSO or ridge because the relative weight between LASSO and ridge has to be selected using cross-validation. But in the absence of prior knowledge, the Elastic Net is generally preferred compared to pure LASSO or ridge regression. The prior knowledge acquired before the regression includes the examples such as

- All covariates (regressors) are relevant and highly correlated. Then variable selection is not needed as we want to keep all information in the result. Under this circumstance, we should turn to the ridge regression to avoid unnecessary variable selection which could remove useful information hidden in some variables.

- Some regressors are completely irrelevant. We need to do variable selection using LASSO to leave them out in the next step of research. This is what we did in the preliminary analysis part of next chapter. We have known that most of the genes are irrelevant to the development of one certain kind of cancer, so we may exclude them from the connection to one certain anti-cancer drug. A strong variable selection is required here, thus, LASSO regression is treated as the priority over other regression methods.

To uncover the screening power of the Elastic Net when different $\alpha$s are applied to the model, we performed variable selection using the *glmnet* package on 227 anti-cancer drugs respectively. The number of variables (genes) that remains after the regression for each drug is shown in Figure 2.2.3. Fewer variables are screened out with more weight put on $l_1-$norm penalty until it becomes a pure LASSO regression where $\alpha$ is set as 1.

### 2.2.4 Adaptive LASSO

Although LASSO is a powerful screening method to produce sparse outcomes, it can still come out with many noise covariates (the ones screened out but close to 0). So we need optimisation methods which can reduce the number of noise covariates to produce sparser results.

To optimise the inadequate screening behaviour of LASSO (noise covariates are selected besides relevant variables), Adaptive LASSO (Zou, 2006) is introduced here which replaces the

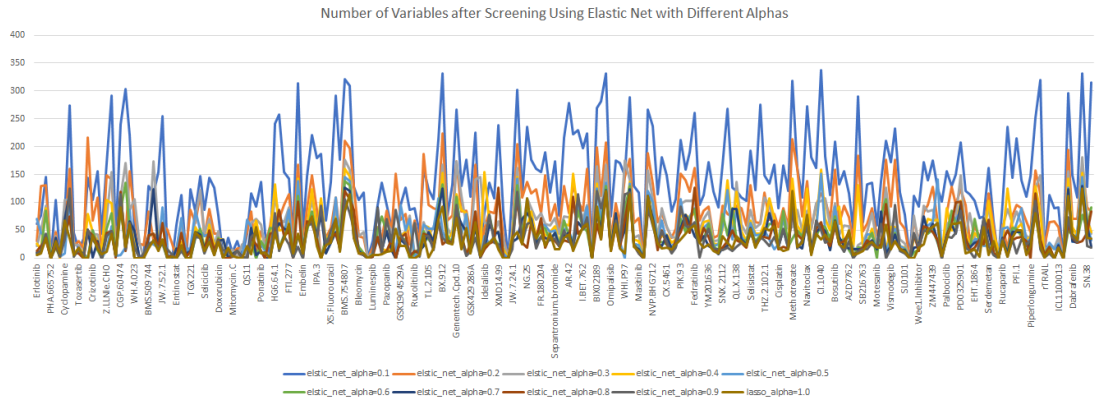



Figure 2.2.3: The number of remaining genes for different drugs after applying the Elastic Net. We illustrate the variable selection effects (by counting how many genes are screened out for each drug) under different $\alpha$s. Y axis indicates the number of non-zero-coefficient variables (genes) after the Elastic Net screening.

$l_1-$norm penalty with a reweighted version:

$$\hat{\boldsymbol{\beta}}_{\text{adapt}}(\lambda) = \arg\min_{\beta} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\left|\hat{\beta}_{\text{init},j}\right|} \right), \tag{2.2.1}$$

where $\hat{\boldsymbol{\beta}}_{\text{init}}$ is an initial estimator from LASSO:

$$\hat{\boldsymbol{\beta}}_{\text{init}}(\lambda') = \arg\min_{\boldsymbol{\beta}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 / n + \lambda' \|\boldsymbol{\beta}\|_1 \right). \tag{2.2.2}$$

Use cross-validation twice to select parameter $\lambda'$ and $\lambda$ . Firstly, we choose it for the initial parameter $\hat{\boldsymbol{\beta}}_{\text{init}} = \hat{\boldsymbol{\beta}}_{\text{init}}(\lambda')$ from (2.2.2). Secondly, we use cross-validation again to select $\lambda$ for Adaptive LASSO (2.2.1).

The inadequate screening behaviour of LASSO can be explained theoretically that

$$\mathbf{P}\left[\hat{S}(\lambda) \supseteq S_0\right] \to 1 \, (p \geq n \to \infty),$$

$$\lim_{n\to\infty} \sup \mathbf{P}\left[\hat{S}(\lambda) = S_0\right] < 1 \, (p \geq n \to \infty),$$

where $S_0$ is the underlying true set of active variables expressed as $S_0 = \left\{j; \beta_j^0 \neq 0, j = 1, \ldots, p\right\}$. So noise covariates cannot be avoided with the choice of $\lambda$, we can just reduce the number of them. And Adaptive LASSO plays this role. Actually, selecting the regularisation parameters $\lambda$ and $\lambda'$ in a sequential way is computationally cheaper than simultaneous optimisation over two tuning parameters.

Now we illustrate that the Adaptive LASSO can yield a much sparser result than the original

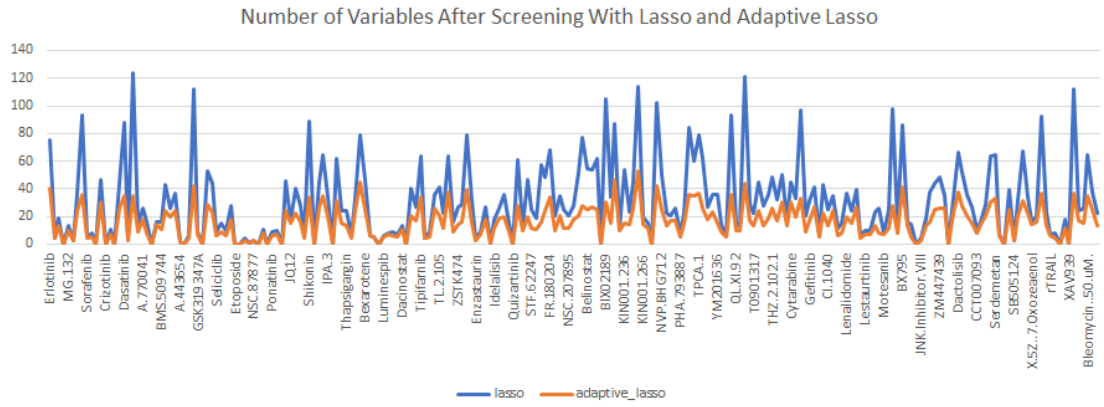



Figure 2.2.4: The number of genes for different anti-cancer drugs after screenings with the original LASSO (blue line) and Adaptive LASSO (orange line). Y axis indicates the number of non-zero variables after the screening.

LASSO. One property of Adaptive LASSO can explain it (Bühlmann and Van De Geer, 2011):

$$\hat{\beta}_{\text{init},j} = 0 \Longrightarrow \hat{\beta}_{\text{adapt},j} = 0.$$

If $\left|\hat{\beta}_{\text{init},j}\right|$ is large, Adaptive LASSO employs a smaller penalty for the $j$th coefficient $\beta_j$ where less bias is implied. Then when $\left|\hat{\beta}_{\text{init},j}\right|$ is small, a heavy penalty could be put on this coefficient $\beta_j$, which may result in a zero coefficient this time. Thus, Adaptive LASSO yields a sparser solution than the original LASSO. It can reduce the number of noise covariates (which are left over after screening but irrelevant) from the first step 'original LASSO'.

We conducted a comparison between the original LASSO and Adaptive LASSO on the real data to show this property. With the help of the *glmnet* package in R, the comparison for the number of active variables left after the screening processes is plotted in Figure 2.2.4. The second step of Adaptive LASSO used the estimates from the original LASSO as the weight to refit. As shown in this graph, the orange line (Adaptive LASSO) is beneath the blue line (the original LASSO) everywhere. Therefore, it can be affirmed that Adaptive LASSO yields sparser solutions than the original LASSO.

For the computation of Adaptive LASSO, we can use rescaling technique on covariates after which it becomes a LASSO problem (Bühlmann and Van De Geer, 2011),

$$\tilde{\mathbf{X}}^{(j)} = \left|\hat{\beta}_{\text{init},j}\right| \mathbf{X}^{(j)}, \qquad \tilde{\beta}_j = \frac{\beta_j}{\left|\hat{\beta}_{\text{init},j}\right|},$$

then the Adaptive LASSO problem of (2.2.1) becomes

$$\min_{\beta} \left( \left\| \mathbf{y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} \right\|_2^2 / n + \lambda \left\| \tilde{\boldsymbol{\beta}} \right\|_1 \right).$$

Here we omit the variables $j$ with $\hat{\beta}_{\text{init},j} = 0$. So the solution of Adaptive LASSO is

$$\hat{\beta}_{\text{adapt},j} = \left| \hat{\beta}_{\text{init},j} \right| \hat{\tilde{\beta}}_j.$$

This Adaptive LASSO can be upgraded to multi-step adaptive LASSO (Bühlmann and Meier, 2008) where every iteration step uses a separate tuning parameter and increases the sparsity in every step. The same as basic Adaptive LASSO, it also benefits from selecting tuning parameters sequentially instead of simultaneously to alleviate the computational complexity.

### 2.2.5 LASSO-OLS Hybrid, Thresholded LASSO and Relaxed LASSO

For the LASSO-OLS hybrid (Guo et al., 2015), it is an ordinary least square estimator on the selected variables following the original LASSO:

$$\hat{\boldsymbol{\beta}}_{\text{LASSO-OLS}} = \left( \mathbf{X}_{\hat{S}}^T \mathbf{X}_{\hat{S}} \right)^{-1} \mathbf{X}_{\hat{S}}^T \mathbf{y},$$

where $\hat{S} = \left\{ j; \hat{\boldsymbol{\beta}}_{\text{LASSO},j} \neq 0 \right\}$ is the set of non-zero coefficients selected from the first step LASSO. And $\mathbf{X}_{\hat{S}}$ is the restriction of $\mathbf{X}$ to columns in $\hat{S}$.

The thresholded LASSO (Zhou, 2010) is another way to obtain a sparser parameter estimate than the original LASSO. At first, the coefficients from the original LASSO are thresholded through the following procedure:

$$\hat{\beta}_{\text{thres},j}(\lambda_{\text{init}}, \delta) = \hat{\beta}_{\text{init},j} \times \mathbb{I}(\left| \hat{\beta}_{\text{init},j} \right| > \delta),$$

where $\hat{\beta}_{\text{init},j} = \hat{\beta}_{\text{init},j}(\lambda_{\text{init}})$, and $\mathbb{I}(\cdot)$ is the indicator function. Then the selected variables are refitted with ordinary least squares like the LASSO-OLS hybrid:

$$\hat{\boldsymbol{\beta}}_{\text{thres-refit}} = \left( \mathbf{X}_{\hat{S}_{\text{thres}}}^T \mathbf{X}_{\hat{S}_{\text{thres}}} \right)^{-1} \mathbf{X}_{\hat{S}_{\text{thres}}}^T \mathbf{y},$$

where $\hat{S}_{\text{thres}} = \left\{ j; \hat{\boldsymbol{\beta}}_{\text{thres},j} \neq 0 \right\}$.

The additional thresholding step leads to a better performance than the LASSO-OLS hybrid. This thresholded LASSO method has inspired the introduction of our two-step initialisation method for the mixture skew normal regression problem (attached in **Appendix D**).

The relaxed LASSO (Meinshausen, 2007) is one more step with a smaller penalty after the

original LASSO in the first step. So in the second step, the estimator is

$$\hat{\boldsymbol{\beta}}_{\hat{S}}(\lambda,\phi) \quad = \quad \arg\min\left\{\left\|\mathbf{y}-\mathbf{X}_{\hat{S}}\boldsymbol{\beta}_{\hat{S}}\right\|_2^2/n+\phi\cdot\lambda\left\|\boldsymbol{\beta}_{\hat{S}}\right\|_1\right\},\ (0\leq\phi\leq 1),$$

where $\hat{S}=\left\{j;\hat{\boldsymbol{\beta}}_{\mathrm{LASSO},j}\neq 0\right\}$. When $\phi=0$, it becomes the LASSO-OLS hybrid. In this method, $\lambda$ and $\phi$ are selected by cross-validation simultaneously unlike Adaptive LASSO. Moreover, the relaxed LASSO and Adaptive LASSO have similar performances in practice. Owing to the interest of my research, the detailed performances of them are not discussed here.

### 2.2.6 Package *glmnet*

*glmnet* (Friedman et al, 2015) is an R package to fit one generalised linear model through penalised maximum likelihood. With the different settings of penalty terms, it can solve LASSO or Elastic Net problem when facing different shapes of data. The options and features in the arguments are introduced below:

- Family: the loss function and the associated model, including Gaussian, Poisson, binomial, multinomial and cox. In our application, the residual sum of squares for Gaussian errors is a kind of deviance measure.
- Penalties: *glmnet* will generate a sequence of 100 values for $\lambda$ to test. $\lambda$ can control how heavy the penalty is. And $\alpha$ in Elastic Net can be chosen by cross-validation with a grid of $\alpha$ and $\lambda$, where $\alpha\in[0,1]$ is the mixture proportion between ridge ($\alpha=0$) and LASSO ($\alpha=1$).
- Coefficient bounds: to accelerate the calculation, there are upper and lower bounds during the coordinate-descent loop.
- Offset: a real-valued number for each observation, which is added to the linear predictor and is not associated with any parameters.
- Matrix input and weights: when response input is a matrix instead of a vector, which indicates the grouped data. Each row of response matrix is filled with weights or weight portions.
- Sparse model: when $p\gg n$, store efficiently by only storing non-zero coefficients. Coordinate descent takes advantage of that sparsity because of one variable one time and inner-product operation.
- Cross-validation: K-fold cross-validation can be performed with a function inside to find the 'best' tuning parameter $\lambda$.

## 2.3 Model-based Clustering and Mixture Model

Clustering or cluster analysis is the identification task to group similar observations into the same group in contrast to the differences from the elements in other groups. Most of the time, it is done through heuristic steps until Fraley and Raftery (2002) composed systematic guidance on cluster analysis to solve the problems such as the number of clusters needed, suitable methods and treating outliers, see also Raftery and Dean (2006). Model-based clustering assumes that the given data are generated through an existing but unknown model, and it tries to recover the true model from the given data. After recovering, the estimated true model defines the clusters and their parametric features. In our research, we use maximum likelihood as the criterion to estimate the true model. We try to find out the parameter set $\mathbf{\Phi}$ to maximise the log-likelihood of generating the given data $\mathbf{Y}$:

$$\hat{\mathbf{\Phi}} = \arg\max_{\mathbf{\Phi}} L(\mathbf{\Phi}|\mathbf{Y}) = \arg\max_{\mathbf{\Phi}} \log \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{\Phi}) = \arg\max_{\mathbf{\Phi}} \sum_{i=1}^{n} \log p(\mathbf{y}_i|\mathbf{\Phi}),$$

where $L(\mathbf{\Phi}|\mathbf{Y})$ is treated as the objective function. When clustering with the same known number of groups $K$, higher $L(\mathbf{\Phi}|\mathbf{Y})$ is preferred.

Different clustering methods can yield different assignments, and the form of clustering could also be very different. For example, fuzzy c-means (Dunn, 1973) produces soft clustering, which means each data point can belong to more than one group using a membership grade indicating the degree of belonging to each group. k-means (MacQueen, 1967) is the other case as it assigns each data point to only one specific group. So k-means produces hard clustering. Mixture modelling is another way to handle clustering problems which may include the EM algorithm (Dempster et al., 1977) to help the computations of interlocking equations, then a soft clustering result is reached owing to the nature of EM algorithm E-step (refer to Section 2.4).

For the mixture model, the first major analysis where it was used can date back over 120 years. Pearson (1894) fitted a mixture model with two normal density functions which have different means, variances and proportions to the data from Weldon (1892, 1893). He used the measurements from 1000 crabs from the Bay of Naples on the ratio of the forehead to body length, from which he found that the asymmetry in the histogram of these data may indicate two new subspecies present in that population. Pearson (1906) is the first research advocating statistical analysis as the primary method to solve biological problems. Contemporarily, McLachlan and Peel (2000) gave a comprehensive review of finite mixture models.

The likelihood function of the mixture model with $K$ groups for data $\mathbf{y}$ is shown here:

$$L(\theta_1, \ldots, \theta_K; \pi_1, \ldots, \pi_K | \mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \theta_k),$$

where $f_k(\cdot)$ is the density function, $\theta_k$ is the parameters in the $k$-th group, and $\pi_k$ is the mixture proportion as a prior probability of $k$-th group which must add up to 1 ($\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$).

When $f_k(\cdot)$ is a multivariate normal density, we have $\mathbf{y}_i \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and

$$f_k(\mathbf{y}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left[-1/2\,(\mathbf{y}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_k)\right]}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}}.$$

The covariance matrix $\boldsymbol{\Sigma}_k$ can determine the geometry of clusters. When $\boldsymbol{\Sigma}_k = \lambda I$, all clusters are sphere-shaped structures of the same size. When $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ (a constant matrix which is a matrix with all its entries equal to one given number), then all clusters have the same geometry which could be sphere-shaped or not. When $\boldsymbol{\Sigma}_k$ is not restricted to one matrix, each cluster can have its own geometry (shape, orientation and volume). Eigenvalue decomposition for the covariance matrix of multivariate normal mixtures (Banfield et al., 1993) can build a general framework for the geometry of clusters:

$$\boldsymbol{\Sigma}_k = \lambda_k O_k D_k O_k^T,$$

where $\lambda_k$ is a constant to control the volume of $k$-th cluster, $O_k$ is the orthogonal matrix of eigenvectors to control the orientation of $k$-th cluster, and $D_k$ controlling the shape is a diagonal matrix with elements proportional to the eigenvalues. From this decomposition, we can deduce that when $\boldsymbol{\Sigma}_k = \lambda_k I$, all clusters are sphere-shaped but with different volumes.

A powerful method called EM algorithm can help us with the mixture model estimation as we can not solve it directly by derivation. The presence of the sum in logarithm $\log L = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \theta_k)\right)$ will lead to a very complicated solution for the maximum likelihood. From next section, we will move to the introduction of EM algorithm.

In our research, for the purpose of comparison, we used the package *mclust* in R to do clustering, where *mclust* (Scrucca et al., 2016) is an R package to perform model-based clustering, classification and density estimation for finite normal mixture models through parameter estimation from EM algorithm. It can cope with different covariance structures and select the optimal model according to BIC values. The eigen-decomposition concerning the within-group covariance matrix $\boldsymbol{\Sigma}_k$ for multidimensional data is also available through the function inside to reveal the geometric features (volume, shape and orientation) of clusters.

### 2.3.1 Mixture Regression Model

When one random variable with a finite mixture distribution depends on some covariates, a finite mixture regression model is obtained (Khalili and Chen, 2007). Now with given data pairs $(\mathbf{x}_i, y_i)$ from variables $(\mathbf{X}, \mathbf{y})$, we can perform a mixture model regression under the assumption that $\mathbf{y}$ follows a mixture distribution. Then the response variable $\mathbf{y}$ and predictor variable $\mathbf{X}$ can be connected by a finite mixture regression. Each data pair $(\mathbf{x}_i, y_i)$ belongs to one of $K$ clusters. Given the information that $(\mathbf{x}_i, y_i)$ comes from cluster $k$ ($k \in [1, 2, \dots K]$), then we have

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_k + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_k^2)$, and the conditional distribution for $\mathbf{y}$ given $\mathbf{X}$ is the mixture model with following density function:

$$f_{\mathbf{y}|\mathbf{X}}(y_i|\mathbf{x}_i) \quad = \quad \sum_{k=1}^{K} \pi_k f_k(y_i|\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2),$$

where $f_k(\cdot|\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)$ can be regarded as the density of normal distribution with mean $\mathbf{x}_i^T \boldsymbol{\beta}_k$ and variance $\sigma_k^2$, or say $y_i \sim N(\mathbf{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)$. And $\pi_k$ is the mixing proportion with properties $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \in [0, 1]$.

When the dimension of covariate $\mathbf{x}_i$ is large, sometimes even larger than the number of observations $n$, it comes to the problem of variable selection in the finite mixture regression model owing to the common practice of including only the important covariates in the model (Khalili and Chen, 2007). Akaike Information Criterion (Akaike, 1973) and Bayes information criterion (Schwarz, 1978) have been used in mixture regression model selections such as the application in finite mixture of Poisson regression models (Wang et al., 1996). But they become more and more computationally intensive with the increase of dimensions. LASSO and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001, 2002) are the later methods suitable for variable selection in high-dimensional problems. See also Khalili and Chen (2007) who introduced a penalised function to automatically select the sub-model and also performed parameter estimations.

## 2.4 Expectation–Maximisation Algorithm

### 2.4.1 Introduction

Maximum likelihood estimation has been the most commonly used approach to estimating the mixture distribution by far since the advent of the EM algorithm (McLachlan et al., 2019).

Expectation-maximisation (EM) algorithm (Dempster et al., 1977) is an iterative method to find the local maximum likelihood estimates of the parameters in a parametric probability distribution when the model depends on unobservable latent variables. It alternates between the E-step and M-step until convergence is reached. In a mixture model, data come from $K$ components, but the group label $z$ of each is not observable, because we don't know which component each observation belongs to. In this circumstance, we turn to the EM algorithm to find the maximum likelihood estimates for these latent variables. The E-step tries to calculate the expectation of the log-likelihood function with the current estimates of parameters and observed data. The M-step tries to maximise the expectation of the log-likelihood function with respect to all parameters needed to estimate.

### 2.4.2 E-step

First, we introduce the **incomplete-data** likelihood function which is the target function we want to maximise:

$$L_{inc}(\mathbf{\Phi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k), \tag{2.4.1}$$

where $\mathbf{\Phi}$ is the set of all parameters and $\boldsymbol{\theta}_k$ is the set of the parameters in $k$th component. But the incomplete-data likelihood function is hard to be maximised. We use another way to maximise it indirectly.

With the latent (hidden) information $z$, (2.4.1) becomes a **complete-data** likelihood function:

$$L_{com}(\mathbf{\Phi}) = \prod_{i=1}^{n} \prod_{k=1}^{K} [\pi_k f_k(\mathbf{y}_i; \boldsymbol{\theta}_k)]^{z_{ik}},$$

where $z_{ik} = 1$ when observation $i$ belongs to group $k$, otherwise $z_{ik} = 0$. It works as an indicator.

Because $z_{ik}$ is unobservable from the data, we use the Bayes theorem to estimate it with the current information (the $v$-th parameter estimates) and the observed data

$$\begin{aligned} \tau_{ik}^{(v)} &= \Pr(z_{ik} = 1 | \mathbf{y}_i, \boldsymbol{\theta}^{(v)}) \\ &= \mathrm{E}(z_{ik} | \mathbf{y}_i, \boldsymbol{\theta}^{(v)}) \\ &= \frac{\pi_k^{(v)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_k^{(v)})}{\sum_{k=1}^{K} \pi_k^{(v)} f_k(\mathbf{y}_i; \boldsymbol{\theta}_k^{(v)})}. \end{aligned}$$

Sometimes $\tau_{ik}^{(v)}$ is called *'membership probability'*. Then the conditional expectation of the **complete-data log-likelihood** function is:

$$
\begin{aligned}
Q(\mathbf{\Phi}|\mathbf{\Phi}^{(v)}) &= E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}^{(v)}}\left[\log\left(L_{com}(\mathbf{\Phi}^{(v)})\right)\right] \\
&= E_{\mathbf{Z}|\mathbf{Y},\boldsymbol{\theta}^{(v)}}\left(\sum_{i=1}^{n}\sum_{k=1}^{K}\left[z_{ik}\log\pi_k + z_{ik}\log f_k(\mathbf{y}_i;\boldsymbol{\theta}_k)\right]\right) \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\mathrm{E}(z_{ik}|\mathbf{y}_i,\boldsymbol{\theta}^{(v)})\times\log\pi_k + \mathrm{E}(z_{ik}|\mathbf{y}_i,\boldsymbol{\theta}^{(v)})\times\log f_k(\mathbf{y}_i;\boldsymbol{\theta}_k)\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\tau_{ik}^{(v)}\log\pi_k + \tau_{ik}^{(v)}\log f_k(\mathbf{y}_i;\boldsymbol{\theta}_k)\right]. \qquad (2.4.2)
\end{aligned}
$$

This is the expectation function which needs to be iteratively maximised and updated in the next **M-step**.

### 2.4.3 M-step

(2.4.2) is maximised with respect to $\mathbf{\Phi}$(all parameters), then the $(v+1)$-th parameter estimate set $\mathbf{\Phi}^{(v+1)}$ is obtained:

$$\mathbf{\Phi}^{(v+1)} = \arg\max_{\mathbf{\Phi}} Q(\mathbf{\Phi}|\mathbf{\Phi}^{(v)}).$$

Thus, for every M-step we have

$$Q(\mathbf{\Phi}^{(v+1)}|\mathbf{\Phi}^{(v)}) \geq Q(\mathbf{\Phi}|\mathbf{\Phi}^{(v)}). \qquad (2.4.3)$$

Then update the parameters in (2.4.2), and carry on the iteration alternating between E-step and M-step until our preset stopping rule is achieved:

$$\left|\log L_{inc}(\mathbf{\Phi}^{(v+1)}) - \log L_{inc}(\mathbf{\Phi}^{(v)})\right| < \epsilon,$$

where $\epsilon$ is a preset tolerance such as $10^{-5}$ or $10^{-10}$.

### 2.4.4 Monotonicity

When the iteration goes on, the $(v+1)$-th incomplete likelihood $L_{inc}(\mathbf{\Phi}^{(v+1)})$ will never be less than the $v$-th one $L_{inc}(\mathbf{\Phi}^{(v)})$, that is

$$L_{inc}(\mathbf{\Phi}^{(v+1)}) \geq L_{inc}(\mathbf{\Phi}^{(v)}).$$

It is called **monotonicity** (Dempster et al., 1977) of the EM algorithm. Gupta et al. (2011) proved it by using the logarithm version $l(\mathbf{\Phi}) = \log L_{inc}(\mathbf{\Phi})$.

First, define $z_{ik} = \begin{cases} 1 & \mathbf{y}_i \in \text{cluster } k \\ 0 & \mathbf{y}_i \notin \text{cluster } k \end{cases}$, $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$ where only one element out of all $K$ elements is 1, others are 0, and $\zeta$ is the space composed of all possible $\mathbf{z}_i$ such as $\underbrace{(1,0,\ldots,0), (0,1,\ldots,0), \ldots, (0,0,\ldots,1)}_{K \text{ different kinds of } \mathbf{z}_i}$. Then we can derive that

$$
\begin{aligned}
l(\mathbf{\Phi}^{(v+1)}) &= \log L_{inc}(\mathbf{\Phi}^{(v+1)}) \\
&= \log f(\mathbf{Y} \mid \boldsymbol{\theta}^{(v+1)}) \\
&= \log \left\{ \sum_{\mathbf{z}_1 \in \zeta, \ldots, \mathbf{z}_n \in \zeta} L_{com}(\mathbf{\Phi}^{(v+1)} \mid \mathbf{Z}) \right\} \\
&\quad \text{augment the incomplete-data likelihood to be a complete-data likelihood} \\
&= \log \left[ \sum_{\mathbf{Z}} f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)}) \right] \\
&= \log \left[ \sum_{\mathbf{Z}} \frac{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)})}{f(\mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\theta}^{(v)})} \times f(\mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\theta}^{(v)}) \right] \\
&= \log E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \frac{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)})}{f(\mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\theta}^{(v)})} \right] \\
&\geq E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \log \frac{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)})}{f(\mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\theta}^{(v)})} \right] \text{ by Jensen's inequality} \\
&= E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \log \frac{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)})}{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v)}) / f(\mathbf{Y} \mid \boldsymbol{\theta}^{(v)})} \right] \text{ by conditional density} \\
&= E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \log \frac{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)}) \times f(\mathbf{Y} \mid \boldsymbol{\theta}^{(v)})}{f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v)})} \right] \\
&= E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \log f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v+1)}) \right] - E_{\mathbf{Z}|\mathbf{Y}, \boldsymbol{\theta}^{(v)}} \left[ \log f(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\theta}^{(v)}) \right] + \log f(\mathbf{Y} \mid \boldsymbol{\theta}^{(v)}) \\
&= Q(\mathbf{\Phi}^{(v+1)} \mid \mathbf{\Phi}^{(v)}) - Q(\mathbf{\Phi}^{(v)} \mid \mathbf{\Phi}^{(v)}) + l(\mathbf{\Phi}^{(v)}) \qquad (2.4.4)
\end{aligned}
$$

From (2.4.3), we can rearrange (2.4.4) and obtain that

$$
l(\mathbf{\Phi}^{(v+1)}) - l(\mathbf{\Phi}^{(v)}) = Q(\mathbf{\Phi}^{(v+1)} | \mathbf{\Phi}^{(v)}) - Q(\mathbf{\Phi}^{(v)} | \mathbf{\Phi}^{(v)}) \geq 0,
$$

which completes the proof.

This monotonicity can guarantee that the parameter estimates will never get worse in terms of the (incomplete-data) likelihood function as the iteration goes on. The convergence of parameter estimate sequence $\left\{ \mathbf{\Phi}^{(v)} \right\}$ also depends on the characters of $l(\mathbf{\Phi})$, $Q(\mathbf{\Phi} | \mathbf{\Phi}^{(v)})$ and the initial parameter values $\mathbf{\Phi}^{(0)}$.

## 2.5 k-means Clustering

### 2.5.1 Introduction

k-means (MacQueen, 1967) is a clustering method for unsupervised learning, which iterates between assigning each data point to the nearest-mean group and calculating the new means until convergence. k-means can be treated as a variant of the generalised expectation-maximisation algorithm, where the assignment step is regarded as the E-step and the update step as the M-step.

#### 2.5.1.1 Assign Data Point

Each data point $y$ is assigned to the nearest centroid which has the least squared Euclidean distance

$$\underset{m_i^{(t)} \in M^{(t)}}{\arg\min}\, d(m_i^{(t)}, y)^2,$$

where $d(\cdot)$ is the Euclidean distance between these two points, and $M^{(t)} = \left\{m_1^{(t)}, m_2^{(t)}, \ldots, m_k^{(t)}\right\}$ is the set of $k$ means for $k$ groups in the $t$-th iterative refinement.

#### 2.5.1.2 Update Centroid

The iterative refinement is done by recomputing the centroid for each group. The centroid of one group is calculated by averaging all data points assigned to that group:

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{y_j \in S_i^{(t)}} y_j,$$

where $S_i^{(t)}$ is the set of assigned data points for $i$-th group in the $t$-th iteration.

#### 2.5.1.3 Stopping Criteria

The iteration will continue between the above two steps until certain stopping rule is met. Mostly we will use the rule that the clustering result is stable, which means no data point changes its assignment anymore.

Moreover, as the k-means algorithm tries to minimise the within-cluster sum of squares:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{y_j \in S_i} \left\|y_j - m_i\right\|^2, \tag{2.5.1}$$

so sometimes the stopping rule is that (2.5.1) is achieved.

Since k-means is not a convex problem (as a discrete problem), the results can often be

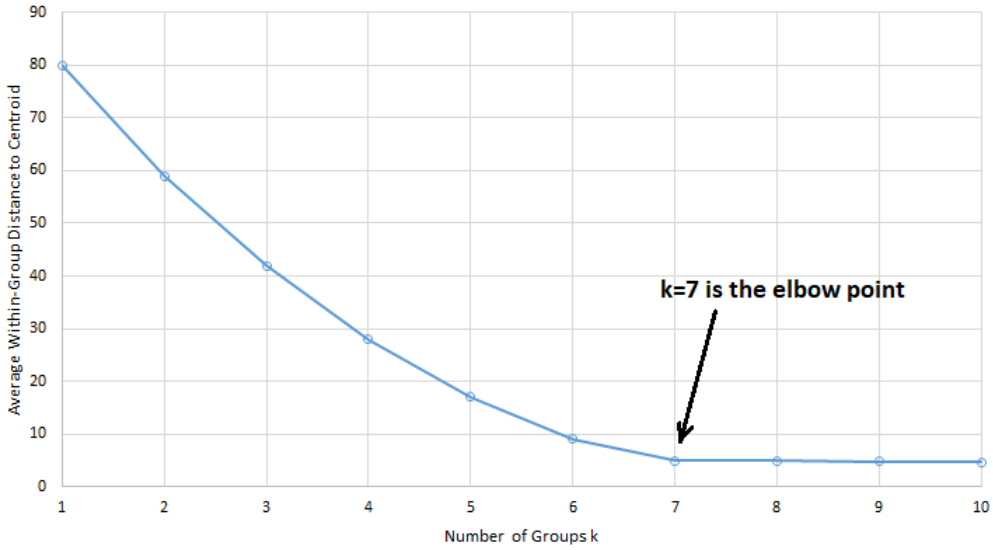



Figure 2.5.1: For different number of groups $k$s(from 1 to 10 groups), this line chart shows the average distance to the within-group centroid for each $k$. Here the within-group centroid is the imaginary or real data point locating at the centre of a cluster.

sub-optimal. We usually restrict it by setting the stopping rule as some maximum number of iterations to allow for repetitive computations.

The convergence of k-means may result in a local optimum which is not a global optimum to produce the best clustering result. So we usually repeat the k-means multiple times with different initial centroids to reach a better clustering result.

#### 2.5.1.4 The Choice of $k$

The k-means algorithm starts with a preset $k$ (the number of groups). To find the best $k$, we need to run the k-means algorithm several times with a range of $k$s and compare the clustering outcome with one certain criterion. Different criteria can bring different $k$s, there is no exact $k$ for most clustering analyses.

Elbow point on the plot of the average distance to the centroid against $k$ is one kind of criterion. An example of one elbow point is shown in Figure 2.5.1. As we can see from this plot, the average distance to the centroid will always decrease when $k$ increases until $k$ equals the number of data points (average distance becomes 0). So the elbow point is chosen ($k = 7$ here) where the decrease rate has a sharp change. Other criteria are also available here such as information criteria and cross-validation.

## 2.6 BIC and K-fold Cross-Validation

### 2.6.1 Introduction to BIC

Bayesian information criterion (BIC; Schwarz, 1978) is an index for model selection among 2 or more alternatives. It is defined as

$$\text{BIC} \quad = \quad k \log(n) - 2 \log(L(\mathbf{\Phi})),$$

where $n$ is the sample size or number of observations, $k$ is the number of parameters estimated in the model, $L(\mathbf{\Phi})$ is the maximised likelihood function. The model with the lowest BIC value is preferred.

When building a model, the likelihood can increase by adding more parameters to it, which can lead to overfitting gradually. To balance good fit with parsimony, a penalty term is added onto the number of parameters which can represent the complexity of the model.

Since the derivation of BIC involves approximations, BIC as a heuristic method is just an approximation to the true distribution of the data. The choice of other criteria such as Akaike information criterion (AIC) are also available, but BIC penalises model complexity more than AIC. The limitation of BIC lies in the sizes of $n$ and $k$, where sample size $n$ needs to be much larger than the number of parameters $k$ in the model.

### 2.6.2 Introduction to K-fold Cross-Validation

Cross-validation is a kind of technique to test the performance of one model's prediction ability by applying it to an independent data set which was not used in estimating this model. Cross-validation can help to assess the model performance and know how accurate predictions this model can give. Data are divided into two parts when fitting a model, one is 'known data' which plays as training data set in the cross-validation to analyse the model, and the other is 'unknown data' which plays as validation data set (testing data set) to validate the model.

K-fold cross-validation is one type of cross-validation, where the given data are partitioned into $K$ equal-sized (similar-sized) subsets randomly, each subset is called one fold. Retain one fold as validation data for testing the model afterwards, the remaining $K-1$ folds are training data for modelling analysis. Then this process is repeated $K$ times to nominate each fold as validation data once in turn. These $K$ results can be averaged to get a more accurate estimate of the model's prediction ability. Usually, multiple rounds of cross-validation are implemented with different partition outcomes to reduce variability. In general, $K = 10$ is commonly used but not fixed, it can be any value depending on the formation of given data such as subset size.

K-fold cross-validation has the advantage that all observations are used both for training the model and validating it in the whole process.

When $K = n$ (the number of observations), K-fold cross-validation becomes leave-one-out cross-validation.

## 2.7 Skew Normal Distribution

Gaussian distribution has been trusted for centuries, but scepticism also exists around there about the validation of normality assumption under many circumstances. To represent a safer initial mental attitude about this doubt, Geary (1947) stated that *'Normality is a myth; there never was, and never will be, a normal distribution'.* Although overstated, it corresponds to the practical viewpoint. Many alternative normal distributions have been put forward to address these non-normality problems with their own desirable properties. Turner (1960) used a heuristic way to look at the estimation method by introducing a generalised density function to include Gaussian (normal) distribution. Prentice (1975) proposed one distribution with a two-parameter density function which can show positive or negative skew from the difference between these two parameters. Azzalini (1985) depicted the asymmetry feature of some skewed normal distributions with a concise density function which can include ordinary normal as a special case. This probability density function for any $Z \sim SN(0, 1, \alpha)$ is given by

$$f(z;\alpha) = 2\phi(z)\,\Phi(\alpha z), \tag{2.7.1}$$

where $\phi$ and $\Phi$ are the standard normal probability density and cumulative distribution functions respectively. Here $\alpha \in (-\infty, +\infty)$ is the parameter regulating skewness. When $\alpha = 0$, (2.7.1) becomes the $N(0, 1)$ normal density function. It resembles ordinary normal distribution in some properties, for instance, $z^2$ follows a degree-one chi-square distribution. And (2.7.1) is strongly unimodal with the potential to help the analysis of the empirically unimodal but skewed data. The skew normal distribution can also be seen as the linear summation of one normal distribution and one half normal distribution. For any $Z \sim SN(0, 1, \alpha)$, if $U_1$ and $U_2$ are independent $N(0, 1)$ variables, then (Azzalini, 1986, 2005)

$$Z = \delta|U_1| + \sqrt{1-\delta^2}U_2,$$

where $\delta = \alpha/\sqrt{1+\alpha^2}$. Now that (2.7.1) is the 'standard' form of skew normal distribution, location ($\mu$) and scale ($\sigma$) parameters are introduced to get the general form (Azzalini and

Capitanio, 1999), the density function of $Y \sim SN\left(\mu, \sigma^2, \alpha\right)$ from $Y = \mu + \sigma Z$ is

$$f\left(y; \mu, \sigma, \alpha\right) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left[\alpha\left(\frac{y-\mu}{\sigma}\right)\right]. \tag{2.7.2}$$

The first three moments and the variance of $Y$ are (Pewsey, 2000)

$$\begin{aligned}
E\left[Y\right] &= \mu + \sigma\delta\sqrt{\frac{2}{\pi}} \\
E\left[Y^2\right] &= \sigma^2 + \mu^2 + 2\sigma\delta\mu\sqrt{\frac{2}{\pi}} \\
E\left[Y^3\right] &= \mu^3 + 3\sqrt{\frac{2}{\pi}}\mu^2\sigma\delta + 3\mu\sigma^2 + 3\sqrt{\frac{2}{\pi}}\sigma^3\delta - \sqrt{\frac{2}{\pi}}\sigma^3\delta^3 \\
\mathrm{Var}\left[Y\right] &= \sigma^2\left(1 - \frac{2\delta^2}{\pi}\right).
\end{aligned}$$

The Pearson's index of skewness of $Y$ is

$$\gamma_1 = \frac{\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right)\delta^3}{\left(1 - \frac{2\delta^2}{\pi}\right)^{3/2}} \in (-0.99527, 0.99527).$$

Arellano-Valle and Azzalini (2008) found a non-quadratic shape of the log-likelihood function when researching skew normal problems. That is, the profile log-likelihood function encounters a local minimum at $\alpha = 0$, which comes from the rank-deficiency of the variance matrix of the score functions when $\alpha = 0$. Here the singularity of the expected Fisher information suggests the difficulty of estimation for the $\alpha$ parameter as it violates the standard asymptotic theory of MLE. Rotnitzky et al. (2000) showed the bimodality for the limiting distribution of estimates and the convergence rate which is slower than usual $O_p\left(n^{-1/2}\right)$ for skewness parameter $\alpha$. To address these problems, Arellano-Valle and Azzalini (2008) proposed an alternative parametrisation 'the standardised version'. By using the mean, variance and Pearson's index of skewness as new parameters, parameter system changes from $\left(\mu, \sigma^2, \alpha\right)$ to $\left(\mu', \sigma'^2, \gamma_1\right)$ where

$$\begin{aligned}
\mu' &= E\left[Y\right] = \mu + \sigma\delta\sqrt{\frac{2}{\pi}}, \\
\sigma'^2 &= \mathrm{Var}\left[Y\right] = \sigma^2\left(1 - \frac{2\delta^2}{\pi}\right), \\
\gamma_1 &= \frac{\sqrt{\frac{2}{\pi}}\left(\frac{4}{\pi} - 1\right)\delta^3}{\left(1 - \frac{2\delta^2}{\pi}\right)^{3/2}}.
\end{aligned}$$

This is defined from $Y = \mu + \sigma Z = \mu' + \sigma' Z_0$ where $Z \sim SN\left(0, 1, \alpha\right)$ and $Z_0 = \frac{Z - \delta\sqrt{2/\pi}}{\sqrt{1 - 2\delta^2/\pi}}$. This parametrisation is named CP (centred parameters) in contrast with the original parametrisation DP (direct parameters) which can be read directly from the original density function (2.7.2).

However, this reparameterisation CP makes it very complicated to carry on the calculations for the statistical inference. In Chapter 4, we adapt it for our calculations in estimation.

## 2.8 Fisher Information and Asymptotic Distribution of MLE

Fisher Information provides a way to measure the amount of information that a random observable variable contains about the underlying unknown parameters from a distribution this random variable is assumed to follow. Given random variable $\boldsymbol{y}$ and parameter $\boldsymbol{\theta}$, the log-likelihood function is $l(\boldsymbol{\theta} \mid \boldsymbol{y})$. Then the Fisher information is the variance of the partial derivative for $\boldsymbol{\theta}$:

$$\mathcal{I}(\boldsymbol{\theta}) = \operatorname{Var}\left[\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta} \mid \boldsymbol{y})\right] = E\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} l(\boldsymbol{\theta} \mid \boldsymbol{y})\right].$$

The matrix form for $N$-parameter case $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$ is a $N \times N$ Fisher information matrix:

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = E\left[\left(\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta} \mid \boldsymbol{y})\right)\left(\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta} \mid \boldsymbol{y})\right)\right] = -E\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta} \mid \boldsymbol{y})\right].$$

The asymptotic distribution of the maximum likelihood estimate is a $N$-dimensional multivariate normal distribution:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{real}\right) \xrightarrow{d} N_N\left(\mathbf{0}, \mathcal{I}_1(\boldsymbol{\theta})^{-1}\right), \tag{2.8.1}$$

where $\boldsymbol{\theta}_{real}$ is the real underlying parameter, $\hat{\boldsymbol{\theta}}$ is the MLE, and $\mathcal{I}_1(\boldsymbol{\theta})$ is the Fisher information for a single observation.

## 2.9 Big-O and Small-O Notations in Probability

$X_n = o_p(a_n)$ means that the set of values $X_n/a_n$ converges to 0 in probability as $n$ approaches a limit. Thus, small-O indicates the convergence in probability. $X_n = O_p(a_n)$ means that the set of values $X_n/a_n$ is stochastically bounded. Thus, big-O notation means stochastic boundedness here. The definition for $O_p(1)$ and $o_p(1)$ are as follows (Dodge, 2003):

$$O_p(1): \forall \varepsilon \quad \exists N_\varepsilon, \delta_\varepsilon \text{ such that } P(|X_n| \geq \delta_\varepsilon) \leq \varepsilon \quad \forall n > N_\varepsilon,$$

$$o_p(1): \forall \varepsilon, \delta \quad \exists N_{\varepsilon,\delta} \text{ such that } P(|X_n| \geq \delta) \leq \varepsilon \quad \forall n > N_{\varepsilon,\delta}.$$

Then we can affirm that if a set of random variables $X_n \sim N(0,1)$, then $X_n = O_p(1)$.

*Proof.* For $\delta_\varepsilon > 0$

$$\begin{aligned} P\left(|X_n| \geq \delta_\varepsilon\right) &= 1 - \Phi\left(\delta_\varepsilon\right) + \Phi\left(-\delta_\varepsilon\right) \\ &= 2 - 2\Phi\left(\delta_\varepsilon\right), \end{aligned}$$

$\forall \varepsilon > 0$ that

$$2 - 2\Phi\left(\delta_\varepsilon\right) \leq \varepsilon \iff \delta_\varepsilon \geq \Phi^{-1}\left(\frac{2-\varepsilon}{2}\right),$$

$$\sup_{\varepsilon>0} P\left(|X_n| \geq \delta_\varepsilon\right) = P\left(|X_n| \geq \Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)\right) = \varepsilon.$$

Thus, it is proved $\forall \varepsilon, \forall n, \exists \delta_\varepsilon = \Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)$, such that

$$P\left(|X_n| \geq \delta_\varepsilon\right) \leq \varepsilon,$$

namely $X_n = O_p\left(1\right)$. □

Moreover, it can also be known that $\frac{X_n}{\sqrt{n}} = o_p\left(1\right)$.

*Proof.* As we have obtained from the previous proof that $\forall \varepsilon$ and $\forall n$,

$$P\left(|X_n| \geq \Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)\right) \leq \varepsilon.$$

Thus, $\forall \delta > 0$ we can calculate

$$P\left(\left|\frac{X_n}{\sqrt{n}}\right| \geq \delta\right) = P\left(|X_n| \geq \sqrt{n}\delta\right) \leq \varepsilon \iff \sqrt{n}\delta \geq \Phi^{-1}\left(\frac{2-\varepsilon}{2}\right) \iff \delta \geq \frac{\Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)}{\sqrt{n}}.$$

Moreover, as $\varepsilon > 0$

$$\lim_{n\to+\infty} \frac{\Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)}{\sqrt{n}} = 0.$$

Therefore, $\forall \delta > 0$ and $\forall \varepsilon > 0$, $\exists N_{\varepsilon,\delta} = \left[\frac{\Phi^{-1}\left(\frac{2-\varepsilon}{2}\right)}{\delta}\right]^2$, such that $\forall n > N_{\varepsilon,\delta}$

$$P\left(\left|\frac{X_n}{\sqrt{n}}\right| \geq \delta\right) \leq \varepsilon,$$

namely $\frac{X_n}{\sqrt{n}} = o_p\left(1\right)$. □

For $y_i \sim SN\left(\tilde{\mu} - \tilde{\sigma}\tilde{\delta}\sqrt{\frac{2}{\pi}}, \tilde{\sigma}^2, \tilde{\alpha} = 0\right) = SN\left(\tilde{\mu}, \tilde{\sigma}^2, 0\right)$, we can derive that

$$y_i \sim N\left(\tilde{\mu}, \tilde{\sigma}^2\right) \Longrightarrow \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \sim N\left(0,1\right) \Longrightarrow \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}}\right) \sim N\left(0,1\right),$$

thus, with known knowledge from previous proof, we can reach

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}}\right) = \frac{\frac{1}{\sqrt{n}} \sum\limits_{i=1}^{n} \left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}}\right)}{\sqrt{n}} = o_p\left(1\right) = O_p\left(n^{-1/2}\right).$$

These conclusions can help understand the proofs in **Appendix B**.

## 2.10 Some Other Knowledge

### 2.10.1 Markov's Inequality

Let $X$ be a non-negative random variable and $\epsilon > 0$, then the following inequality holds

$$\Pr\left(X \geq \epsilon\right) \quad \leq \quad \frac{E\left[X\right]}{\epsilon},$$

where $E\left[X\right]$ is the expectation of $X$.

### 2.10.2 Chebyshev's Inequality

$X$ is a random variable with a finite mean $\mu$ and a finite variance $\sigma^2$. For any real number $\varepsilon > 0$,

$$\Pr\left(\left|X - \mu\right| \geq \varepsilon\right) \quad \leq \quad \frac{\sigma^2}{\varepsilon^2}.$$

Sometimes Markov's inequality is termed *'Chebyshev's First Inequality'*. We can use a proof to connect Chebyshev's inequality with Markov's inequality.

*Proof.* Use Markov's inequality to obtain

$$\Pr\left(\left(X - \mu\right)^2 \geq \varepsilon^2\right) \quad \leq \quad \frac{E\left[\left(X - \mu\right)^2\right]}{\varepsilon^2}.$$

Because $\left(X - \mu\right)^2 \geq \varepsilon^2 \iff \left|X - \mu\right| \geq \varepsilon$, then we can write

$$\Pr\left(\left|X - \mu\right| \geq \varepsilon\right) \quad \leq \quad \frac{E\left[\left(X - \mu\right)^2\right]}{\varepsilon^2}.$$

Since we have known the variance definition

$$E\left[(X-\mu)^2\right] \quad = \quad \sigma^2,$$

therefore, it is proved that

$$\Pr\left(|X-\mu| \geq \varepsilon\right) \quad \leq \quad \frac{\sigma^2}{\varepsilon^2}.$$

□

### 2.10.3 Slutsky's Theorem

Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random elements (scalar/vector/matrix). If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{P} c$ ($c$ is a constant), then the continuous function $g(x,y)$ has the property

$$g\left(X_n, Y_n\right) \quad \xrightarrow{d} \quad g\left(X, c\right),$$

where '$\xrightarrow{d}$' denotes convergence in distribution and '$\xrightarrow{P}$' denotes convergence in probability.

If we replace '$\xrightarrow{d}$' with '$\xrightarrow{P}$' everywhere, this theorem still remains valid.

# Chapter 3

# Multivariate Mixture Regression Models (with and without Missingness)

## 3.1 Complete Case Analysis by Penalised Likelihood Fusion

### 3.1.1 Introduction

Nowadays, high-throughput technology has made massive biological and chemical tests possible in a short time, which could provide enormous multivariate data to researchers. But most of the time, we got lost in these massive datasets, like what Rutherford D. Rogers has said *'We are drowning in information and starving for knowledge'*. That is why we need statistical knowledge to sort out essentials through oceans of information. As for our research, stress is laid on anti-cancer drugs and different genes. There are over 30000 genes in the human body, but we don't expect all of them are connected with the growth of a certain kind of cancer. With a penalised model, we can screen out a small proportion of them to conduct informative clustering and classification to power the discovery of anti-cancer drugs in the future.

### 3.1.2 Preprocess Data

Our research uses the dataset as introduced in Section 1.2, which has two parts: drug sensitivity and gene expressions. To connect these two parts, we keep the common observations (cell lines) in these two matrices after removing these rows with missingness for the drug data matrix ($\mathbf{Y}$)

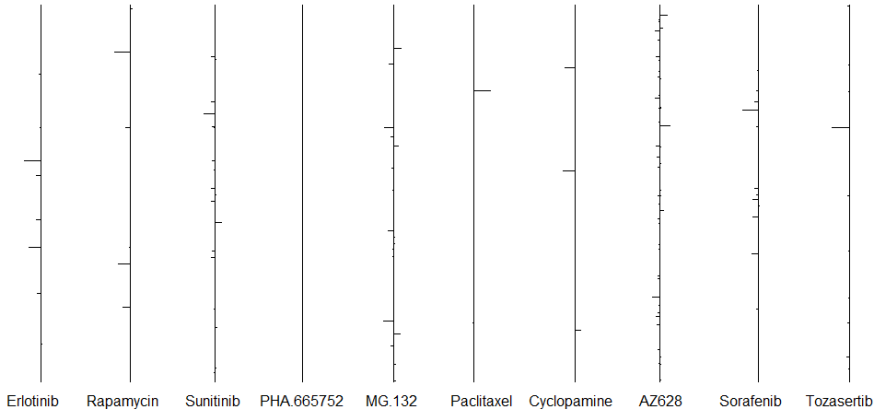



Figure 3.1.1: The regression coefficient plot from LASSO regression for 10 drugs. All 17737 genes (unlabelled) run from top to bottom in each vertical long line. Small line segments pointing to the right indicate positive weights, and to the left, negative weights.

to make it a complete-case problem. After deletions, 227 anti-cancer drugs, 111 cell lines and 17737 genes are screened out for the next step of our research. Here two big matrices $\mathbf{Y}$ and $\mathbf{X}$ are constructed from the aforementioned two parts. $\mathbf{Y}$ is a $111 \times 227$ response matrix which consists of 111 cell lines (COSMIC_ID) tested on 227 anti-cancer drugs. $\mathbf{X}$ is a $111 \times 17737$ covariate matrix where the corresponding 111 cell lines (COSMIC_ID) are tested to obtain the expression data across 17737 genes.

### 3.1.3 Preliminary Analysis of Group Structure

Anti-cancer drugs always manifest a hidden group structure in which the drugs in one group may help with the same treatment or share similar chemical characteristics. To identify the underlying group structures, we carried out the following preliminary analysis with the help of LASSO from the *glmnet* package and the Jaccard Index (Jaccard, 1912) for similarity comparison. Then we grouped the drugs into reasonable clusters and displayed them with clustering plots plotted by Gephi (Bastian et al., 2009).

Here LASSO regression is carried out on each drug with all applicable genes. The bar plots are drawn to indicate the weights of coefficients for each anti-cancer drug. Figure 3.1.1 takes out 10 drugs as examples. They show the feature of sparsity for LASSO regression, where only a handful of gene weights are set as non-zero for each anti-cancer drug. This sparsity feature can help building our model in the next section.

The Jaccard index (Jaccard, 1912) is a statistic used to compare the similarity between two sample sets, it is defined as

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

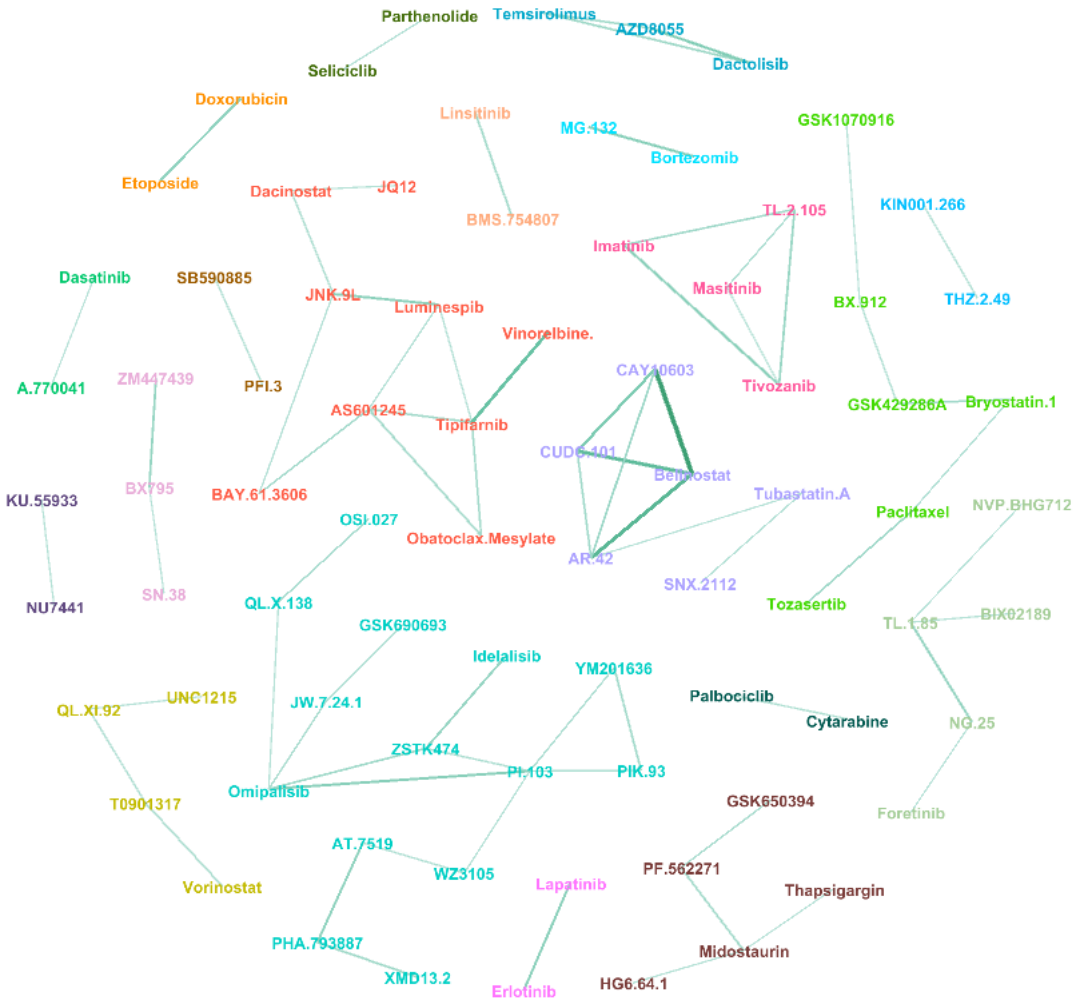



Figure 3.1.2: 79 anti-cancer drugs are grouped into 20 clusters. Darker green the edge between two nodes is, the greater weight they have. For the name tags, different colours are assigned to different clusters.

Obviously, $0 \leq J(A,B) \leq 1$. For our analysis, $A$ and $B$ are the sets of active genes for two drugs respectively. And $J(A,B)$ is set as 0 when $A$ and $B$ are both empty since some cell lines are deleted to exclude missingness before the LASSO regression.

To group these drugs, we need a Jaccard index benchmark $J$. We set $J = 10\%$ as a default value to indicate the group structure of the drugs in the next step. Then two drugs with the Jaccard index of no less than $J$ are considered coming from the same cluster. Figure 3.1.2 is plotted using Gephi to show the clustering results with $J = 10\%$.

Using the same dataset, Iorio et al. (2016) performed an anti-cancer drug clustering based on the similarity of the area under curve (AUC) pattern. In their results, 223 drugs are grouped into 26 clusters. Compared with my result about clustering, there are some similar clusters, which can partly support my clustering result based on LASSO regression and Jaccard index.

Here we put forward a method to realise clustering which is easy to carry out. This method can yield reasonable results. As for the Jaccard index benchmark $J$, it can determine both

the number of clusters and the number of drugs participating in our clustering. After all, the underlying group structure within these 227 anti-cancer drugs can be revealed, which can enhance the theoretical support for our penalised mixture modelling in the remaining parts of this chapter.

### 3.1.4 Penalised Likelihood Fusion

As shown from our preliminary analysis, anti-cancer drugs indeed manifest a hidden group structure. Recent research has verified that drug combination treatment is more effective not only in killing tumour cells but also in drug resistance reduction when certain drugs are applied simultaneously (Carter et al., 2016). Anti-cancer drugs exert their functions by binding with protein targets. Correlating drugs with biomarkers can help identify their molecular targets to power the discovery of new drugs and disease-specific therapies. In our research, it is a multivariate mixture regression problem, so we turn to the multivariate Gaussian mixture model for our next modelling.

Similar to the finite mixture model structure, we propose the likelihood fusion model (Oftadeh, 2017) to start our regression. The likelihood fusion model is a mixture model composed of many likelihood density functions, where different anti-cancer drugs in our data are combined in one model. As shown in the matrix $\mathbf{Y}$, each anti-cancer drug out of $J = 227$ drugs has recorded $n = 111$ observations, which means each column $\mathbf{y}_j = (y_{1j}, \ldots, y_{nj})^T$ is a $n \times 1$ vector. Then the conditional density of anti-cancer drug $j$ under the condition of $k$th component $\left(\mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2\right)$ is

$$\begin{aligned} f_k(\mathbf{y}_j|\alpha_j, \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2) =& \prod_{i=1}^{n} f_k^i(y_{ij}|\alpha_j, \mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2) \\ =& \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma_k^2\right)^{1/2}} \exp\left\{-\frac{\left(y_{ij} - \alpha_j - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2}\right\} \\ =& \frac{1}{\left(2\pi\sigma_k^2\right)^{n/2}} \exp\left\{-\frac{\left(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right\}, \end{aligned}$$

where $\boldsymbol{\beta}_k$ is the coefficient and $\sigma_k^2$ is the variance for component $k$, and $\alpha_j$ is the intercept. This is a multivariate normal density with respect to the error term $\boldsymbol{\varepsilon}_k$ in $\mathbf{y}_j - \alpha_j\mathbf{1} = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k$, where $\alpha_j$ is the intercept, and $\mathbf{1}$ is the vector of $n$ ones. And it can be expressed as $\boldsymbol{\varepsilon}_k \sim N_{n=111}\left(\mathbf{0}, \mathrm{diag}(\sigma_k^2)\right)$, where $\mathbf{0}$ is the vector of $n$ zeros, and independence is assumed for each dimension from the diagonal covariance matrix setting. The likelihood function can be regarded as a joint distribution with $n$ independent, identically distributed (i.i.d.) observations. If we assume that drug $j$ has the probability $\pi_k$ of originating from component $k$, then the likelihood

fusion model is constructed:

$$f(\mathbf{y}_j|\mathbf{X},\mathbf{\Phi}_j) \quad = \quad \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}_j|\alpha_j, \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2),$$

where $\mathbf{\Phi}_j$ stands for all the parameters needed to estimate such as $\alpha_j, \boldsymbol{\beta}, \boldsymbol{\pi}$ and $\boldsymbol{\sigma}$. It is a weighted summation of $K$ density functions, and many properties can be inherited from the ordinary Gaussian mixture model. Inspired by the Laplace-inverse-gamma priors and the penalised likelihood function in the literature of Zhang (2017), we impose a penalty term onto the likelihood fusion model to realise variable selection purpose. The $l_1-$norm penalty has been used before in multivariate mixture model problems, see Pan and Shen (2007), Zhou and Pan (2009). Then we have the following fused likelihood:

$$\underbrace{\prod_{j=1}^{J}}_{\text{All Drugs}} \underbrace{\sum_{k=1}^{K} \frac{\pi_k}{(2\pi\sigma_k^2)^{n/2}} \exp\left\{-\frac{(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{2\sigma_k^2}\right\}}_{\text{Likelihood Fusion}} \underbrace{\times \exp\left(-\frac{n\lambda\|\boldsymbol{\beta}_k\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^2}}_{\text{Penalty Term}}. \tag{3.1.1}$$

It comes from the penalisation on each observation of each drug, which can be decomposed into

$$\begin{aligned} g_k(\mathbf{y}_j|\alpha_j, \mathbf{X}\boldsymbol{\beta}_k, \sigma_k^2) &= \prod_{i=1}^{n} f_k^i(y_{ij}|\alpha_j, \mathbf{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2) \times \exp\left(-\frac{\lambda\|\boldsymbol{\beta}_k\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^{2/n}} \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_k^2)^{1/2}} \exp\left\{-\frac{(y_{ij} - \alpha_j - \mathbf{x}_i^T\boldsymbol{\beta}_k)^2}{2\sigma_k^2}\right\} \exp\left(-\frac{\lambda\|\boldsymbol{\beta}_k\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^{2/n}} \\ &= \frac{1}{(2\pi\sigma_k^2)^{n/2}} \exp\left\{-\frac{(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T(\mathbf{y}_j - \alpha_j\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{2\sigma_k^2}\right\} \\ &\quad \times \exp\left(-\frac{n\lambda\|\boldsymbol{\beta}_k\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^2}. \end{aligned}$$

Generally speaking, our target is to find maximum likelihood estimates of the parameters in a penalised likelihood function which is a mixture of multivariate normal density. And the expectation-maximisation (EM) algorithm is a good choice for solving this kind of mixture problem. McLachlan and Basford (1988) firstly demonstrated the use of EM algorithm to fit the normal mixture models with the data from different fields. EM algorithm can do iterations between the expectation step and the maximisation step until convergence is reached. In our calculation, the *glmnet* package in R is used to help solve the LASSO problem when we are updating the parameter $\boldsymbol{\beta}_k$, which will be explained afterwards.

(3.1.1) is the **incomplete-data** likelihood function, it comes from the combination of $J$

independent drugs:

$$\begin{aligned} pL_{inc}(\mathbf{\Phi}) &= \prod_{j=1}^{J} g(\mathbf{y}_j|\mathbf{X},\mathbf{\Phi}_j) \\ &= \prod_{j=1}^{J}\sum_{k=1}^{K} \pi_k g_k(\mathbf{y}_j|\alpha_j,\mathbf{X}\boldsymbol{\beta}_k,\sigma_k^2) \\ &= \prod_{j=1}^{J}\sum_{k=1}^{K} \frac{\pi_k}{\left(2\pi\sigma_k^2\right)^{n/2}} \exp\left\{-\frac{\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right\} \\ &\quad \times \exp\left(-\frac{n\lambda\left\|\boldsymbol{\beta}_k\right\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^2}, \end{aligned}$$

where $\mathbf{\Phi}$ represents $(\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_J)$.

With the hidden (latent) variable $z$ which is the group label for each drug $j$, we can enhance the **incomplete-data** likelihood function to a **complete-data** likelihood function:

$$\begin{aligned} pL_{com}(\mathbf{\Phi}) &= \prod_{j=1}^{J}\prod_{k=1}^{K}\left[\pi_k g_k(\mathbf{y}_j|\alpha_j,\mathbf{X}\boldsymbol{\beta}_k,\sigma_k^2)\right]^{z_{jk}} \\ &= \prod_{j=1}^{J}\prod_{k=1}^{K}\left\{\pi_k\prod_{i=1}^{n}\frac{1}{\left(2\pi\sigma_k^2\right)^{1/2}}\exp\left[-\frac{\left(y_{ij}-\alpha_j-\mathbf{x}_i^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2}\right]\right. \\ &\quad \left.\times\exp\left(-\frac{\lambda\left\|\boldsymbol{\beta}_k\right\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^{2/n}}\right\}^{z_{jk}}, \end{aligned}$$

where $z_{jk} = 1$ when drug $j$ belongs to group $k$, otherwise $z_{jk} = 0$. It works as an indicator.

Then the corresponding **complete-data log-likelihood** function is

$$\begin{aligned} pl_{com}(\mathbf{\Phi}) &= \log\left(pL_{com}(\mathbf{\Phi})\right) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\log\left(\pi_k\right) + \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{jk}\log\left(\frac{1}{\sqrt{2\pi}\sigma_k^{1+2/n}}\right) \\ &\quad -\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{jk}\frac{\left(y_{ij}-\alpha_j-\mathbf{x}_i^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2} - \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n} z_{jk}\frac{\lambda\left\|\boldsymbol{\beta}_k\right\|_1}{\sigma_k} \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log\left(\pi_k\right) - \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log\left((2\pi)^{n/2}\sigma_k^{n+2}\right) \\ &\quad -\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{z_{jk}\cdot\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)}{2\sigma_k^2} \\ &\quad -\sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\frac{n\lambda\left\|\boldsymbol{\beta}_k\right\|_1}{\sigma_k}. \end{aligned} \tag{3.1.2}$$

The step-by-step calculations for this EM algorithm is put in **Appendix A** as it is similar to our Section 3.2.2 in the next Incomplete Case part.

### 3.1.5 The Choices of $\lambda$ and $K$

To find the best number of components $K$ and the tuning constant $\lambda$, we used Bayesian information criterion (Schwarz, 1978; see Section 2.6.1). First, construct a grid of $\lambda$s (from 0.003 to 0.06 with interval 0.003 for our dataset, it comes from many empirical pilot tests) and $K$ (integers from 1 to 10). Then start the EM algorithm with each $K$ and $\lambda$, calculate the $\text{BIC}_{\lambda,K}$ every time with the estimated parameters $\hat{\boldsymbol{\beta}}_k$, $\hat{\sigma}_k^2$ and $\hat{\pi}_k$ using the formula below

$$\text{BIC}_{\lambda,K} = \left( K - 1 + K + \sum_{i=1}^{p}\sum_{k=1}^{K} \mathbb{I}(\hat{\beta}_{ki} \neq 0) \right) \times \frac{\log(J)}{J} - \frac{2}{J} \times l_{inc}(\hat{\boldsymbol{\Phi}}),$$

where $\mathbb{I}(x)$ is the indicator function to count how many coefficients are non-zero, and $l_{inc}(\hat{\boldsymbol{\Phi}})$ is

$$l_{inc}(\hat{\boldsymbol{\Phi}}) = \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \frac{\hat{\pi}_k}{(2\pi\hat{\sigma}_k^2)^{n/2}} \exp \left[ - \frac{\left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\right)^T \left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\right)}{2\hat{\sigma}_k^2} \right] \right\}.$$

After that, the minimum $\text{BIC}_{\lambda,K}$ value is chosen and the best $K$ and $\lambda$ are determined from it.

### 3.1.6 Simulation

To test the screening ability and clustering accuracy, we carried out this simulation. Firstly, we created 3 groups with sparse matrix $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ where active points are marked with red hollow triangles in Figure 3.1.3, and the error term $\boldsymbol{\varepsilon}_k \sim N_n\left(\mathbf{0}, \text{diag}(\sigma_k^2)\right)$ where $(\sigma_1, \sigma_2, \sigma_3) = (0.5, 1, 1.5)$. Then we extracted a small part from gene expression data matrix $\mathbf{X}$ as $\mathbf{X}'_{25\times200}$. With assignment possibilities $\boldsymbol{\pi} = (20\%,\ 30\%,\ 50\%)$, we generated the response vectors $\mathbf{y}_j - \bar{y_j}\mathbf{1}$ for 50 drugs using $\mathbf{y}_j - \bar{y_j}\mathbf{1} = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k$. After that, with these $\mathbf{X}'_{25\times200}$ and $\mathbf{Y}'_{25\times50}$, we performed the penalised likelihood fusion modelling and estimated it with the EM algorithm as introduced. Through re-running with the 'best' $\lambda$ and $K$ after our BIC selection, these drugs are clustered into 3 groups. The estimate $\hat{\boldsymbol{\beta}}$ (black hollow dots) and its preset true value $\boldsymbol{\beta}$ (red hollow triangles) are plotted in Figure 3.1.3. As shown on the graphs, estimates come out with some noise variables, but noises are much smaller than our estimates. Almost all preset active (non-zero) variables are detected with shrunken estimates. The result of this simulation can verify that when the data have a hidden group structure and a sparse-coefficient feature, our modelling can classify these data and screen out these relevant variables with good accuracy.

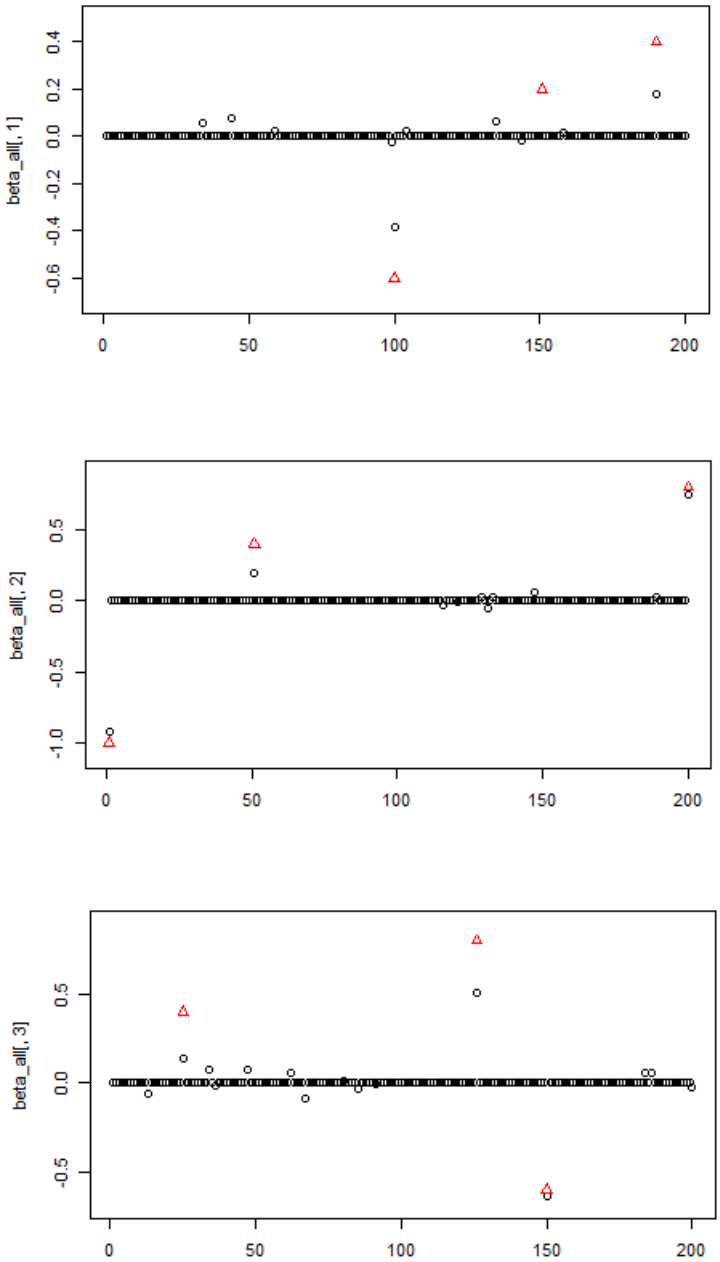



Figure 3.1.3: The coefficient plots for all three drug groups to show the performance of penalised likelihood fusion model. Red hollow triangular markers are the preset true values of $\boldsymbol{\beta}$. Black hollow circular markers are our estimates $\hat{\boldsymbol{\beta}}$.

### 3.1.7 Real Data Performance

We have introduced the structures of real data matrices at the beginning of this chapter (Section 3.1.2). We start from these two matrices $\mathbf{X}_{111\times17737}$ and $\mathbf{Y}_{111\times227}$ here. Due to the large number of genes, it is a high-dimensional problem so it is not a good choice to apply all 17737 genes into our modelling considering the computing capacity of our computers. Dimension-reduction methods on explanatory variables are needed before our penalised regression modelling.

#### 3.1.7.1 Individual LASSO

With the help of the package *glmnet* (Friedman et al., 2015), we apply LASSO to each drug across all genes (drug $j$ is shown here):

$$\hat{\boldsymbol{\beta}}_j(\lambda_j) = \arg\min_{\boldsymbol{\beta}} \left( \left\| \mathbf{y}_j - \mathbf{X}\boldsymbol{\beta}_j \right\|_2^2 / n + \lambda_j \left\| \boldsymbol{\beta}_j \right\|_1 \right), \tag{3.1.3}$$

where $\mathbf{X}$ is the column-wise standardised matrix, and $\mathbf{y}_j$ is the centralised vector for drug $j$ namely the centralised $j$th column of $\mathbf{Y}$. After these operations, the intercept for (3.1.3) becomes zero and can be removed from the formula.

After 227 drug-wise LASSO regressions, we can construct a coefficient matrix $\boldsymbol{B}_{17737\times227}$. By checking this matrix, the genes which have no effects on any drugs are removed from our data. After this step, 3800 out of 17737 genes are retained for our next screening step.

#### 3.1.7.2 Marginal Likelihood Fusion

To screen out fewer genes for our next main analysis, we implement a marginal regression (the EM algorithm calculations here are similar to those in **Appendix A**) with each gene across all drugs using the likelihood fusion model. The **incomplete-data** likelihood for the $i$-th gene is

$$L_{inc}(\boldsymbol{\Phi}_i) = \prod_{j=1}^{J}\sum_{k=1}^{K} \frac{\pi_k}{(2\pi\sigma_k^2)^{n/2}} \exp\left[ -\frac{(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{x}_i\beta_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{x}_i\beta_k)}{2\sigma_k^2} \right].$$

With the latent variable $z_{jk}$, the **complete-data** likelihood for the $i$-th gene is

$$L_{com}(\boldsymbol{\Phi}_i) = \prod_{j=1}^{J}\prod_{k=1}^{K} \left\{ \frac{\pi_k}{(2\pi\sigma_k^2)^{n/2}} \exp\left[ -\frac{(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{x}_i\beta_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{x}_i\beta_k)}{2\sigma_k^2} \right] \right\}^{z_{jk}}.$$

For the convenience of calculation, we use the **complete-data log-likelihood**

$$\begin{aligned} l_{com}(\mathbf{\Phi}_i) &= \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log(\pi_k) - \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log\left((2\pi)^{n/2}\sigma_k^n\right) \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{z_{jk}\cdot(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k)^T(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k)}{2\sigma_k^2}. \end{aligned}$$

Then replace $z_{jk}$ with *membership probability* $\tau_{jk}^{(v)}$ to reach the expectation function $Q$:

$$\begin{aligned} Q &= Q(\mathbf{\Phi}_i|\mathbf{\Phi}_i^{(v)}) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log(\pi_k) - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left((2\pi)^{n/2}\sigma_k^n\right) \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}\cdot(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k)^T(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k)}{2\sigma_k^2}, \end{aligned}$$

where

$$\begin{aligned} \tau_{jk}^{(v)} &= \frac{\pi_k^{(v)} f_k(\mathbf{y}_j|\mathbf{x}_i\beta_k^{(v)},\sigma_k^{(v)2})}{\sum_{k=1}^{K}\pi_k^{(v)} f_k(\mathbf{y}_j|\mathbf{x}_i\beta_k^{(v)},\sigma_k^{(v)2})} \\ &= \frac{\frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n/2}}\exp\left\{-\frac{\left(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k^{(v)}\right)^T\left(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k^{(v)}\right)}{2\sigma_k^{(v)2}}\right\}}{\sum_{k=1}^{K}\frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n/2}}\exp\left\{-\frac{\left(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k^{(v)}\right)^T\left(\mathbf{y}_j-\bar{y_j}\mathbf{1}-\mathbf{x}_i\beta_k^{(v)}\right)}{2\sigma_k^{(v)2}}\right\}}. \end{aligned}$$

From here, the EM algorithm can start by updating the parameter estimates as below:

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{j=1}^{J}\tau_{jk}^{(v)}}{J}, \qquad \hat{\beta}_k^{(v+1)} = \frac{\mathbf{X}^{*T}\mathbf{Y}^*}{\mathbf{X}^{*T}\mathbf{X}^*},$$

where

$$\mathbf{X}^* = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}}\cdot\mathbf{x}_i \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}}\cdot\mathbf{x}_i \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}}\cdot\mathbf{x}_i \end{pmatrix}, \qquad \mathbf{Y}^* = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}}\cdot(\mathbf{y}_1-\bar{y_1}\mathbf{1}) \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}}\cdot(\mathbf{y}_2-\bar{y_2}\mathbf{1}) \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}}\cdot(\mathbf{y}_J-\bar{y_J}\mathbf{1}) \end{pmatrix}.$$

For given $k$, $\mathbf{X}^*$ is a $nJ\times 1$ vector here compared with the $nJ\times p$ matrix (6.2.4) in the full model, because only one gene is applied into the modelling for the marginal analysis, where

$p = 1$ here. And $\mathbf{Y}^*$ is a $nJ \times 1$ vector the same as (6.2.5).

$$\hat{\sigma}_k^{(v+1)} = \sqrt{\frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot (\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \beta_k)^T (\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \beta_k)}{n \sum_{j=1}^{J} \tau_{jk}^{(v)}}}.$$

We use the stopping rule with default $\epsilon = 10^{-4}$ as follows:

$$\left| \frac{l_{inc}(\mathbf{\Phi}_i^{(v+1)}) - l_{inc}(\mathbf{\Phi}_i^{(v)})}{l_{inc}(\mathbf{\Phi}_i^{(v)})} \right| < \epsilon,$$

where

$$l_{inc}(\mathbf{\Phi}_i) = \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \frac{\pi_k}{(2\pi\sigma_k^2)^{n/2}} \exp \left[ -\frac{(\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \beta_k)^T (\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \beta_k)}{2\sigma_k^2} \right] \right\}.$$

Then the EM algorithm for the $i$-th gene applied in a marginal model is done. The estimates of parameters $\hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2, \hat{\pi}_k$ are used to calculate $\mathrm{BIC}_{i,K}$ value:

$$\mathrm{BIC}_{i,K} = \left( K - 1 + K + \sum_{k=1}^{K} \mathbb{I}(\hat{\beta}_k \neq 0) \right) \times \frac{\log(J)}{J} - \frac{2}{J} \times l(\hat{\mathbf{\Phi}}_i),$$

where

$$l(\hat{\mathbf{\Phi}}_i) = \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \frac{\hat{\pi}_k}{(2\pi\hat{\sigma}_k^2)^{n/2}} \exp \left[ -\frac{\left(\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \hat{\beta}_k\right)^T \left(\mathbf{y}_j - \bar{y}_j \mathbf{1} - \mathbf{x}_i \hat{\beta}_k\right)}{2\hat{\sigma}_k^2} \right] \right\}.$$

In our programming, we set $K = 1, 2, \ldots, 16$ (chosen from many pilot tests), and the minimum BIC value for each gene is kept to continue our next step, for the $i$th gene,

$$\mathrm{BIC}_i = \min_{1 \leq K \leq 16} (\mathrm{BIC}_{i,K}).$$

After getting these 3800 selected $\mathrm{BIC}_i$s, our next step continues by using the reciprocals of them, for the $i$th gene,

$$\mathrm{rBIC}_i = \frac{1}{\mathrm{BIC}_i}.$$

To cluster these 3800 genes into two groups, we utilised k-means clustering method on $\mathrm{rBIC}_i$s. The ordered $\mathrm{rBIC}_i$s and the separator line (red) are plotted in Figure 3.1.4. The left part of the red line is the group with a larger cluster mean. In this group, 872 genes with the larger $\mathrm{rBIC}_i$ values are screened out for our next full model analysis.

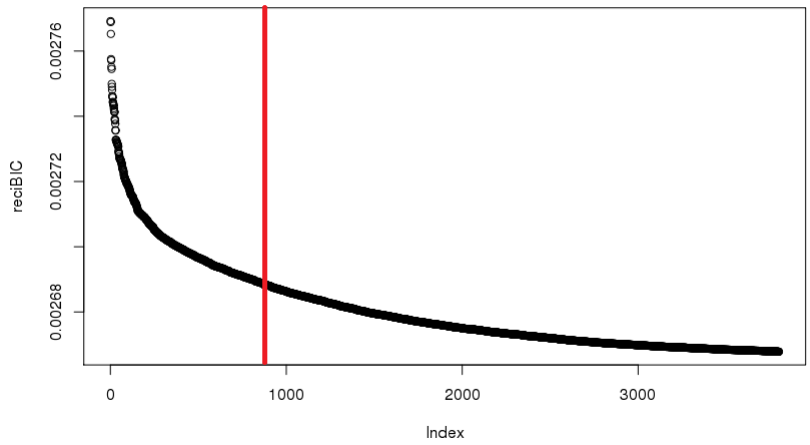



Figure 3.1.4: The rBIC values for different genes from high to low with red vertical line classifying them into two groups based on k-means clustering method.

#### 3.1.7.3 Applying to the Full Model

After the screening with the marginal likelihood fusion, 872 genes are held for the next penalised likelihood fusion modelling. Thus, $\mathbf{X}_{111\times872}$ and $\mathbf{Y}_{111\times227}$ are applied to our full model (3.1.1). After computation, the lowest (best) BIC locates at $(\lambda = 0.015, K = 9)$, which means 227 drugs are clustered into 9 groups. This result is plotted in Figure 3.1.5 where all anti-cancer drug names are coloured with nine different colours to represent nine groups, and these clusters have different sizes. For example, Group I has just one drug inside.

In Figure 3.1.6, $\boldsymbol{\beta}$ (coefficient) is estimated for each group. These coefficients can indicate the degrees of relevance to the selected genes. The red horizontal lines are the testing bounds (the hypothesis test assuming the coefficient is not 0) with a significance level of 5%. The significant genes are screened out by locating outside of the bounds, while the noise variables locate within the bounds.

Then we exhibit these significant genes marked with their corresponding group names and colours in Figure 3.1.7. The correlation between the drugs and genes can be obtained when connecting with Figure 3.1.5.

To give an overview of the results, we summarise these 3 graphs in Table 3.1. Since we have made an independence assumption for these drugs, the covariance matrix is a diagonal matrix $\sigma_k^2 I_n$ for each group. So we only need to estimate $\sigma_k^2$ for the covariance here.

#### 3.1.7.4 Cross-Validation and Comparison with *mclust*

Then we divide $J = 227$ drugs into 5 folds randomly to start our cross-validation. Four out of five folds are set as training data $\mathbf{Y}_{-v}$, the remaining one is set as validation data $\mathbf{Y}_v$. The

Figure 3.1.5: The clustering result for anti-cancer drugs. There are 9 groups in total with each group displayed in a different colour.

| Group Name | Group Size (Drugs) | Number of Significant Genes | $\sigma_k^2$ in $\sigma_k^2 I_n$ | Marking Colour |
|---|---|---|---|---|
| A | 15 | 6 | 0.22 | |
| B | 25 | 6 | 0.41 | |
| C | 43 | 8 | 0.80 | |
| D | 32 | 3 | 0.81 | |
| E | 36 | 4 | 1.75 | |
| F | 37 | 5 | 1.72 | |
| G | 34 | 4 | 1.63 | |
| H | 4 | 5 | 1.09 | |
| I | 1 | 11 | 0.06 | |

Table 3.1: The summary table for the clustering and variable selection results using our penalised likelihood fusion.

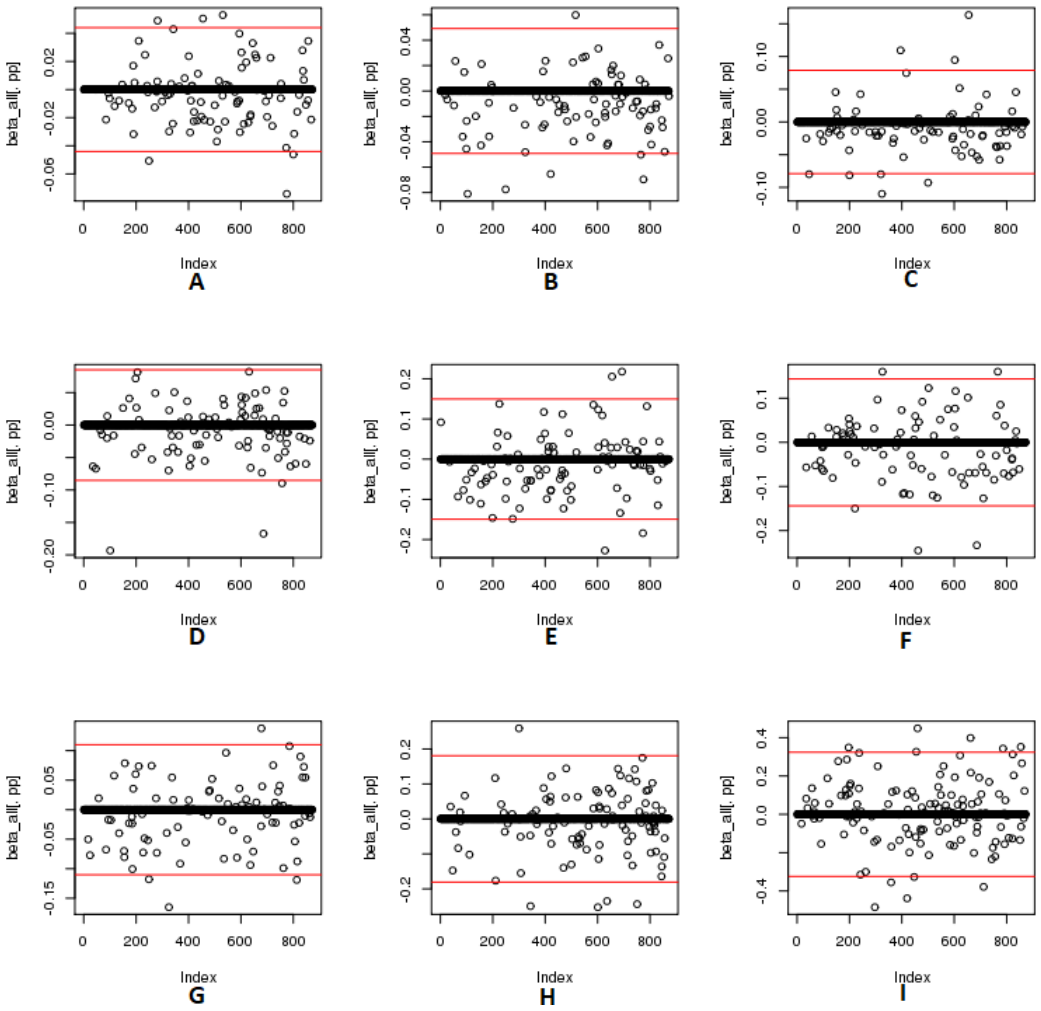



Figure 3.1.6: The coefficient plot for each anti-cancer drug group with 5% significance level bounds (red lines).

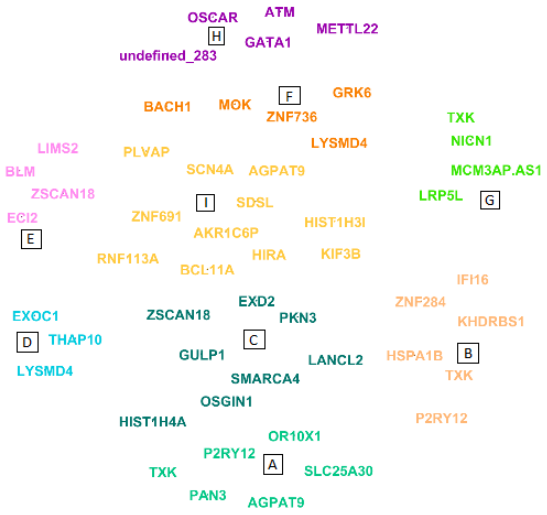



Figure 3.1.7: The exhibition of significant genes related to specific anti-cancer drug groups from the screening results in Figure 3.1.6.

numbers of drugs for these 5 folds are $J_v = (45, 45, 45, 46, 46)$. Each fold plays as validation data in turn, which means the training and validation are performed 5 times in total to reach an average result. Every time, we calculate the log-likelihood with the formula below for the validation data using the estimates from training.

$$l_v\left(\hat{\Phi} \mid \mathbf{Y}_v\right) = \sum_{j=1}^{J_v} \log\left\{\sum_{k=1}^{\hat{K}} \frac{\hat{\pi}_k}{\left(2\pi\hat{\sigma}_k^2\right)^{n/2}} \exp\left[-\frac{\left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\right)^T\left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\hat{\boldsymbol{\beta}}_k\right)}{2\hat{\sigma}_k^2}\right]\right\},$$

where $J_v$ is the number of validation drugs, $\mathbf{y}_j$ is the observations of validation drugs from validation matrix $\mathbf{Y}_v$. These estimates $\hat{K}, \hat{\pi}_k, \hat{\sigma}_k, \hat{\boldsymbol{\beta}}_k$ are from the estimation with training data $\mathbf{Y}_{-v}$. Here $J_{-v} = J - J_v$.

*mclust* (Fraley and Raftery, 2002; Scrucca et al., 2016) is a popular R package for model-based clustering, classification, and density estimation based on finite normal mixture models.

To compare with *mclust*, we input training data $\mathbf{Y}_{-v}$ into R package *mclust* to get new estimates and calculate the validation data log-likelihood with:

$$l_v^\star\left(\hat{\Phi}^\star \mid \mathbf{Y}_v\right) = \sum_{j=1}^{J_v} \log\left\{\sum_{k=1}^{\hat{K}^\star} \frac{\hat{\pi}_k^*}{\sqrt{(2\pi)^n\left|\hat{\boldsymbol{\Sigma}}^\star\right|}} \exp\left[-\frac{1}{2}\left(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_k^\star\right)^T \hat{\boldsymbol{\Sigma}}^{\star-1}\left(\mathbf{y}_j - \hat{\boldsymbol{\mu}}_k^\star\right)\right]\right\}.$$

where $\left|\hat{\boldsymbol{\Sigma}}^\star\right| = \det \hat{\boldsymbol{\Sigma}}^\star$ is the determinant of $\hat{\boldsymbol{\Sigma}}^\star$. These new estimates $\hat{K}^\star, \hat{\pi}_k^\star, \hat{\boldsymbol{\Sigma}}^\star, \hat{\boldsymbol{\mu}}_k^\star$ are from the outputs of the function *mclust* with training data $\mathbf{Y}_{-v}$.

The results are exhibited in Figure 3.1.8. Our estimates can always yield higher likelihood

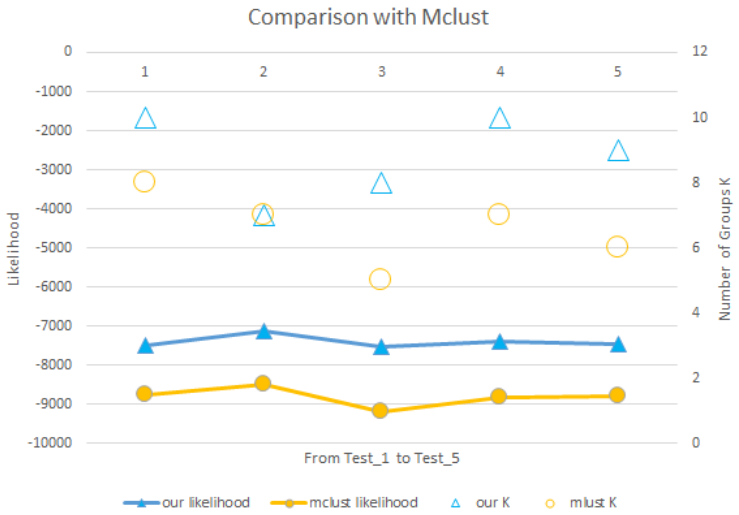



Figure 3.1.8: From test 1 to test 5, the likelihood values and numbers of groups for the results of our penalised likelihood fusion modelling and *mclust* respectively. The line graphs are the likelihood values $l_v\left(\hat{\Phi} \mid \mathbf{Y}_v\right)$ and $l_v^{\star}\left(\hat{\Phi}^{\star} \mid \mathbf{Y}_v\right)$. Blue one is from our modelling, yellow one is from *mclust*. The triangle (our modelling) and circle (*mclust*) markers show the comparison of group numbers.

values for validation data, and anti-cancer drugs are clustered into more groups compared with the outcomes from *mclust*.

### 3.1.8 Conclusions

- Anti-cancer drugs manifest a group structure and are relevant to a small number of genes out of all body genes. The selected genes play a key role in the functioning of anti-cancer drugs.
- The comparison with R package *mclust* can indicate that gene expression data can contribute to drug clustering work, since this additional information can help yield a more detailed clustering outcome. The classification for new drugs also becomes more accurate regarding likelihood values.
- The performance in clustering indicates that our penalised likelihood fusion model and dimension-reduction methods applied above are effective in both screening out the relevant genes and clustering the drugs with drug sensitivity and gene expression data.

## 3.2 Incomplete Case with Missingness

When studying with high-dimensional data problems, missing data problem always becomes an annoying barrier which hinders the use of full given data. Since the complete-case analysis (Little and Rubin, 2014) leaves out all the individuals which have missing entries, it will result

in a high chance of the deletion of useful information. As the illustration from Zhu, Wang and Samworth (2019), when a $n \times J$ data matrix has an independent probability of only 0.01 for each entry that there exists missingness, it can lead to a huge difference as $n$ increases. When $n = 5$, there will be around 95% of individuals (out of all $J$ individuals) having complete data. But when $n$ increases to 300, there will be only 5% of all $J$ individuals remaining complete (no missingness). In our analysis, we try to use all given data and perform a multivariate mixture regression with the LASSO penalty and $l_q-$norm penalty to realise clustering and variable selection simultaneously. Our data come from anti-cancer drug studies (Iorio et al., 2016) and consist of the logarithm of half maximal inhibitory concentration (drug sensitivity) and RMA process basal expression (gene expression). There is missingness in the drug data. Then it becomes necessary for us to introduce the new model from this part to perform a more complete analysis by not constraining to the rectangle shape of a matrix. We keep nearly all columns by just removing the useless rows from drug data. Thus, more data can participate in the analysis without compromising on the missing patterns. In this chapter, we will also compare the penalty $q = 1$ (LASSO) and the penalty $q = 0.5$ ($l_q$) to reveal the estimation accuracy and sparsity attribute within them, after which we will combine these penalties to propose an original methodology for better solving clustering and variable selection problems.

As introduced in Section 1.2, our research data have two parts $\mathbf{Y}_{n\times J}$ and $\mathbf{X}_{n\times p}$. There are missing entries (denoted as NA) in this matrix $\mathbf{Y}_{n\times J}$. There are no missing entries in matrix $\mathbf{X}_{n\times p}$. These two data matrices are connected by sharing the same cell lines, thus, $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times J}$ have the same row length. As the complete case problem introduced in previous section, we use our likelihood fusion model again, but different from Section 3.1.4, for response matrix $\mathbf{Y}_{n\times J}$ this time we don't need to do any column or row deletions to make it a complete-case matrix. We use subscripts to avoid the missing entries in each column, and then most columns with small proportions of missingness can be retained for our numerical analysis. Thus, the possibility of the deletion of useful information could be greatly reduced.

For the response matrix $\mathbf{Y}_{n\times J}$, each column is a vector, such as the $j$th column $\mathbf{y}_j = (y_{1j}, \ldots, \text{NA}, \ldots, y_{nj})^T$. After avoiding all missing entries in it, we obtain a new $n_{(j)} \times 1$ vector denoted as $\mathbf{y}_{(j)} = \left(y_{(j)1}, \ldots, y_{(j)n_{(j)}}\right)^T$ where $n_{(j)}$ is the cardinal number of $j$th column without any missing entries. The conditional density of $j$th drug, under the condition that this drug

belongs to $k$th component following $\left(\mathbf{X}_{(j)}\boldsymbol{\beta}_k, \sigma_k^2\right)$, can be written as

$$\begin{aligned} f_k\left(\mathbf{y}_{(j)} \mid \alpha_j, \mathbf{X}_{(j)}\boldsymbol{\beta}_k, \sigma_k^2\right) &= \prod_{i=1}^{n_{(j)}} f_k^i\left(y_{(j)i} \mid \alpha_j, \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k, \sigma_k^2\right) \\ &= \prod_{i=1}^{n_{(j)}} \frac{1}{\left(2\pi\sigma_k^2\right)^{1/2}} \exp\left\{-\frac{\left(y_{(j)i} - \alpha_j - \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2}\right\} \\ &= \frac{1}{\left(2\pi\sigma_k^2\right)^{n_{(j)}/2}} \\ &\quad \times \exp\left\{-\frac{\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right\}, \end{aligned} \tag{3.2.1}$$

where $\beta_k$ is the coefficient vector and $\sigma_k^2$ is the variance for the $k$th component, and $\alpha_j$ is the intercept.

Here $\mathbf{X}_{(j)}$ is a sub-matrix of the design matrix $\mathbf{X}_{n\times p}$, and $\mathbf{X}_{(j)}$ shares the same rows as vector $\mathbf{y}_{(j)}$, so matrix $\mathbf{X}_{(j)}$ has a size $n_{(j)} \times p$. And $\mathbf{x}_{(j)i}$ is the $i$th row vector of $\mathbf{X}_{(j)}$, then $\alpha_j$ is a tuning constant which will be determined in our next estimation. If the $j$-th drug has a probability of $\pi_k$ that it belongs to the $k$-th component, then this likelihood fusion model is constructed as

$$f\left(\mathbf{y}_{(j)} \mid \mathbf{X}_{(j)}, \boldsymbol{\Phi}_{(j)}\right) = \sum_{k=1}^{K} \pi_k f_k\left(\mathbf{y}_{(j)} \mid \alpha_j, \mathbf{X}_{(j)}\boldsymbol{\beta}_k, \sigma_k^2\right).$$

For convenience, we denote all parameters needed to estimate as $\boldsymbol{\Phi}_{(j)}$, such as $\alpha_j, \boldsymbol{\beta}, \boldsymbol{\pi}$ and $\boldsymbol{\sigma}$.

It is the same as Section 3.1.4 where we impose a penalty term upon the likelihood fusion model to give it the variable screening ability. But this time a $l_q-$norm penalty is introduced instead of the previous LASSO ($l_1-$norm) penalty term, which can bring in more flexibility for our analysis. Hence we have the following fused likelihood:

$$\begin{aligned} {}_{pL_{inc}}(\boldsymbol{\Phi}) = &\underbrace{\prod_{j=1}^{J}}_{\text{All Drugs}} \underbrace{\sum_{k=1}^{K} \frac{\pi_k}{\left(2\pi\sigma_k^2\right)^{n_{(j)}/2}} \exp\left\{-\frac{\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right\}}_{\text{Likelihood Fusion}} \\ &\underbrace{\times \exp\left(-\frac{n_{(j)}\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right)\frac{1}{\sigma_k^2}}_{\text{Penalty Term}}, \end{aligned} \tag{3.2.2}$$

where $\boldsymbol{\Phi}$ represents $\left(\boldsymbol{\Phi}_{(1)}, \boldsymbol{\Phi}_{(2)}, \ldots, \boldsymbol{\Phi}_{(J)}\right)$.

It comes from the penalisation onto each observation of each drug, which can be decomposed

into

$$\begin{aligned}
&\prod_{i=1}^{n_{(j)}} f_k^i\left(y_{(j)i} \mid \alpha_j, \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k, \sigma_k^2\right) \times \exp\left(-\frac{\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n_{(j)}}} \\
= &\prod_{i=1}^{n_{(j)}} \frac{1}{\left(2\pi\sigma_k^2\right)^{1/2}} \exp\left\{-\frac{\left(y_{(j)i} - \alpha_j - \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2}\right\} \times \exp\left(-\frac{\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n_{(j)}}} \\
= &\frac{1}{\left(2\pi\sigma_k^2\right)^{n_{(j)}/2}} \exp\left\{-\frac{\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right\} \\
&\times \exp\left(-\frac{n_{(j)}\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^2}.
\end{aligned}$$

From next section, we will use Expectation-Maximisation algorithm to solve the optimisation problem with interlocking equations under the circumstances of $q = 1$ and $q = 0.5$ respectively.

### 3.2.1 Algorithm and Methodology

To find the maximum likelihood estimates of the parameters in a penalised likelihood function which is a mixture of multivariate normal density, we resort to the EM algorithm. (3.2.2) is an **incomplete-data** likelihood function short of group label $z$ for each drug. After imposing hidden (latent) variables $z$ to indicate the affiliation with groups for each drug, we enhance the **incomplete-data** likelihood function to a **complete-data** likelihood function

$$\begin{aligned}
pL_{com}(\boldsymbol{\Phi}) = &\prod_{j=1}^{J}\prod_{k=1}^{K}\left[\pi_k f_k\left(\mathbf{y}_{(j)} \mid \alpha_j, \mathbf{X}_{(j)}\boldsymbol{\beta}_k, \sigma_k^2\right) \times \exp\left(-\frac{\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n_{(j)}}}\right]^{z_{jk}} \\
= &\prod_{j=1}^{J}\prod_{k=1}^{K}\left\{\pi_k \prod_{i=1}^{n_{(j)}} \frac{1}{\left(2\pi\sigma_k^2\right)^{1/2}} \exp\left[-\frac{\left(y_{(j)i} - \alpha_j - \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k\right)^2}{2\sigma_k^2}\right]\right. \\
&\left.\times \exp\left(-\frac{\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^{2/n_{(j)}}}\right\}^{z_{jk}} \\
= &\prod_{j=1}^{J}\prod_{k=1}^{K}\left\{\frac{\pi_k}{\left(2\pi\sigma_k^2\right)^{n_{(j)}/2}} \exp\left[-\frac{\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)} - \alpha_j\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}\right]\right. \\
&\left.\times \exp\left(-\frac{n_{(j)}\lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}\right) \frac{1}{\sigma_k^2}\right\}^{z_{jk}},
\end{aligned}$$

where $z_{jk} = 1$ when $j$th drug belongs to group $k$, otherwise $z_{jk} = 0$. For the convenience of calculation, we use the **log-likelihood** function:

$$\begin{aligned} pl_{com}(\mathbf{\Phi}) &= \log\left(pL_{com}(\mathbf{\Phi})\right) \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log\left(\pi_k\right) - \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\log\left((2\pi)^{n_{(j)}/2}\sigma_k^{n_{(j)}+2}\right) \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\frac{\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2} \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K} z_{jk}\cdot\frac{n_{(j)}\lambda_q\left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}. \end{aligned} \tag{3.2.3}$$

**E-step**

Under the current information (the $v$-th parameter estimates) and the given data, from the Bayes theorem, the hidden variable indicator $z_{jk}$ can have a proportional expectation with weights $\pi_k^{(v)}$:

$$\begin{aligned} \tau_{jk}^{(v)} &= \Pr(z_{jk}=1 \mid \mathbf{y}_{(j)}, \mathbf{X}_{(j)}, \mathbf{\Phi}_{(j)}^{(v)}) \\ &= \mathrm{E}(z_{jk} \mid \mathbf{y}_{(j)}, \mathbf{X}_{(j)}, \mathbf{\Phi}_{(j)}^{(v)}) \\ &= \frac{\pi_k^{(v)} f_k(\mathbf{y}_{(j)} \mid \alpha_j, \mathbf{X}_{(j)}\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2})}{\sum_{k=1}^{K}\pi_k^{(v)} f_k(\mathbf{y}_{(j)} \mid \alpha_j, \mathbf{X}_{(j)}\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2})} \\ &= \frac{\frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n_{(j)}/2}}\exp\{-\Delta\}\times\exp\left(-\frac{n_{(j)}\lambda_q\left\|\boldsymbol{\beta}_k^{(v)}\right\|_q^q}{\sigma_k^{(v)}}\right)\frac{1}{\sigma_k^{(v)2}}}{\sum_{k=1}^{K}\frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n_{(j)}/2}}\exp\{-\Delta\}\times\exp\left(-\frac{n_{(j)}\lambda_q\left\|\boldsymbol{\beta}_k^{(v)}\right\|_q^q}{\sigma_k^{(v)}}\right)\frac{1}{\sigma_k^{(v)2}}}, \\ &\quad \left(\Delta = \frac{\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k^{(v)}\right)^T\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k^{(v)}\right)}{2\sigma_k^{(v)2}}\right). \end{aligned}$$

After replacing $z_{jk}$ in (3.2.3) with this $\tau_{jk}^{(v)}$, we reach the conditional expectation function which needs to be iteratively maximised and updated:

$$\begin{aligned} Q\left(\mathbf{\Phi} \mid \mathbf{\Phi}^{(v)}\right) &= E_{\mathbf{Z}|\mathbf{Y},\mathbf{X},\mathbf{\Phi}^{(v)}}\left[\log\left(pL_{com}(\mathbf{\Phi})\right)\right] \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left(\pi_k\right) - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left((2\pi)^{n_{(j)}/2}\sigma_k^{n_{(j)}+2}\right) \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\frac{\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_{(j)}-\alpha_j\mathbf{1}-\mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2} \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\frac{n_{(j)}\lambda_q\left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k}. \end{aligned}$$

**M-step**

To update the estimate $\mathbf{\Phi}$, we maximise $Q(\mathbf{\Phi} \mid \mathbf{\Phi}^{(v)})$ with respect to $\mathbf{\Phi}$ block by block. Starting from $\alpha_j$, we set

$$\begin{aligned}
\frac{\partial Q}{\partial \alpha_j} &= \frac{\partial \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\tau_{jk}^{(v)} \cdot \left(\mathbf{y}_{(j)} - \alpha_j \mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_{(j)} - \alpha_j \mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}}{\partial \alpha_j} \\
&= -\sum_{k=1}^{K}\left[\frac{\tau_{jk}^{(v)}}{\sigma_k^2}\sum_{i=1}^{n_{(j)}}\left(y_{(j)i} - \alpha_j - \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k\right)\right] \\
&= -\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}}{\sigma_k^2}\left(\sum_{i=1}^{n_{(j)}} y_{(j)i} - n_{(j)}\alpha_j - \sum_{i=1}^{n_{(j)}}\mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k\right) \\
&= 0.
\end{aligned}$$

Here if we standardise $\mathbf{X}_{(j)}$ by column-wise before the EM algorithm, which means $\sum_{w=1}^{n_{(j)}} \mathbf{x}_{(j)wi} = 0$, then we can have $\sum_{i=1}^{n_{(j)}} \mathbf{x}_{(j)i}^T\boldsymbol{\beta}_k = 0$, hence $\sum_{i=1}^{n_{(j)}} y_{(j)i} - n_{(j)}\alpha_j = 0$. Then we can reach

$$\alpha_j = \frac{\sum_{i=1}^{n_{(j)}} y_{(j)i}}{n_{(j)}} = \bar{y}_{(j)}.$$

It suggests that $\alpha_j$ is the average of elements in $\mathbf{y}_{(j)}$. Therefore, the calculation $\mathbf{y}_{(j)} - \alpha_j\mathbf{1}$ is a centralisation operation on each column of $\mathbf{Y}$, and $\alpha_j$ just depends on the original data $\mathbf{Y}$ which is independent of other parameters such as $\tau_{jk}, \sigma_k, \boldsymbol{\beta}_k, \pi_k, \lambda_q$. Accordingly, we don't need to update $\alpha_j$ in the iteration of the EM algorithm. We replace it with $\bar{y}_{(j)}$ in our next steps.

To update $\pi_k$ with the constraint $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k \geq 0$, we turn to the Lagrange multiplier for help. The corresponding Lagrangian function is

$$L(\pi_k, \Lambda) = \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)} \cdot \log(\pi_k) - \Lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).$$

To maximise this function, we differentiate it with respect to $\pi_k$ and make it equal to 0:

$$\frac{\partial L(\pi_k, \Lambda)}{\partial \pi_k} = \sum_{j=1}^{J}\frac{\tau_{jk}^{(v)}}{\pi_k} - \Lambda = 0,$$

after rearrangement, it gives

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{j=1}^{J}\tau_{jk}^{(v)}}{\Lambda}, \tag{3.2.4}$$

then

$$\sum_{k=1}^{K} \hat{\pi}_k^{(v+1)} = \sum_{k=1}^{K} \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)}}{\Lambda} = 1,$$

which gives $\Lambda = \sum_{k=1}^{K} \sum_{j=1}^{J} \tau_{jk}^{(v)}$. From the definition of $\tau_{jk}$, we have $\sum_{k=1}^{K} \tau_{jk}^{(v)} = 1$. Thus $\Lambda = \sum_{j=1}^{J} 1 = J$. So we update (3.2.4) in the $v$-th EM iteration to reach the $(v+1)$-th estimate of $\pi_k$:

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)}}{J}.$$

Afterwards we update $\boldsymbol{\beta}_k$ by

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\beta}_k} &= \frac{\partial \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\tau_{jk}^{(v)} \cdot \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}}{\partial \boldsymbol{\beta}_k} \\ &\quad + \frac{\partial \sum_{j=1}^{J} \sum_{k=1}^{K} \tau_{jk}^{(v)} \cdot \frac{n_{(j)} \lambda_q \|\boldsymbol{\beta}_k\|_q^q}{\sigma_k}}{\partial \boldsymbol{\beta}_k} \\ &= 0. \end{aligned} \tag{3.2.5}$$

To solve the equation (3.2.5), it suffices to solve the following minimisation problem:

$$\hat{\boldsymbol{\beta}}_k^{(v+1)} = \arg\min_{\boldsymbol{\beta}_k} \left\{ \frac{1}{2} \|\mathbf{Y}^* - \mathbf{X}^* \boldsymbol{\beta}_k\|_2^2 + \lambda_q \sum_{j=1}^{J} \left( \tau_{jk}^{(v)} \cdot n_{(j)} \right) \cdot \|\boldsymbol{\beta}_k\|_q^q \right\}, \tag{3.2.6}$$

where

$$\mathbf{X}^* = \operatorname{diag}\left( \frac{\tau_{1k}^{(v)}}{\sigma_k}, \frac{\tau_{2k}^{(v)}}{\sigma_k}, \ldots, \frac{\tau_{Jk}^{(v)}}{\sigma_k} \right)^{1/2} \begin{pmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \\ \vdots \\ \mathbf{X}_{(J)} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}} \cdot \mathbf{X}_{(1)} \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}} \cdot \mathbf{X}_{(2)} \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}} \cdot \mathbf{X}_{(J)} \end{pmatrix},$$

$$\mathbf{Y}^* = \operatorname{diag}\left( \frac{\tau_{1k}^{(v)}}{\sigma_k}, \frac{\tau_{2k}^{(v)}}{\sigma_k}, \ldots, \frac{\tau_{Jk}^{(v)}}{\sigma_k} \right)^{1/2} \begin{pmatrix} \mathbf{y}_{(1)} - \bar{y}_{(1)}\mathbf{1} \\ \mathbf{y}_{(2)} - \bar{y}_{(2)}\mathbf{1} \\ \vdots \\ \mathbf{y}_{(J)} - \bar{y}_{(J)}\mathbf{1} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}} \cdot \left(\mathbf{y}_{(1)} - \bar{y}_{(1)}\mathbf{1}\right) \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}} \cdot \left(\mathbf{y}_{(2)} - \bar{y}_{(2)}\mathbf{1}\right) \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}} \cdot \left(\mathbf{y}_{(J)} - \bar{y}_{(J)}\mathbf{1}\right) \end{pmatrix},$$

and $\mathbf{X}^*$ is a $\sum_{j=1}^{J} n_{(j)} \times p$ matrix, $\mathbf{Y}^*$ is a $\sum_{j=1}^{J} n_{(j)} \times 1$ vector.

**q=1 Case**

When $q = 1$, (3.2.6) is a LASSO problem with $\sum_{j=1}^{J} n_{(j)}$ observations and $p$ covariates. Generally we can use the coordinate descent method to solve this LASSO problem. The detailed calculation is attached in **Appendix A**. This optimisation is implemented by using the *glmnet* package (Friedman et al., 2010), which makes use of warm starts and active-set convergence to save much time.

**q=0.5 Case**

When $q = 0.5$, (3.2.6) is a nonconvex problem, and $\|\boldsymbol{\beta}_k\|_{0.5}^{0.5} = \sum_{i=1}^{p} \sqrt{|\beta_{ki}|}$ is a $\mathcal{L}_q$ quasi-norm, $0 < q < 1$. To solve this tricky problem, we use the IRucLq-v algorithm (Lai et al., 2013). Here, to make it differentiable, a smoothing parameter $\delta_v$ is added to make $\|\boldsymbol{\beta}_k\|_q^q = \sum_{i=1}^{p} \left(\beta_{ki}^2 + \delta_v^2\right)^{q/2}$ where $\delta_v$ starts at a large value and is gradually reduced towards 0. Moreover, $\delta_v$ can also help avoid being trapped into local optimum in the next iterations. From empirical results (Lai et al., 2013), this approximation can yield a very good estimate for $\boldsymbol{\beta}_k$ under $0 < q < 1$ especially $q = 0.5$. Our updating process follows the first-order optimality condition:

$$\frac{\partial \left\{ \frac{1}{2} \|\mathbf{Y}^* - \mathbf{X}^* \boldsymbol{\beta}_k\|_2^2 + \lambda_q \sum_{j=1}^{J} \left(\tau_{jk}^{(v)} \cdot n_{(j)}\right) \cdot \sum_{i=1}^{p} \left(\beta_{ki}^2 + \delta_v^2\right)^{q/2} \right\}}{\partial \boldsymbol{\beta}_k} = 0.$$

Then $\boldsymbol{\beta}_k$ can be updated to $\boldsymbol{\beta}_k^{(v+1)}$ by solving

$$-\mathbf{X}^{*T}\left(\mathbf{Y}^* - \mathbf{X}^* \boldsymbol{\beta}_k^{(v+1)}\right) + \left[ \frac{\lambda_q \cdot q \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)} \cdot \beta_{ki}^{(v+1)}}{\left(\beta_{ki}^{(v)2} + \delta_v^2\right)^{1-q/2}} \right]_{1 \le i \le p} = 0.$$

Rearrange it to reach

$$\mathbf{X}^{*T}\mathbf{Y}^* = \left( \mathbf{X}^{*T}\mathbf{X}^* + \operatorname{diag}\left[ \frac{\lambda_q \cdot q \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)}}{\left(\beta_{ki}^{(v)2} + \delta_v^2\right)^{1-q/2}}, i = 1, 2, \ldots, p \right] \right) \boldsymbol{\beta}_k^{(v+1)}.$$

$\delta_v$ is updated by $\delta_{v+1} = \min\left\{\delta_v, \alpha \cdot r(\boldsymbol{\beta}_k^{(v+1)})_{s+1}\right\}$, where $\alpha \in (0, 1)$ is a step-size constant. $s$ is an estimated sparsity level to guide the stoppage of iterations which is inspired from Daubechies et al. (2010). And $r(\boldsymbol{\beta}_k^{(v+1)})_{s+1}$ is the $(s+1)$th largest absolute-value entry of $\boldsymbol{\beta}_k$ after $(v+1)$th iteration. The iteration process will stop until $\delta_{v+1} = 0$ or within a reasonable time, after which the latest $\boldsymbol{\beta}_k^{(v+1)}$ is returned as the output.

Insofar, we have finished the update of $\boldsymbol{\beta}_k$ under the circumstances of $q = 1$ and $q = 0.5$

respectively. From now on, we start updating $\sigma_k$ by solving

$$\frac{\partial Q}{\partial \sigma_k^2} = 0.$$

After rearrangement, it becomes

$$\begin{aligned} 0 = & \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \left(n_{(j)} + 2\right) \cdot \sigma_k^2 - \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)} \lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q \cdot \sigma_k \\ & - \sum_{j=1}^{J} \tau_{jk}^{(v)} \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right). \end{aligned}$$

Solve this quadratic equation with respect to $\sigma_k$, we can reach

$$\hat{\sigma}_k^{(v+1)} = \frac{\mathcal{B} + \sqrt{\mathcal{B}^2 + 4\mathcal{A} \cdot \mathcal{C}}}{2\mathcal{A}},$$

where

$$\begin{aligned} \mathcal{A} = & \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \left(n_{(j)} + 2\right), \\ \mathcal{B} = & \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)} \lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q, \\ \mathcal{C} = & \sum_{j=1}^{J} \tau_{jk}^{(v)} \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right). \end{aligned}$$

**Stopping Rule**

The above algorithm alternates between E-step and M-step until convergence is reached. The stopping rule imposed to confirm the convergence is based on the relative changes in the log-likelihood. The convergence is attained when the following equality holds:

$$\left| \frac{pl_{inc}(\boldsymbol{\Phi}^{(v+1)}) - pl_{inc}(\boldsymbol{\Phi}^{(v)})}{pl_{inc}(\boldsymbol{\Phi}^{(v)})} \right| < \epsilon,$$

where $\epsilon$ is a reasonably small value (generally $10^{-5}$ or $10^{-10}$) and $pl_{inc}(\boldsymbol{\Phi})$ is the penalised incomplete-data likelihood defined by

$$\begin{aligned} pl_{inc}(\boldsymbol{\Phi}) = & \sum_{j=1}^{J} \log \left\{ \sum_{k=1}^{K} \frac{\pi_k}{(2\pi\sigma_k^2)^{n_{(j)}/2}} \right. \\ & \times \exp\left[ -\frac{\left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\boldsymbol{\beta}_k\right)}{2\sigma_k^2} \right] \\ & \left. \times \exp\left( -\frac{n_{(j)} \lambda_q \left\|\boldsymbol{\beta}_k\right\|_q^q}{\sigma_k} \right) \frac{1}{\sigma_k^2} \right\}. \end{aligned}$$

#### 3.2.1.1 Thresholding the Regression Coefficients

After the estimation with $q = 0.5$, due to the limitation of iteration times and floating point issue in computer, the inactive points (zero entries) in $\boldsymbol{\beta}$s are not estimated as exact zeros. Then the use of $\text{BIC}_{\lambda_{0.5},K}$ in next section to select number of components $K$ and tuning constant $\lambda_{0.5}$ will have a problem because the term $\sum_{i=1}^{p}\sum_{k=1}^{K}\mathbb{I}\left(\hat{\beta}_{ki} \neq 0\right)$ in $\text{BIC}_{\lambda_{0.5},K}$ cannot reflect the actual sparsity. Now we need to filter out these noises by setting thresholds for the estimates $\hat{\beta}_{ki}$s.

In empirical studies, we like to use three-sigma rule of thumb $\mu \pm 3\sigma$ to describe the range to contain most values in a normal distribution, where 'three' is used to guarantee that $\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.9973$. Back to our problem, here we treat it as a robust estimation problem, so the median is used as a robust measure of central tendency $\mu$, and median absolute deviation (MAD) is used as a robust measure of statistical dispersion $\sigma$. Then the threshold becomes

$$\hat{\boldsymbol{\beta}}_{ki}^{(new)} = \begin{cases} 0 & \text{median}\left(\hat{\boldsymbol{\beta}}_{ki}\right) - a \times \text{MAD}\left(\hat{\beta}_{ki}\right) \leq \hat{\beta}_{ki} \leq \text{median}\left(\hat{\beta}_{ki}\right) + a \times \text{MAD}\left(\hat{\beta}_{ki}\right), \\ \hat{\beta}_{ki} & \text{otherwise.} \end{cases}$$

$a$ is the accuracy factor to gauge the confidence. Generally $a$ can be calculated by $c \times \Phi^{-1}(1 - 0.001/(2p))$ which is composed of the scale factor $c$ and a quantile function with Bonferroni correction (Dunn, 1961) for multiple tests (as $p$ comparisons are performed). Here significance level is assumed as 0.1%, and it is a two-tailed test. The scale factor $c$ depending on the distribution is to make MAD a consistent estimator when using it as an estimation of the standard deviation $\sigma$ (Rousseeuw and Croux, 1993). To make sure 50% (the interquartile range from 25% to 75%) of the standard normal cumulative distribution function is covered, we set

$$\text{P}(|X - \mu| \leq \text{MAD}) = \text{P}\left(\left|\frac{X-\mu}{\sigma}\right| \leq \frac{\text{MAD}}{\sigma}\right) = \text{P}\left(|Z| \leq \frac{\text{MAD}}{\sigma}\right) = 50\%,$$

then it is

$$\Phi\left(\frac{\text{MAD}}{\sigma}\right) - \Phi\left(-\frac{\text{MAD}}{\sigma}\right) = 0.5.$$

Therefore, $\text{MAD}/\sigma = \Phi^{-1}(0.75) = 0.67449 = 1/c$, that is $c = 1.4826$. So in normal distributed data we set $\hat{\sigma} = c \cdot \text{MAD} = 1.4826 \cdot \text{MAD}$.

In the simulation section, under $\boldsymbol{X}_{50\times500}$ and $\boldsymbol{Y}_{50\times20}$, $a = c \times \Phi^{-1}(1 - 0.001/(2p)) = 1.4826 \times \Phi^{-1}(1 - 0.001/(2 \times 500)) = 7.047427$. So for convenience we apply $\text{median}\left(\hat{\beta}_{ki}\right) \pm 7 \times \text{MAD}\left(\hat{\beta}_{ki}\right)$ to the simulation as the threshold for filtering out noises.

#### 3.2.1.2 The Choices of $\lambda_q$ and $K$

To find the best number of components $K$ and penalty tuning constant $\lambda_q$, we decide to use the BIC value (Schwarz, 1978; see Section 2.6.1). First, construct a grid of $\lambda_q$ and $K$ respectively. Then implement the aforementioned EM algorithm with each pair of $\lambda_q$ and $K$, calculate the $\text{BIC}_{\lambda_q,K}$ each time with estimated parameters $\hat{\mathbf{\Phi}}$ and given data $\mathbf{X}, \mathbf{Y}$ by

$$\text{BIC}_{\lambda_q,K} = \left(K - 1 + K + \sum_{i=1}^{p}\sum_{k=1}^{K} \mathbb{I}(\hat{\beta}_{ki}^{(new)} \neq 0)\right) \times \frac{\log\left(\sum_{j=1}^{J} n_{(j)}\right)}{J} - \frac{2}{J} \times l(\hat{\mathbf{\Phi}}),$$

where

$$\begin{aligned} l(\hat{\mathbf{\Phi}}) = & \sum_{j=1}^{J} \log\left\{\sum_{k=1}^{K} \frac{\hat{\pi}_k}{(2\pi\hat{\sigma}_k^2)^{n_{(j)}/2}}\right. \\ & \left. \times \exp\left[-\frac{\left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\hat{\boldsymbol{\beta}}_k^{(new)}\right)^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\mathbf{1} - \mathbf{X}_{(j)}\hat{\boldsymbol{\beta}}_k^{(new)}\right)}{2\hat{\sigma}_k^2}\right]\right\}. \end{aligned}$$

$K - 1$ represents the number of parameters estimated in $\boldsymbol{\pi}_{1:K} = (\pi_1, \pi_2 \ldots, \pi_K)$ ($\sum_{k=1}^{K} \pi_k = 1$), and $K$ represents the number of parameters estimated in $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_K)$. Here $\sum_{i=1}^{p}\sum_{k=1}^{K} \mathbb{I}(\hat{\beta}_{ki}^{(new)} \neq 0)$ uses the indicator function to count how many coefficients are non-zero. $\sum_{j=1}^{J} n_{(j)}$ is the number of observations. In the experiments, this $\lambda_q-$and$-K$ grid needs to be amended to make the smallest $\text{BIC}_{\lambda_q,K}$ locate in the middle of this grid, then the smallest BIC value indicates the best $\lambda_q$ and $K$.

#### 3.2.1.3 Combine $q = 1$ Case and $q = 0.5$ Case

As far as we know, $q = 1$ case will yield a decent estimation with a much faster speed than $q = 0.5$ case. This may owe to the high-speed computing ability of the *glmnet* package. But $q = 0.5$ case can give a more accurate estimation concerning the variable selection and estimation abilities on $\boldsymbol{\beta}$, see Section 3.2.3.3. Meanwhile, $q = 1$ case and $q = 0.5$ case both demonstrate strong ability in clustering. Therefore, $q = 1$ case can provide a good warm start of $\boldsymbol{\pi}_{1:K}, \boldsymbol{\beta}, \boldsymbol{\sigma}$ and decent estimates of $K$ and $s$ for $q = 0.5$ case. After that, $q = 0.5$ case is performed to reach a more accurate result and much time is saved. Moreover, $\lambda_1$ can guide our selection process of $\lambda_{0.5}$ when drawing the $\lambda_{0.5}-$and$-\hat{K}$ grid for $q = 0.5$ case.

### 3.2.2 Initialisation

Here we propose a good starting point for our EM algorithm when starting the esimation for $q = 1$ case. Firstly, without prior information, we consider the group size equally. So for any $k \in (1, 2, \ldots, K)$ the initial value of $\pi_k^{(0)}$ is set as $\frac{1}{K}$, which means $j$th drug has a probability $\frac{1}{K}$ coming from group $k$ (also a probability $\frac{1}{K}$ from any other group). Then, with the linear assumption we have $\mathbf{y}_{(j)} = \mathbf{X}_{(j)}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k$ where the error term $\boldsymbol{\varepsilon}_k$ follows a multivariate normal distribution as $\boldsymbol{\varepsilon}_k \sim \mathcal{N}_{n_{(j)}}\left(\mathbf{0}, \sigma_k^2 \mathbf{I}_{n_{(j)}}\right)$. Here $\mathbf{0}$ is a $n_{(j)}-$size zero vector, and $\mathbf{I}_{n_{(j)}}$ is a $n_{(j)}-$dimensional identity matrix. $\sigma_k^2 \mathbf{I}_{n_{(j)}}$ makes up a covariance matrix. Secondly, we aim to reach a sparse estimate of $\boldsymbol{\beta}_k$ by the LASSO penalty term. Thus, most entries in $\boldsymbol{\beta}_k$ will be estimated as 0 in the end. We set $\boldsymbol{\beta}_k^{(0)}$ as a $p-$size zero vector $\mathbf{0}$ which can be updated to a sparse vector easily. From that, we can get $\mathbf{y}_{(j)} = \mathbf{X}_{(j)}\boldsymbol{\beta}_k^{(0)} + \boldsymbol{\varepsilon}_k^{(0)} = \boldsymbol{\varepsilon}_k^{(0)}$, then we can use the standard deviation of $\mathbf{y}_{(j)}$ as a reference for the initial value $\sigma_k^{(0)}$. We calculate the variances of $\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(J)}$ to find the range of their variances to construct a grid for $\sigma_1^{(0)}, \sigma_2^{(0)}, \ldots, \sigma_K^{(0)}$. We denote the maximum and minimum of these standard deviations by $M$ and $m$, then we let $\sigma_1^{(0)} = m$, $\sigma_2^{(0)} = m + \frac{M-m}{K-1}$, $\sigma_3^{(0)} = m + 2 \cdot \frac{M-m}{K-1}, \ldots, \sigma_K^{(0)} = m + (K-1) \cdot \frac{M-m}{K-1} = M$, from this setting we have a rough coverage for the value of $\sigma_k^{(0)}$. Our initialisation for $q = 1$ case is finished here. As mentioned in last section, $q = 0.5$ case will have a good starting point (warm start) inherited from the result of $q = 1$ case.

### 3.2.3 Simulation Studies

#### 3.2.3.1 Selection Accuracy Evaluation

Before numerical simulation studies, we define several measures to evaluate our estimation performances. First of all, we introduce criteria – sensitivity and specificity, to quantify the variable selection accuracy of our method. Sensitivity is the effective selection rate of true active (non-zero) variables. And specificity is the effective selection rate of true zero variables. When one variable has a non-zero coefficient, we regard it as active. When the corresponding coefficient of one variable is zero, we regard it as a non-active or zero variable. In our simulation settings, we use $V_k$ and $\hat{V}_k$ to denote the locations of true underlying and estimated active variables respectively in group $k$. $V_k^c$ and $\hat{V}_k^c$ are the complements representing the locations of true and estimated zero variables respectively in group $k$. Then sensitivity and specificity of group $k$ can be expressed as the ratio of set cardinality (size):

$$\mathrm{SEN}_k = \frac{\left|V_k \cap \hat{V}_k\right|}{|V_k|}, \qquad \mathrm{SPE}_k = \frac{\left|V_k^c \cap \hat{V}_k^c\right|}{|V_k^c|}.$$

Obviously $V_k \cup V_k^c = \hat{V}_k \cup \hat{V}_k^c = \{1, 2, \ldots, p\}$ representing all the locations of the coefficients in vector $\boldsymbol{\beta}_k$. For the reason of sparsity, SPE is often close to 1. The overall sensitivity and specificity of one estimation are expressed as

$$\text{SEN} = \sum_{k=1}^{K} \pi_k \cdot \frac{\left|V_k \cap \hat{V}_k\right|}{|V_k|}, \qquad \text{SPE} = \sum_{k=1}^{K} \pi_k \cdot \frac{\left|V_k^c \cap \hat{V}_k^c\right|}{|V_k^c|}.$$

They are the weighted summations of sensitivity and specificity for each group. The simulation performs well when sensitivity and specificity are both of high values.

#### 3.2.3.2 Simulations on $n$, $p$, $J$ for $q = 1$ Case

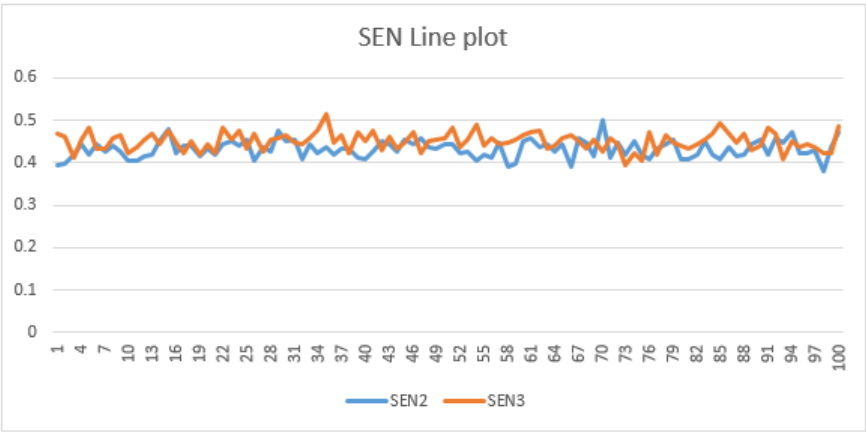
$|V|$ is the cardinal (size) of true active variables. And $\left|\hat{V}\right| \leq n$, it means the cardinal of estimated active variable is always less than or equal to $n$. In our dataset, $p$ is larger than $n$, and when compared with $n$, the larger $p$ is, the larger $\lambda_1$ is, to penalise on the variable selection process. When $n$ is close to $p$, the specificity (SPE) will not be so close to 1 anymore. In general, when $p \gg |V| + n$, the specificity (SPE) is close to 1. When the number of drugs '$J$' increases, the number of groups '$K$' is likely to increase since more anti-cancer drugs participate in our clustering. In this section, we will generate datasets to simulate the changes of $n$, $p$ or $J$ to illustrate the effects on our estimations.

Here we discuss the performances of variable selection (for $q = 1$ case) by changing $n$, $p$ or $J$ respectively under the same parameters $\boldsymbol{\sigma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_{1:K}$. Firstly, start our simulation by using a small part of the real data as $\mathbf{X}_{n\times p}^{(real)}$ and $\mathbf{Y}_{n\times J}^{(real)}$ where $n = 200, p = 1000$ and $J = 50$. Perform our penalised regression to obtain estimates $\hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}_{1:\hat{K}}$ and group amount $\hat{K}$. Every time under the probability of $\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{\hat{K}}$, we generate $j$th column of response matrix $\mathbf{Y}_{n\times J}^{(sim)}$ with real design matrix $\mathbf{X}_{(j)}^{(real)}$ by $\mathbf{y}_{(j)} = \mathbf{X}_{(j)}^{(real)}\hat{\boldsymbol{\beta}}_k + \boldsymbol{\varepsilon}_{jk}$ where error term $\boldsymbol{\varepsilon}_{jk}$ is sampled from multivariate normal distribution $N_{n_{(j)}}(\mathbf{0}, \hat{\sigma}_k^2\mathbf{I}_{n_{(j)}})$, and $\mathbf{I}_{n_{(j)}}$ is identity matrix to make $\hat{\sigma}_k^2\mathbf{I}_{n_{(j)}}$ a diagonal matrix. As mentioned before, $\mathbf{y}_{(j)}$ is the $j$th column of matrix $\mathbf{Y}_{n\times J}^{(sim)}$ after avoiding missing entries in it by using subscripts. $\mathbf{y}_{(j)}$ has a column length $n_{(j)}$. And $\mathbf{X}_{(j)}^{(real)}$ is the sub-matrix of $\mathbf{X}_{n\times p}^{(real)}$ with the same cell lines as vector $\mathbf{y}_{(j)}$. And $\mathbf{X}_{(j)}^{(real)}$ has been standardised by column-wise (each column) before generating $\mathbf{y}_{(j)}$.

**Increase the Number of Drugs $J$**

We increase the number of drugs $J$ from 50 to 70. After generating $\mathbf{Y}_{n\times J}^{(sim)}$ with $\mathbf{y}_{(j)} = \mathbf{X}_{(j)}^{(real)}\hat{\boldsymbol{\beta}}_k + \boldsymbol{\varepsilon}_{jk}$, we perform penalised regression ($q = 1$) with $\mathbf{X}_{n\times p}^{(real)}$ and $\mathbf{Y}_{n\times J}^{(sim)}$ to compare selection accuracies by calculating SENs and SPEs of the results (only sensitivities are plotted below). Repeat trials with each setting ($J = 50$ and $J = 70$) one hundred times re-

| J=50 | | | J=70 | | |
|---|---|---|---|---|---|
| SEN2 mean | SPE2 mean | SEN2 std error | SEN3 mean | SPE3 mean | SEN3 std error |
| 43.22% | 95.39% | 0.00211 | 45.09% | 95.79% | 0.00214 |



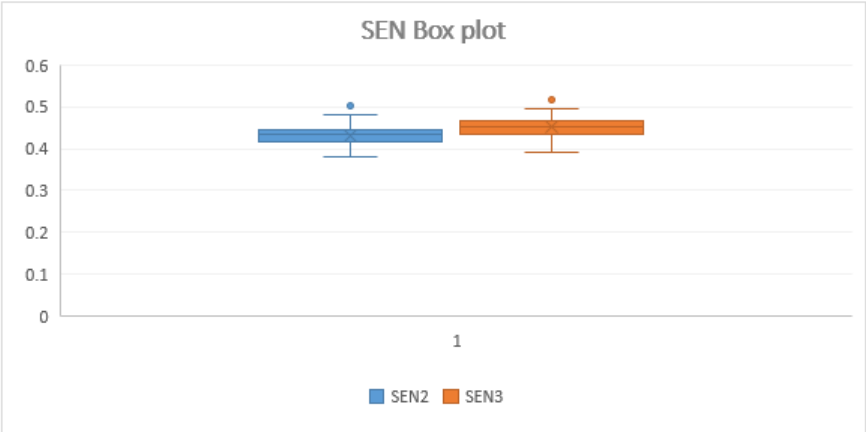



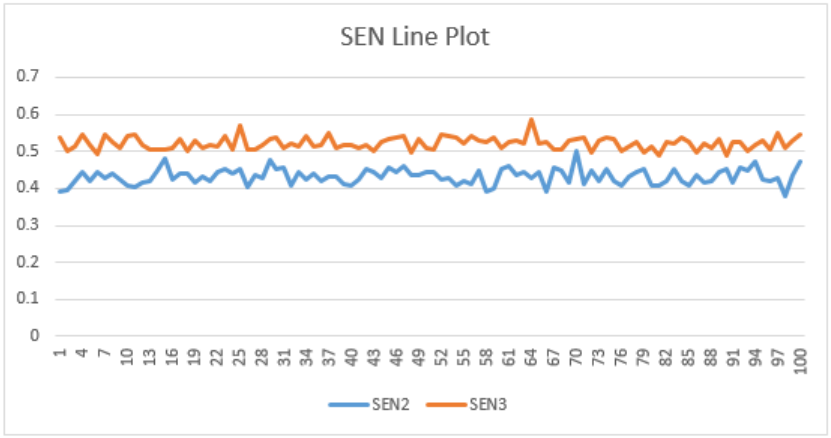

Figure 3.2.1: The sensitivity table, line chart and box plot for 100 trials under the numbers of drugs $J = 50$ (blue) and $J = 70$ (orange).

spectively to remove the randomness of error term $\boldsymbol{\varepsilon}_k$. The results are illustrated in Figure 3.2.1. Seen from the graph, when more drugs are put into our clustering with a definite number of groups $\hat{K}$ and group characteristics ($\boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\pi}_{1:\hat{K}}$), the variable selection accuracy will be improved. This is predictable since more drugs being added to each group will increase group distinction, so it becomes easier to cluster them together.

### Increase the Number of Cell Lines $n$

We increase the number of cell lines $n$ from 200 to 300. The new design matrix $\mathbf{X}^{(real)}_{300\times p}$ is a extension of the old one $\mathbf{X}^{(real)}_{200\times p}$. To extend it, we just need to extract a larger part from the original real data $\mathbf{X}$. The generated response matrix $\mathbf{Y}^{(sim)}_{n\times J}$ has the same missingness positions as real data matrix $\mathbf{Y}^{(real)}_{n\times J}$. After generating $\mathbf{Y}^{(sim)}_{n\times J}$ with $\mathbf{y}_{(j)} = \mathbf{X}^{(real)}_{(j)}\hat{\boldsymbol{\beta}}_k + \boldsymbol{\varepsilon}_{jk}$, again we perform penalised regression with $\mathbf{X}^{(real)}_{n\times p}$ and $\mathbf{Y}^{(sim)}_{n\times J}$ to compare selection accuracies

| n=200 | | | n=300 | | |
|---|---|---|---|---|---|
| SEN2 mean | SPE2 mean | SEN2 std error | SEN3 mean | SPE3 mean | SEN3 std error |
| 43.22% | 95.39% | 0.00211 | 52.19% | 93.16% | 0.00171 |



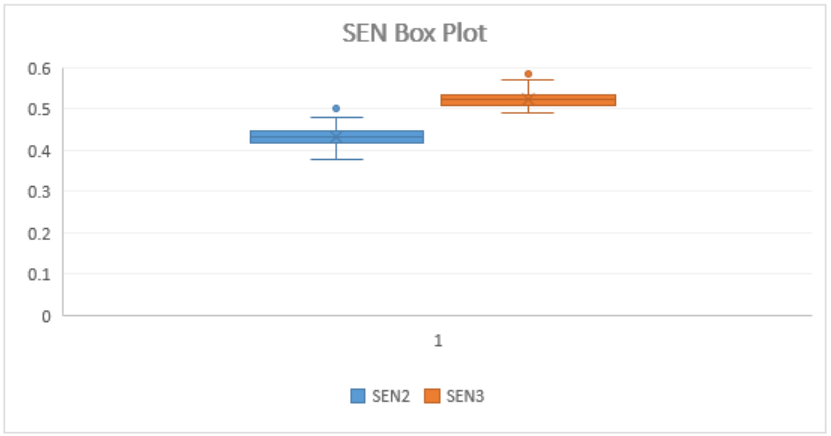




Figure 3.2.2: The sensitivity table, line chart and box plot for 100 trials under the sample sizes $n = 200$ (blue) and $n = 300$ (orange).

by SENs and SPEs of the results (only sensitivities are plotted below). Repeat trials with each setting ($n = 200$ and $n = 300$) one hundred times respectively to remove the randomness of error term $\boldsymbol{\varepsilon}_k$. The results are displayed in Figure 3.2.2. When more cell lines are observed in our simulation with a definite number of groups $\hat{K}$ and group characteristics ($\boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\pi}_{1:\hat{K}}$), variable selection accuracy will be improved obviously. This is also predictable since more observations will bring in more information for us to exploit.

### Increase the Number of Genes $p$

Then we increase the number of genes $p$ from 1000 to 1300. The new design matrix $\mathbf{X}^{(real)}_{n\times 1300}$ is also a extension of the old one $\mathbf{X}^{(real)}_{n\times 1000}$. To make it compatible, the length of $\hat{\boldsymbol{\beta}}$ is extended to 1300 for each $\hat{\boldsymbol{\beta}}_k$ by filling in zeros. Here $\hat{\boldsymbol{\beta}}_k$ used to be a 1000−dimension vector, now it becomes a 1300−dimension vector with 300 more zeros in the rear. After generating $\mathbf{Y}^{(sim)}_{n\times J}$

| p=1000 | | | p=1300 | | |
|---|---|---|---|---|---|
| SEN2 mean | SPE2 mean | SEN2 std error | SEN3 mean | SPE3 mean | SEN3 std error |
| 43.22% | 95.39% | 0.00211 | 36.31% | 95.39% | 0.00208 |

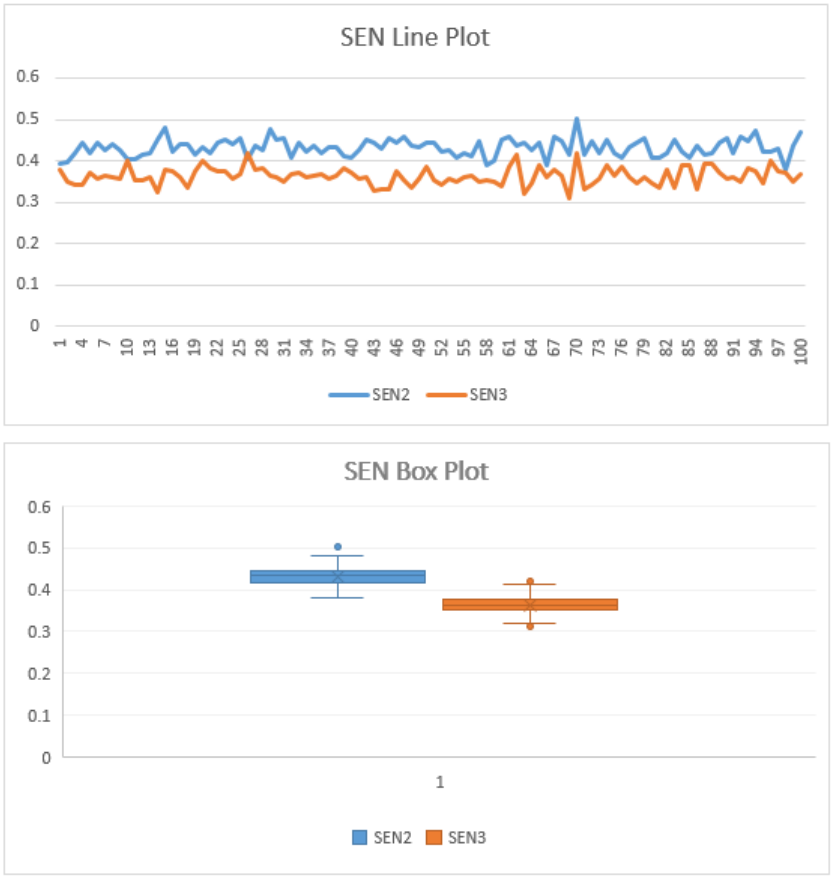



Figure 3.2.3: The sensitivity table, line chart and box plot for 100 trials under the numbers of genes $p = 1000$ (blue) and $p = 1300$ (orange).

with $\mathbf{y}_{(j)} = \mathbf{X}_{(j)}^{(real)}\hat{\boldsymbol{\beta}}_k + \boldsymbol{\varepsilon}_{jk}$, same as above, we perform penalised regression with $\mathbf{X}_{n\times p}^{(real)}$ and $\mathbf{Y}_{n\times J}^{(sim)}$ to compare selection accuracies by SENs and SPEs (only sensitivities are plotted below). Repeat trials with each setting ($p = 1000$ and $p = 1300$) one hundred times respectively to remove the randomness of error term $\boldsymbol{\varepsilon}_k$. The results are shown in Figure 3.2.3. When more genes are added into our simulation data generation process as noises with a definite number of groups $\hat{K}$ and group characteristics ($\boldsymbol{\beta}, \boldsymbol{\sigma}$ and $\boldsymbol{\pi}_{1:\hat{K}}$), variable selection accuracy will distinctly drop. This happens as a result of the dilution of active data because the added new data act as noises to reduce the ratio of active (non-zero) variables, which leads to $\boldsymbol{\beta}$ sparser and increases between-group homogeneity. Then our clustering and variable selection become trickier to handle.

#### 3.2.3.3 Compare the Performances of $q = 0.5$ Case and $q = 1$ Case

As the intuition that a smaller $q$ would produce a more sparse vector, SEN and SPE are not suitable anymore for the comparison of $q = 0.5$ case and $q = 1$ case. As a consequence, here we define a new measure to evaluate our estimation accuracy by quantifying the difference between our estimated coefficients and the underlying true ones.

$$\text{Relative Error} \quad := \quad \sum_{k=1}^{K} \hat{\pi}_k \frac{\left\| \hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^r \right\|_2}{\|\boldsymbol{\beta}_k^r\|_2},$$

where $\boldsymbol{\beta}_k^r$ is the true underlying coefficients, and in the simulations we simulate the active point values by sampling from $N(0,1)$ under a certain sparsity level $s^r$.

Firstly, we start our simulation by assuming that the given $J = 20$ drugs come from $K = 2$ groups with a probability of 60% from Group 1 and a probability of 40% from Group 2. From part of original matrix $\mathbf{X}^{(real)}$, we construct matrix $\mathbf{Y}^{(sim)}$ with $\mathbf{y}_j^{(sim)} = \mathbf{X}_{50\times 500}^{(real)} \hat{\boldsymbol{\beta}}_k + \boldsymbol{\varepsilon}_{jk}$ where $\boldsymbol{\varepsilon}_{jk}$ is sampled from the multivariate normal distribution $N_{n=50}\left(\mathbf{0}, \sigma_k^2 \mathbf{I}\right)$ with $(\sigma_1, \sigma_2) = (0.5, 0.6)$. $\mathbf{I}$ is a $50-$dimensional identity matrix. Each time we set a real sparsity level $s^r$ from $(5, 10, 15, 20, 25, 30, 35, 40, 45, 50)$, and generate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ with $(s^r, s^r)$ active points whose values are sampled from standard normal distribution $N(0,1)$. After constructing $\mathbf{Y}_{50\times 20}^{(sim)}$, we fabricate some missing data artificially by assuming that there is an equal chance for each entry to become missing in each column. And the missingness proportion for each drug (column) is sampled from uniform distribution $U(0, 20\%]$, thus we control our missingness proportion within 20%. Then we perform penalised regression with $q = 1$ to select $\hat{K}$ and $\lambda_1$, meanwhile, produce a warm start $\hat{\boldsymbol{\Phi}}_1$ for $q = 0.5$ penalised regression. After that, we select the best $\lambda_{0.5}$ and use it to reach our final estimate $\hat{\boldsymbol{\Phi}}_{0.5}$. To evaluate the performances of $q = 0.5$ case (with the warm start) and $q = 1$ case, we calculate the Relative Error with $\hat{\boldsymbol{\Phi}}_1$ and $\hat{\boldsymbol{\Phi}}_{0.5}$ respectively. After repeating each trial 50 times under one certain sparsity level $s^r$, we take their average and plot Figure 3.2.4 to compare their performances under different sparsity levels. As shown in the plot, the estimation accuracy of both cases will drop dramatically as $\boldsymbol{\beta}$ gets less sparse. Until $\frac{50}{500} = 10\%$ of the data entries are active (sparsity level 50), the Relative Error rises to 1 which means no estimation made, as $\hat{\boldsymbol{\beta}}_k = \mathbf{0}$ will lead to Relative Error equal to 1 as well. And $q = 0.5$ case (with the warm start) can always outperform $q = 1$ case in terms of estimation accuracy measured by Relative Error.

#### 3.2.3.4 Evaluate the Efficacy of the Warm Start

Under the setting that real sparsity level $s^r = 5$ and active points of $\boldsymbol{\beta}$ sampling from $U(-0.5, 0.5)$ (instead of $N(0,1)$ as above to produce a weaker signal), each time we generate simulation data-

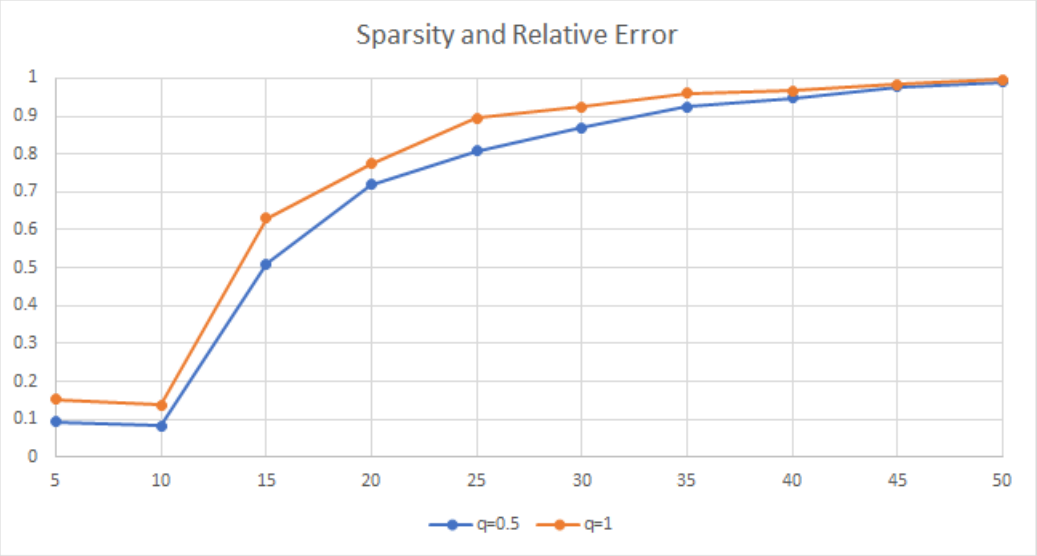



Figure 3.2.4: The Relative Error comparison for $q = 0.5$ case (with the warm start) and $q = 1$ case against the increase of underlying coefficient sparsity (from 5 to 50 active coefficients).

| | SEN | SPE | Relative Error |
|---|---|---|---|
| $q = 1$ case | 0.8875 | 0.9774 | 0.2636 |
| $q = 0.5$ without warm start | 0.8665 | 0.9739 | 0.2544 |
| $q = 0.5$ with the warm start | 0.9055 | 0.9897 | 0.1371 |

Table 3.2: The summary table about SEN, SPE and Relative Error for different penalty and initialisation settings.

set $\mathbf{Y}^{(sim)}$ as explained above. To show the efficacy of our warm start, we use three different methods to perform our regressions, i.e. $q = 1$ case, $q = 0.5$ case without a warm start and $q = 0.5$ case with the warm start. After 20 trials, the average results are laid out in Table 3.2. As shown, after using the result of $q = 1$ case as a warm start, $q = 0.5$ case can reach a higher performance in terms of SEN, SPE and Relative Error.

### 3.2.4 Numerical Results on Anti-Cancer Drug Data

With the same dataset matrix $\mathbf{Y}_{n\times J}$, Iorio et al (2016) performed a drug clustering based on the area under curve (AUC) pattern similarity. In their results, 223 drugs were clustered into 26 groups. This drug heterogeneity suggests that a single regression model may not be proper. Our mixture modelling can just deal with this hidden group structure. Gupta and Ibrahim (2007) successfully demonstrated a unified mixture model which could perform clustering and variable selection simultaneously. Inspired by this aim, we build our model and realise the same functions as theirs. Faced with the same situation as Khalili et al (2011) that the number of potential features $p$ is much larger than the number of observations $n$, we use another method to

| | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $K=5$ | $K=6$ |
|---|---|---|---|---|---|---|
| $\lambda_1=0.010$ | 1306.529 | 1250.557 | 1225.578 | 1228.122 | 1249.891 | 1272.006 |
| $\lambda_1=0.015$ | 1308.901 | 1248.825 | 1223.047 | 1223.795 | 1235.240 | 1253.420 |
| $\lambda_1=0.020$ | 1311.940 | 1250.225 | 1223.359 | 1222.517 | 1224.951 | 1243.708 |
| $\lambda_1=0.025$ | 1315.127 | 1253.470 | 1225.351 | 1223.360 | 1220.526 | 1241.193 |
| $\lambda_1=0.030$ | 1318.933 | 1256.267 | 1246.357 | 1224.856 | 1220.610 | 1238.878 |
| $\lambda_1=0.035$ | 1322.561 | 1259.620 | 1250.306 | 1227.059 | 1223.836 | 1239.937 |
| $\lambda_1=0.040$ | 1325.838 | 1263.040 | 1253.644 | 1230.487 | 1225.592 | 1240.111 |

Table 3.3: BIC grid for $q=1$ case on real data to show that $(\lambda_1=0.025, K=5)$ has the minimum value 1220.526 which is regarded as the optimal.

diminish $p$ to alleviate the computational burden in such a high-dimensional problem. Starting with $\mathbf{X}_{983\times17737}$ and $\mathbf{Y}_{983\times251}$, all $p=17737$ genes are our potential features. Considering the computing capability of computers, our composite matrix $\mathbf{X}^*_{N\times p}$ $(N=\sum_{j=1}^{J} n_{(j)})$ (see Section 3.2.1) becomes $\mathbf{X}^*_{190628\times17737}$ which excessively exceeds the allocation capacity of our computers. To make it possible to handle, we first implement individual LASSO onto each drug to pick out the genes related to each drug. Then the remaining genes which have no relation to any drugs are deleted from all $p$ genes. After individual LASSO, $p$ is diminished from 17737 to 9083. Since there are many missing entries in matrix $\mathbf{Y}_{983\times251}$, considering the effect of outliers, empirically we just keep the drugs and cell lines containing more than 80% information to reduce $J$ from 251 to 187 and $n$ from 983 to 795, where 80% means there are 80% entries in that row or column being not missing. For $\mathbf{X}_{795\times9083}$ and $\mathbf{Y}_{795\times187}$, we calculate the row variances for matrix $\mathbf{Y}_{795\times187}$, then use k-means clustering method to divide these 795 variances into two groups and keep the higher-value group to reduce $n$ again from 795 to 397. After that, we conduct individual drug analysis with $q=0.5$ penalty by using IRucLq-v algorithm (Lai et al., 2013). The result $\left(\boldsymbol{\beta}_1^{(ind)}, \boldsymbol{\beta}_2^{(ind)}, \ldots, \boldsymbol{\beta}_{187}^{(ind)}\right)$ will indicate the related features (genes) to each drug, thus the top 1000 features (by adding up to the total weights for each gene) are kept. From the results of regression with $q=1$, we have learned that each drug is generally related to $100\sim200$ genes. So keeping 1000 genes as candidates for our next analysis is adequate and appropriate considering the computation speed of computers. Finally these dimension-reduction methods lead us to $\mathbf{X}_{397\times1000}$ and $\mathbf{Y}_{397\times187}$. With $q=1$ case, the smallest BIC value sits at $(\lambda_1=0.025, K=5)$ as shown in the Table 3.3. From this $\hat{K}=5$, we construct a BIC-value grid for possible $\lambda_{0.5}$, then using $q=0.5$ case penalised regression analysis we obtain the BIC-value grid as Table 3.4. Using the optimal point from Table 3.4, the final clustering and gene selection outcomes are listed in Table 3.5, Table 3.6 and Table 3.7. The related-gene coefficients for each group are plotted in Figure 3.2.5. In **Appendix A**, to compare with our final results, we also list the results from $q=1$ case after calculating with $(\lambda_1=0.025, K=5)$.

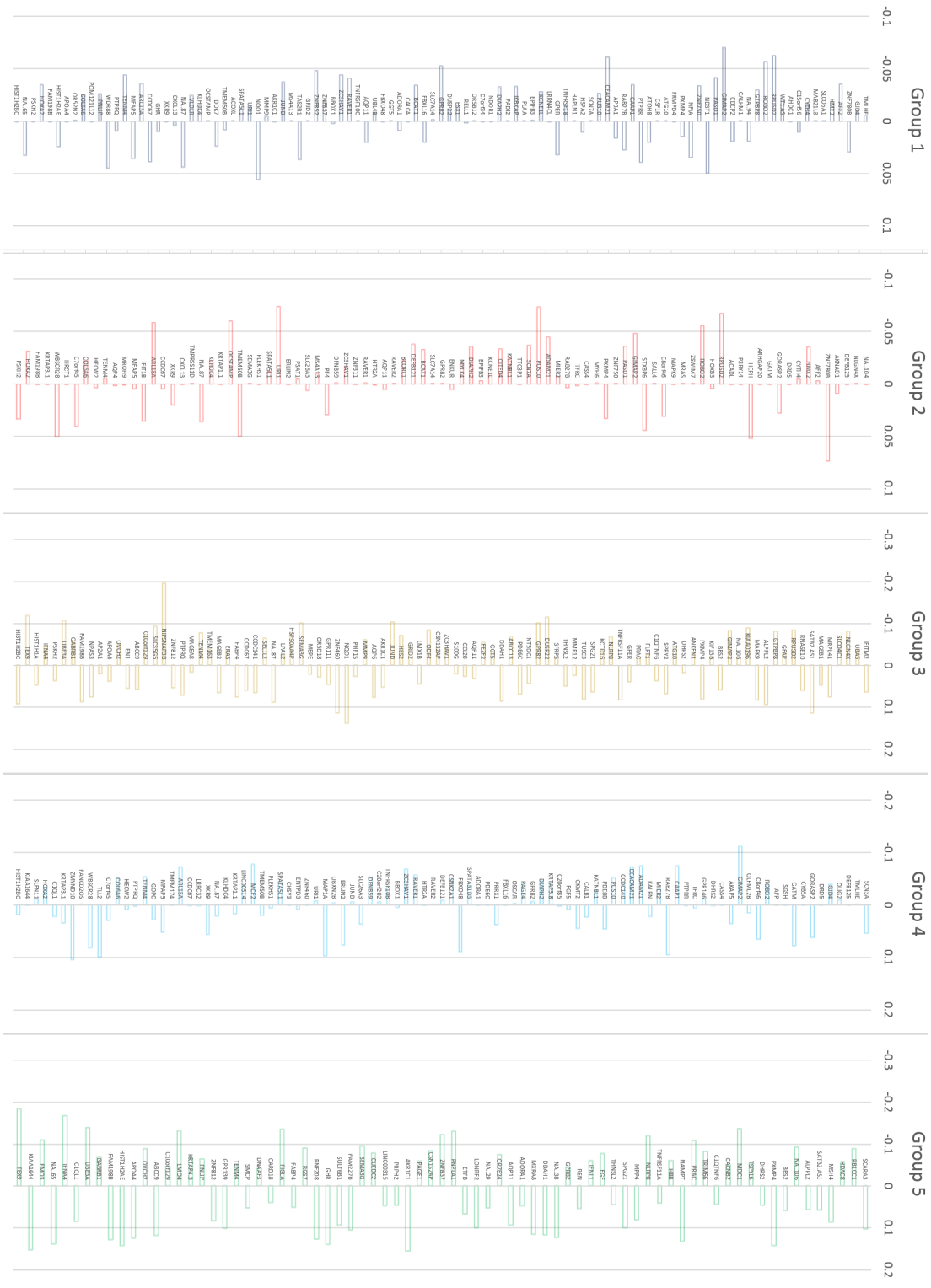



Figure 3.2.5: Coefficient bar-plot for each group. Selected genes are labelled and run from top to bottom in each vertical long line. Small hollow bars pointing to the right indicate positive weights, and those to the left indicate negative weights.

| | $K=1$ | $K=2$ | $K=3$ | $K=4$ | $\hat{K}=5$ | $K=6$ |
|---|---|---|---|---|---|---|
| $\lambda_{0.5}=0.001$ | 1350.307 | 1270.243 | 1288.399 | 1272.772 | 1262.720 | 1285.654 |
| $\lambda_{0.5}=0.002$ | 1304.643 | 1242.177 | 1266.980 | 1251.313 | 1234.266 | 1255.397 |
| $\lambda_{0.5}=0.003$ | 1303.190 | 1238.743 | 1253.645 | 1239.584 | 1231.022 | 1240.678 |
| $\lambda_{0.5}=0.004$ | 1311.045 | 1243.675 | 1246.849 | 1225.467 | 1216.436 | 1225.252 |
| $\lambda_{0.5}=0.005$ | 1314.379 | 1248.228 | 1245.891 | 1228.392 | 1202.692 | 1234.803 |
| $\lambda_{0.5}=0.006$ | 1316.136 | 1254.292 | 1247.376 | 1239.010 | 1205.142 | 1215.820 |
| $\lambda_{0.5}=0.007$ | 1320.542 | 1258.935 | 1222.912 | 1216.696 | 1208.888 | 1216.942 |
| $\lambda_{0.5}=0.008$ | 1324.985 | 1262.924 | 1227.683 | 1222.031 | 1216.382 | 1222.082 |
| $\lambda_{0.5}=0.009$ | 1329.003 | 1267.183 | 1233.683 | 1226.407 | 1220.988 | 1226.697 |
| $\lambda_{0.5}=0.010$ | 1333.158 | 1270.512 | 1237.936 | 1229.972 | 1226.992 | 1230.691 |

Table 3.4: BIC values show that $\left(\lambda_{0.5}=0.005, \hat{K}=5\right)$ has the minimum value 1202.692, it is an optimal point. Generally we just need to compute the $\hat{K}=5$ column to locate the optimal point, which can save much time.

### 3.2.5 Discussion and Conclusion

We have proposed a methodology to address high-dimensional clustering and variable selection problems when given information has missing data. We fuse all drugs into one model to realise clustering and variable selection simultaneously. And from the comparison of two different non-convex penalisation terms, readers can have a general view of the performances of mixture models under the different penalised regression systems. Meanwhile, from initialisation to thresholding, our novel methodology can help researchers when faced with some non-convex penalised mixture modellings. During our research, imputation methods were also attempted to complete the matrix $\mathbf{Y}_{n\times J}$ first, then applied the complete matrix to our methodology for discussing their performances. These imputation methods included Soft Impute (Mazumder, Hastie and Tibshirani, 2010) and primePCA (Zhu, Wang and Samworth, 2019). After imputation, the complete dataset did not outperform our original dataset when applied to our methodology (so these imputation analyses are not included in this thesis), because the imputed entries might contaminate the whole data matrix by bringing in new errors. But it still needs time to testify whether there exists a kind of imputation method which surely improves our methodology under different simulation settings.

Table 3.5: Drugs in each group and some group characters. (Part 1)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| 1 | LFM.A13, SGC0946, Ruxolitinib, FMK, CP724714, STF.62247, Zibotentan, VNLG.124, OSI.930, Tivozanib, T0901317, Selisistat, Veliparib, Lenalidomide, Vismodegib, SL0101, IOX2, UNC1215, Temozolomide, PFI.3, GW.2580, XMD14.99, KIN001.236, FR.180204, KIN001.270, XMD15.27, Tamoxifen, VX.702, GSK1904529A. | 29 | 103 | 0.155 | 0.686 |
| 2 | NSC.87877, Midostaurin, JNK.9L, PF.562271, FTI.277, OSU.03012, Shikonin, Embelin, FH535, Dacinostat, Bexarotene, PAC.1, Luminespib, Linifanib, Bryostatin.1, TAK.715, AS601245, Alectinib, GSK429286A, WHI.P97, Quizartinib, KIN001.266, AICA.Ribonucleotide, Vinblastine, Cisplatin, Docetaxel, Tretinoin, Gefitinib, Vorinostat, Axitinib, GW441756, Lestaurtinib, Motesanib, KU.55933, BX795, NU7441, Doramapimod, JNK.Inhibitor.VIII, PD173074, ZM447439, RO.3306, MK.2206, Dactolisib, AZD8055, CCT007093, EHT.1864, | 57 | 89 | 0.305 | 1.01 |

Table 3.6: Drugs in each group and some group characters. (Part 2)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| | Cetuximab, PF.4708671, Serdemetan, TW.37, CCT.018159, Rucaparib, SB505124, PFI.1, YK.4.279, XAV939, Piperlongumine. | | | | |
| 3 | CP466722, BMS.345541, TL.2.105, Idelalisib, Cabozantinib, JW.7.24.1, NPK76.II.72.1, NG.25, TL.1.85, Tubastatin.A, I.BET.762, BIX02189, KIN001.244, Masitinib, KIN001.260, PIK.93, MPS.1.IN.1, TPCA.1, NVP.BHG712, BX.912, Fedratinib, Foretinib, Y.39983, YM201636, QL.XI.92, XMD13.2. | 26 | 94 | 0.139 | 1.116 |
| 4 | Doxorubicin, Etoposide, Gemcitabine, Vinorelbine, Mitomycin.C, QS11, Ponatinib, HG6.64.1, JQ12, GSK650394, DMOG, BAY.61.3606, IPA.3, Thapsigargin, Obatoclax.Mesylate, BMS.754807, Linsitinib, Bleomycin, Phenformin, Pazopanib, Epothilone.B, Tipifarnib, AS605240, Enzastaurin, VX.11e, NSC.207895, SB52334, Amuvatinib, Cytarabine, Methotrexate, Nilotinib, Navitoclax, CI.1040, Temsirolimus, Bosutinib, AZD7762, Tanespimycin, Elesclomol, Nutlin.3a, Palbociclib, PD0325901, | 50 | 95 | 0.267 | 1.570 |

Table 3.7: Drugs in each group and some group characters. (Part 3)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| | SB590885, X.5Z..7.Oxozeaenol, Talazoparib, rTRAIL, ICL1100013, SN.38, Trametinib, Dabrafenib, Bleomycin..50.uM. | | | | |
| 5 | X5.Fluorouracil, AR.42, Ispinesib.Mesylate, AT.7519, GSK1070916, Daporinad, QL.XII.47, ZSTK474, WZ3105, Genentech.Cpd.10, Sepantronium.bromide, CUDC.101, Belinostat, CAY10603, Pelitinib, Omipalisib, OSI.027, CX.5461, PHA.793887, PI.103, GSK690693, SNX.2112, QL.X.138, THZ.2.49, THZ.2.102.1. | 25 | 75 | 0.134 | 1.620 |

## 3.3 Discussions about Imputation and Component Number Selection

### 3.3.1 Introduction

The matrix completion problem has drawn large attention over years. Formerly, we have multiple imputation methods (Rubin, 1987) which replace each missing value with multiple possible values. Missing values here mean the unobserved data that would be meaningful for analysis if observed, or rather, a missing value masks an underlying meaningful value (Little and Rubin, 2014). Nowadays when faced with this kind of problem, we usually use **softImpute** algorithm and **hardImpute** algorithm (Mazumder, Hastie and Tibshirani, 2010; Hastie, Mazumder, Lee and Zadeh, 2015), which impute (namely fill in) the missing entries of one matrix by solving an optimisation problem in a series of iterations. And these methods have reached decent results. After all, imputation can help a lot. As these missing values are once filled in, many complete-data methods for analysis can be adopted (Rubin, 1987), moreover, most computer programs can be run without error reports regarding incomplete data input. To this advantage, more efforts are taken by worldwide researchers recently. Zhu, Wang and Samworth (2019) developed one algorithm called *Projected Refinement for Imputation of Missing Entries in PCA* (**primePCA**), which can be regarded as state of the art for matrix completion problems. It is an iterative algorithm using singular value decomposition (SVD) to refine the principal compon-

ents of the target matrix. **PrimePCA** is also the first method designed for high-dimensional PCA to deal with heterogeneous missingness settings (Zhu, Wang and Samworth, 2019). With the application to Million Song Dataset, **primePCA** showed strong performance and yielded a reasonable result. But as a nuisance parameter, the choice of $\hat{K}$ (the number of the principal components of interest) has not been discussed in their article. Moreover, the model can be augmented by including additional explanatory variables with fixed effects $\boldsymbol{\beta}$, which could capture relations between covariates and responses. To realise these aims, the algorithm is explained next, and after the discussions, we will propose another method about the selection of the number of components $\hat{K}$.

### 3.3.2 Algorithm

#### 3.3.2.1 Regression and Imputation

We are provided with a design matrix $\mathbf{X}_{n\times p}$ and a response matrix with missing entries $\mathbf{Y}_{n\times J}$. We assume that they have a linear relationship after imputing missing entries in $\mathbf{Y}_{n\times J}$, which is $\hat{\mathbf{Y}}_{n\times J} = \boldsymbol{\alpha}\mathbf{1}^\intercal + \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times J} + \mathbf{W}_{n\times J} + \mathbf{Z}_{n\times J}$ where $\boldsymbol{\beta}_{p\times J}$ is a column-wise coefficient matrix representing fixed effects, $\boldsymbol{\alpha}\mathbf{1}^\intercal$ is an intercept term, $\mathbf{W}_{n\times J}$ is a matrix representing random effects with $K$ principal components, and $\mathbf{Z}_{n\times J}$ is a standard normal error term. Our targets are

- Estimate the coefficient matrix $\boldsymbol{\beta}_{p\times J}$;
- Estimate the nuisance parameter $K$;
- Impute the missing entries in $\mathbf{Y}_{n\times J}$ to reach a complete matrix $\hat{\mathbf{Y}}_{n\times J}$.

To fulfil these aims, we need to carry out a regression. Pick out all complete rows in $\mathbf{Y}_{n\times J}$ and denote the new matrix by $\mathbf{Y}_I$. A complete row means there is no missing entry in that row of $\mathbf{Y}_{n\times J}$. Here $I$ in $\tilde{\mathbf{Y}}_I$ is the ordinal number of all non-missing rows. We use vector $\bar{Y}$ to represent the mean of each row. With centralisation on rows, we estimate the residuals by

$$\hat{\boldsymbol{\varepsilon}}_I = \mathbf{Y}_I - \bar{Y}\cdot\mathbf{1}^\intercal_{|I|\times 1} - \mathbf{X}_I\hat{\boldsymbol{\beta}},$$

where $\mathbf{X}_I$ is the sub-matrix of $\mathbf{X}_{n\times p}$ with corresponding rows as $\mathbf{Y}_I$, and $|I|$ is the cardinal number of $I$. Here $\hat{\boldsymbol{\beta}}$ is from the multi-response LASSO estimation that

$$\hat{\boldsymbol{\beta}} \quad = \quad \underset{\boldsymbol{\beta}\in\mathbb{R}^{p\times J}}{\arg\min}\frac{1}{2}\left\|\mathbf{Y}_I - \bar{Y}\cdot\mathbf{1}^\intercal - \mathbf{X}_I\hat{\boldsymbol{\beta}}\right\|_F^2 + \lambda\sum_{u=1}^{p}\left\|\hat{\boldsymbol{\beta}}_u\right\|_2.$$

The complementary residual part is computed by

$$\hat{\boldsymbol{\varepsilon}}_{I^c} = \mathbf{Y}_{I^c} - \bar{Y} \cdot \mathbf{1}^{\intercal}_{|I^c| \times 1} - \mathbf{X}_{I^c}\hat{\boldsymbol{\beta}},$$

which contains the same missing entries as $\mathbf{Y}_{I^c}$. After that, we reach the estimate of residuals

$$\hat{\boldsymbol{\varepsilon}}_{n \times J} = \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_I \\ \hat{\boldsymbol{\varepsilon}}_{I^c} \end{pmatrix} \text{(assembled as the same row order as } \mathbf{Y}_{n \times J}).$$

With the help of package **primePCA** and **Algorithm 1** from Zhu, Wang and Samworth (2019), we impute the missing entries in $\hat{\boldsymbol{\varepsilon}}_{n \times J}$ with estimate $\hat{\hat{\varepsilon}}_{ij}$ and keep the existing entries in $\hat{\boldsymbol{\varepsilon}}_{n \times J}$ unchanged. Finally missing entries in our observed response matrix $\mathbf{Y}_{n \times J}$ are imputed via

$$\hat{y}_{ij} \quad = \quad \bar{Y}_i + \mathbf{x}_i\hat{\boldsymbol{\beta}}_j + \hat{\hat{\varepsilon}}_{ij}.$$

Hence, we obtain a complete matrix $\hat{\mathbf{Y}}_{n \times J}$.

We can calculate the mean absolute error (MAE) for imputed values in $\hat{\mathbf{Y}}_{n \times J}$ to quantify the accuracy when underlying missing values $y_{ij}^{(m)}$s are given in simulations, which is (assume that there are $N$ missing entries totally in $\mathbf{Y}_{n \times J}$)

$$\text{MAE} \quad = \quad \frac{\sum_{m=1}^{N} \left| \hat{y}_{ij}^{(m)} - y_{ij}^{(m)} \right|}{N}.$$

#### 3.3.2.2 Estimate $K$

When using package **primePCA**, the argument $\hat{K}$ (the number of the principal components of interest) has to be determined. When underlying missing values $y_{ij}^{(m)}$s are known in simulations, the MAE values can be calculated under different $\hat{K}$s. And the chosen $\hat{K}$ locates at the smallest MAE. This is done in our next simulation part and the result is shown in Figure 3.3.1.

But in real data analysis, missing values will never be known. We need to find some counterparts to replace unknown missing values to realise the same function. Here we can regard adjacent entries as missing ones which are the entries next to true missing ones. We set these adjacent entries in $\mathbf{Y}_{n \times J}$ as NA (treat them the same way as true missing ones) before implementing package **primePCA**. Then impute these adjacent entries as the way mentioned above but keep the true missing ones still as NA. Then use these known 'missing' entries to calculate MAE (true missing ones are set as 0) and select the $\hat{K}$s from the smallest MAE. This is also carried out in the next simulation part.

### 3.3.3 Simulation

First of all, under $n = 100, p = 50, J = 20$, we sample a $n \times p$ matrix $\mathbf{X}_{n\times p}$ by independently generating $\mathbf{x}_i$ from $N_p(\mathbf{0}, \Sigma_x)$, here $\mathbf{x}_i$ ($i \in \{1, 2, \ldots, n\}$) denotes the $i$-th row of $\mathbf{X}_{n\times p}$, $\mathbf{0}$ is the $p$-sized zero vector ($\mathbf{0} = (0, 0, \ldots, 0)^\intercal$), and

$$\Sigma_x^{p\times p} = \begin{pmatrix} 1 & 0.1 & 0.1^2 & 0 & \cdots & 0 \\ 0.1 & 1 & 0.1 & 0.1^2 & \ddots & \vdots \\ 0.1^2 & 0.1 & 1 & 0.1 & \ddots & 0 \\ 0 & 0.1^2 & 0.1 & 1 & \ddots & 0.1^2 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0.1 \\ 0 & \cdots & 0 & 0.1^2 & 0.1 & 1 \end{pmatrix}.$$

Then a coefficient matrix $\boldsymbol{\beta}_{p\times J} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_J)$ is built, where the column vectors have sparse alternate entries as

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 2 \\ -2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \boldsymbol{\beta}_2 = \begin{pmatrix} 0 \\ -2 \\ 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \boldsymbol{\beta}_3 = \begin{pmatrix} 0 \\ 0 \\ 2 \\ -2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \ldots$$

Let $\mathbf{w}_i$ denote the $i$-th row of $\mathbf{W}_{n\times J}$, to make $K = 3$, the $\mathbf{w}_i$s are independently sampled from $N_J(\mathbf{0}, \mathbf{o}_1 \cdot \mathbf{o}_1^\intercal + \mathbf{o}_2 \cdot \mathbf{o}_2^\intercal + \mathbf{o}_3 \cdot \mathbf{o}_3^\intercal)$ where $\mathbf{0}$ is the $J$-sized zero vector and $\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3$ are orthogonal with each other:

$$\mathbf{o}_1 = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \mathbf{o}_2 = \begin{pmatrix} 2 \\ \vdots \\ 2 \\ -2 \\ \vdots \\ -2 \end{pmatrix}, \qquad \mathbf{o}_3 = \begin{pmatrix} 3 \\ -3 \\ \vdots \\ 3 \\ -3 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

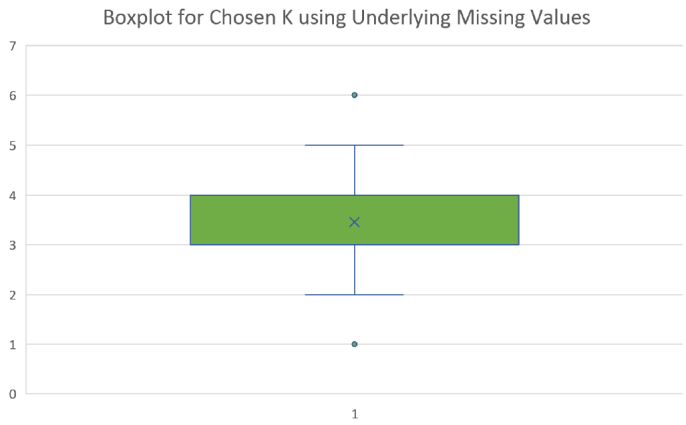



Figure 3.3.1: Box-and-whisker plot for chosen $\hat{K}$ from 50 sample trials. MAE made use of the underlying true missing entries.

Let $\mathbf{z}_i$ denote the $i$-th row of $\mathbf{Z}_{n\times J}$, the $\mathbf{z}_i$s are independently generated from $N_J(\mathbf{0}, I_J)$, here $I_J$ is an identity matrix of size $J$. With a $n$-sized intercept vector $\boldsymbol{\alpha} = (2, 2, \ldots, 2)^\intercal$, we compute the complete response matrix $\tilde{\mathbf{Y}}_{n\times J}$ by

$$\tilde{\mathbf{Y}}_{n\times J} \quad = \quad \boldsymbol{\alpha}\mathbf{1}^\intercal + \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times J} + \mathbf{W}_{n\times J} + \mathbf{Z}_{n\times J}.$$

To simulate the missing values, we sample a missing-pattern matrix $\mathbf{Q}_{n\times J}$ from binomial distribution, where each entry $Q_{ij}$ follows the binomial distribution $B(1, 0.99)$, which means there is a possibility of 0.01 for each entry being missing. Then we get our observed response matrix $\mathbf{Y}_{n\times J}$ by the Hadamard product of $\tilde{\mathbf{Y}}_{n\times J}$ and $\mathbf{Q}_{n\times J}$ (missing entries are denoted as 0 because of $\mathbf{Q}_{n\times J}$):

$$\mathbf{Y}_{n\times J} \quad = \quad (y_{ij})_{n\times J} = \tilde{\mathbf{Y}}_{n\times J} \circ \mathbf{Q}_{n\times J} = (\tilde{y}_{ij} \cdot Q_{ij})_{n\times J}\,.$$

Then we carry out our aforementioned algorithm with these $\mathbf{X}_{n\times p}$ and $\mathbf{Y}_{n\times J}$.

Figure 3.3.1 is the box-and-whisker plot after conducting 50 sample trials by sampling different design matrices and error terms. We also put the statistics for these 50 sample trials in Table 3.8. Here in Figure 3.3.1 and Table 3.8, we suppose that underlying missing values are all known to us. So MAEs are calculated based on them. As seen from Figure 3.3.1, most of the chosen $\hat{K}$ fall into 3 or 4, and the average lies at around 3.5, so when provided with underlying missing values in simulations, MAE is a good statistic to identify $\hat{K}$. Table 3.8 again supports this idea by showing that the smallest Average MAE appears at $\hat{K} = 3$.

As mentioned, missing values cannot be known to us in real data analysis. So we conducted other 100 sample trials and covered adjacent entries (the right or left entries next to true missing ones) as artificial missingness, but uncover them when calculating MAE. Figure 3.3.2 lays out

| Stats for Figure 3.3.1 | $\hat{K}=1$ | $\hat{K}=2$ | $\hat{K}=3$ | $\hat{K}=4$ | $\hat{K}=5$ | $\hat{K}=6$ |
|---|---|---|---|---|---|---|
| Average MAE | 2.421 | 1.851 | 1.657 | 1.761 | 1.809 | 1.877 |
| Standard Errors | 0.070 | 0.049 | 0.042 | 0.047 | 0.043 | 0.041 |

Table 3.8: Statistics table as complementary information for 50 sample trials related to Figure 3.3.1.

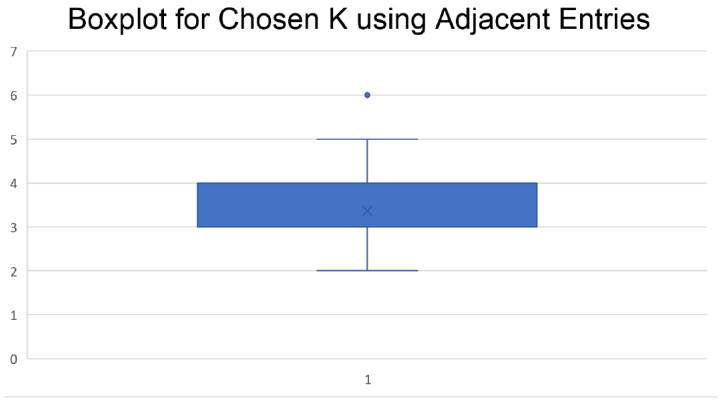



Figure 3.3.2: Box-and-whisker plot for chosen $\hat{K}$ from 100 sample trials. MAE made use of the entries adjacent to missing ones.

these simulations and Table 3.9 shows the statistics behind them. As seen in Figure 3.3.2, this box plot has nearly the same pattern as Figure 3.3.1. So it suggests that our adjacent-entry method works well for estimating this $K$. Table 3.9 also supports our method as $\hat{K}=3$ has the smallest Average MAE.

This adjacent-entry method is actually a kind of cross-validation. But the adjacent-entry method uses specific subsamples as a validation set. Generally we can choose other entries instead of adjacent ones as our validation set, such as random ones in every row which has missing entries. But when the missing proportion increases to more than half of the row size, we can not find the same number of counterparts as the missing ones. Adjacent entries will always be found in this circumstance as overlapping is considered, and here overlaps (true values cannot be known) should be retained as 0 after imputation to avoid mismatch.

### 3.3.4 Complete-Case Problem

This MAE statistic can also be generalised to solve the complete-case problem where there are no missing entries in the target matrix and we just want to find out the possible $K$. In this

| Stats for Figure 3.3.2 | $\hat{K}=1$ | $\hat{K}=2$ | $\hat{K}=3$ | $\hat{K}=4$ | $\hat{K}=5$ | $\hat{K}=6$ |
|---|---|---|---|---|---|---|
| Average MAE | 2.327 | 1.838 | 1.600 | 1.697 | 1.768 | 1.865 |
| Standard Errors | 0.043 | 0.032 | 0.029 | 0.031 | 0.033 | 0.037 |

Table 3.9: Statistics table as complementary information for 100 sample trials related to Figure 3.3.2.

| Missing Percentage | Success Rate | Average Chosen $\hat{K}$ | Standard Errors |
|---|---|---|---|
| 0.05% | 30.0% | 3 | 0.494 |
| 0.15% | 20.0% | 3 | 0.471 |
| 0.50% | 20.0% | 3.8 | 0.291 |
| 1.00% | 70.0% | 3.3 | 0.153 |
| 1.50% | 80.0% | 3.2 | 0.133 |
| 2.00% | 80.0% | 3.2 | 0.133 |
| 2.50% | 80.0% | 3.3 | 0.213 |
| 4.00% | 80.0% | 3.3 | 0.213 |
| 5.00% | 70.0% | 3.6 | 0.306 |
| 7.50% | 60.0% | 3.6 | 0.267 |
| 10.00% | 70.0% | 3.5 | 0.269 |

Table 3.10: Statistics for different missing percentage levels with respect to indicator matrix $\mathbf{Q}$.

case, we have to fabricate missing entries first, which can be based on an indicator matrix $\mathbf{Q}$ from a binomial distribution. Then calculated MAE after imputation will pick out $\hat{K}$ and the boxplot will be like Figure 3.3.1. The indicator matrix $\mathbf{Q}$ should be sampled several times, and the final box plots will be compared with each other to reduce the randomness brought about by $\mathbf{Q}$. Now the only question lies in how to build up this indicator matrix $\mathbf{Q}$, which means how many percentages of missing entries are needed to best help us in identify this $\hat{K}$. To address this problem, we sample $\mathbf{Q}$ with different missing percentage levels from 0.05% to 10% and implement 10 trials with each missing percentage level respectively. The missing percentage level is not set to exceed 10% here because all non-missing rows are used to estimate $\hat{\boldsymbol{\beta}}$, larger missing percentage level probably results in no non-missing rows left. So under $\mathbf{Q}_{100\times 20}$ and the underlying $K = 3$ (the same setting as $\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3$), we obtained some statistics in Table 3.10. As seen from Table 3.10, when the missing percentage level was set as 1.5% or 2.0%, it has the highest success rate of selecting $\hat{K}$ (80.0% as displayed), and the Average Chosen $\hat{K}$ is closest to the underlying true $K = 3$ along with minimising the Standard Errors. Therefore, when encountering a complete-case problem, the missing percentage level can be fabricated as around 2.0% to facilitate the estimation of $K$ which is the number of the principal components of interest.

This result conforms to our expectation. On the one hand, when there is an excessively low missing percentage level, the effect of randomness will largely affect our choice of $K$ from calculated MAE values. For instance, 0.05% means only $nJ \times 0.05\% = 100 \times 20 \times 0.05\% = 1$ entry is fabricated as missing. Then in the calculation of MAE, only one imputed value will be involved. So the chosen $\hat{K}$ shows a high Standard Error in the first line of Table 3.10. At the same time, Success Rate can not be accepted. That Average Chosen $\hat{K}$ going to exact 3 just comes from the trade-off in the over- or under-estimate of $\hat{K}$. On the other hand, when there is an excessively high missing percentage level, the missing entries will cut down the amount of information by excluding more rows in our regression part, which results in a less accurate

estimate of $\hat{\boldsymbol{\beta}}$. More entries needed to impute will increase the difficulty of imputation from **primePCA**. Therefore, a mild missing percentage level is desired for our estimation of the number of components $K$.

In summary, in this section, we propose a good method of selecting the number of components for clustering problems, which is different from the BIC method we have used in previous sections. This new method is based on the matrix imputation method **primePCA** (Zhu, Wang and Samworth, 2019), and it can be used in not only complete-case problems where there is no missingness but also incomplete-case problems.

# Chapter 4

# Estimation of Skew Normal Model

## 4.1 Introduction

In the previous chapter, we conducted our analyses under the assumption that the error terms (residuals) of the regression follow a normal or mixture normal distribution. However, there are many situations where the assumption is invalid. There is a knowledge that the incorrectness of the underlying assumption may temper the utilisation and application of the methods (Shaphiro and Wilk, 1965). For example, observations from asymmetric distribution could mislead the inferences and distort the shape of mixture normal modelling. Thus, we turned to the skew normal distribution (Azzalini, 1985) to address the potential skewness feature within the dataset. The skew normal distribution extends the normal distribution function by including one more parameter called shape parameter $\alpha$ which is used to gauge the skewness. From this feature, the normal distribution function becomes a special case of the skew normal function when $\alpha = 0$. Hence skew normal is a more reasonable and flexible model to describe the asymmetric observations. At the same time, Azzalini (1985) found the peculiar singularity property for the Fisher information matrix as the shape parameter $\alpha$ approaches 0, then he proposed an alternative re-parametrisation system which is a kind of standardised version (called CP short for centred parameters, see Section 2.7) with a 'normalised' variable. Arellano-Valle and Azzalini (2008) explained that the standard skew normal density representation $f(y;\xi,\sigma,\alpha) = \frac{2}{\sigma}\phi\left(\frac{y-\xi}{\sigma}\right)\Phi\left(\alpha\left(\frac{y-\xi}{\sigma}\right)\right),\ y \in \mathbb{R}$ is problematic when applied to likelihood-based inference due to the unsuitable parametrisation in the model. Then our research starts by adapting this problematic parameter system. Inspired by this CP parametrisation, a different location-scale version of the skew normal density function is proposed in this chapter. Similar to CP version our location-scale version also benefits from the orthogonal feature of $\mu$ and $\sigma$ (they are diagonal in Fisher information matrix, see Section 4.3 below). As known,

the orthogonal parameters have independent maximum likelihood estimates, so each parameter can be estimated separately. Even after the introduction of the $\alpha-$penalty quadratic term, this problematic inference problem is addressed. Compared with the CP, our location-scale version has a simpler form, and thus it is easier in the computation when applied to likelihood-based inference such as estimation with EM algorithm. Moreover, most of the articles relating to skew normal distribution and EM algorithm used the linear summation form of normal and half normal distributions, see Lin et al. (2007), Baghfalaki (2011) and Teimouri (2020). In this chapter, a different EM algorithm is developed which is distinct from the conventional EM algorithm for skew normal estimation. In this method, we exploit Fisher transformation (Fisher, 1915) and logarithm transformation to help with the calculations in the EM algorithm.

## 4.2 Methodology

### 4.2.1 Location-Scale Skew Normal Model

Let $Z \sim \mathrm{SN}(0,1,\alpha)$ be the standard skew Normal with density $f_Z(z;\alpha) = 2\phi(z)\Phi(\alpha z), z \in \mathbb{R}$, where $\phi(z)$ and $\Phi(z)$ are the density and cumulative distribution function respectively, and $\alpha$ is the shape parameter. It is known (see Section 2.7) that $Z$ can be expressed as a linear combination of a half standard normal $Z_+$ and the standard normal $Z_0$ in the form $Z = \delta Z_+ + \sqrt{1-\delta^2}Z_0$, where $\delta = \alpha/\sqrt{1+\alpha^2}$ and $Z_+$ has the density $2\phi(z), z \geq 0$. We have $\mathrm{var}(Z) = 1 - 2\delta^2/\pi$ and $E[Z] = \delta\sqrt{2/\pi}$.

Let $Y = \mu + \sigma(Z - \delta\sqrt{2/\pi})$, where $\mu$ and $\sigma$ are location and scale parameters respectively. Then $Y$ has density function:

$$g\left(y;\mu,\sigma,\alpha\right) \quad = \quad \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\right)$$

with $E[Y] = \mu$, $\mathrm{var}(Y) = \sigma^2(1 - 2\delta^2/\pi)$.

To facilitate the calculations in EM algorithm, we introduce a new reparametrisation system by using Fisher transformation and logarithm transformation.

#### Re-parametrisation

Re-parametrise $\delta$ by using Fisher transformation

$$\theta = 0.5\log\left(\frac{1+\delta}{1-\delta}\right),\ \theta \in \mathbb{R},\ \delta \in (-1,1)\,,$$

thus, shape parameter $\alpha$ can also be expressed with $\theta$

$$\delta = \frac{e^{2\theta}-1}{e^{2\theta}+1} = 1 - \frac{2}{1+e^{2\theta}}, \qquad \alpha = \frac{1}{2}(e^{\theta} - e^{-\theta}),$$

then we re-parametrise scale parameter $\sigma$ by taking its logarithm, $\eta = \log(\sigma)$.

We can write our location-scale skew normal density function as

$$g\left(y;\mu,\eta,\theta\right) = \frac{2}{\sigma(\eta)} \cdot \phi\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right) \cdot \Phi\left(\alpha(\theta)\cdot\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)\right),$$

where

$$\sigma(\eta) = e^{\eta}, \qquad \delta(\theta) = 1 - \frac{2}{1+e^{2\theta}}, \qquad \alpha(\theta) = \frac{1}{2}\left(e^{\theta} - e^{-\theta}\right).$$

### 4.2.2 Estimation with EM Algorithm

#### Maximum Likelihood Estimation

Given independent observations $\mathbf{y} = (y_i)_{1\leq i\leq n}$, we derive the parameter estimates for our location-scale skew normal model through maximising likelihood:

$$\begin{aligned}\left(\hat{\mu},\hat{\eta},\hat{\theta}\right) &= \underset{(\mu,\eta,\theta)}{\arg\max} l_{inc}\left(\mu,\eta,\theta \mid \mathbf{y}\right) \\ &= \underset{(\mu,\eta,\theta)}{\arg\max} \sum_{i=1}^{n} \log g\left(y_i;\mu,\eta,\theta\right),\end{aligned}$$

where $l_{inc}\left(\mu,\eta,\theta \mid \mathbf{y}\right)$ is the **incomplete data log-likelihood** with

$$\begin{aligned}l_{inc}\left(\mu,\eta,\theta \mid \mathbf{y}\right) &= \sum_{i=1}^{n} \log\left\{\frac{2}{\sigma(\eta)}\phi\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\theta)\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)\right)\right\} \\ &= \frac{n}{2}\log\left(\frac{2}{\pi\sigma(\eta)^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)^2 \\ &\quad + \sum_{i=1}^{n}\log\left(\Phi\left(\alpha(\theta)\cdot\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)\right)\right).\end{aligned}$$

After deriving the first derivative to each parameter we arrive at the simultaneous equations. For this kind of interlocking equations, we use EM algorithm to solve them. Our EM algorithm consists of four parts: E-step, M-step, initialisation and stopping rule. To begin with, we augment the data $\mathbf{y}$ with another variable $W$ to form the following joint density.

**Complete Data Likelihood**

We introduce a positive random variable $W$ so that $(Y, W)$ has the **joint** density

$$g(y,w) = \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\phi\left(w - \alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\right), \qquad y \in \mathbb{R}, w \in (0,\infty).$$

Then the marginal density is

$$\begin{aligned}
\int_0^\infty g(y,w)dw &= \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma} + \delta\sqrt{2/\pi}\right)\int_0^\infty \phi\left(w - \alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{2/\pi}\right)\right)dw \\
&= \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma} + \delta\sqrt{2/\pi}\right)\left(1 - \Phi\left(-\alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{2/\pi}\right)\right)\right) \\
&= g(y). \\
g(w \mid y) &= \frac{g(y,w)}{g(y)} = \frac{\phi\left(w - \alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{2/\pi}\right)\right)}{\Phi\left(\alpha\left(\frac{y-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\right)}, \qquad w \in (0,\infty).
\end{aligned}$$

Augmenting $\mathbf{y}$ by $\mathbf{w} = (w_i)_{1\le i\le n}$, we form the complete data $(\mathbf{y}, \mathbf{w}) = (y_i, w_i)_{1\le i\le n}$. The **complete-data log-likelihood** is

$$\begin{aligned}
l_{com}(\mu,\eta,\theta \mid \mathbf{y},\mathbf{w}) &= \sum_{i=1}^n \log\left(g(y_i,w_i;\mu,\eta,\theta)\right) \\
&= \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\pi^2}\right) - \frac{1}{2}\sum_{i=1}^n\left(\frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^n\left(w_i - \alpha(\theta)\cdot\left(\frac{y-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}\right)\right)^2.
\end{aligned}$$

For the convenience of notation, we define

$$b_i = \frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}, \qquad b_i^{(v)} = \frac{y_i-\hat{\mu}^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}},$$

where $(v)$ means the $v$-th iteration. Then we start from the expectation step.

**E Step**

Given $\hat{\mu}^{(v)}$, $\hat{\eta}^{(v)}$ and $\hat{\theta}^{(v)}$ the estimates obtained in the $v$-th iteration, then the conditional expectation is

$$
\begin{aligned}
& \Psi(\mu,\eta,\theta \mid \hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)}) \\
= \;& E_{\mathbf{w}|\mathbf{y},\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)}}\left[l_{com}(\mu,\eta,\theta \mid \mathbf{y},\mathbf{w})\right] \\
= \;& n\log\left(\frac{1}{\sigma(\eta)\cdot\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{2/\pi}\right)^2 \\
& -\frac{1}{2}\sum_{i=1}^{n} E_{w_i|y_i,\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)}}\left[\left(w_i - \alpha(\theta)\cdot\left(\frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{2/\pi}\right)\right)^2\right] \\
= \;& \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\cdot\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \\
& \times\int_0^{\infty}\left(w_i-\alpha(\theta)\cdot b_i\right)^2\phi\left(w_i-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)dw_i \\
= \;& \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\cdot\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \\
& \times\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\left(w_i+\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)^2\phi(w_i)dw_i \\
= \;& \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\cdot\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \\
& \times\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\left[w_i^2+2\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)w_i\right]\phi(w_i)dw_i \\
& -\sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)^2\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\phi(w_i)dw_i \\
= \;& \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\cdot\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \\
& \times\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty} w_i^2\phi(w_i)dw_i - \sum_{i=1}^{n}\frac{\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i}{\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty} w_i\phi(w_i)dw_i \\
& -\sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)^2\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\phi(w_i)dw_i \\
= \;& \frac{n}{2}\log\left(\frac{1}{\sigma(\eta)^2\cdot\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \sum_{i=1}^{n}\frac{1}{2\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)} \\
& \times\left[-\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)\cdot\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} + \Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right) + 2\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)\right. \\
& \left.\times\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right) + \left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)^2\cdot\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)\right] \\
= \;& n\log(\frac{1}{\pi}) - n\eta - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \frac{1}{2}\sum_{i=1}^{n}\left\{1+\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-2\alpha(\theta)\cdot b_i\right)\frac{\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)}\right. \\
& \left.+\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}-\alpha(\theta)\cdot b_i\right)^2\right\}.
\end{aligned}
$$

where

$$\begin{aligned}
\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty} w_i^2\phi(w_i)dw_i &= -\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty} w_i\phi'(w_i)dw_i \\
&= -\left[\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)\cdot\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} - \int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\phi(w_i)dw_i\right] \\
&= -\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right)\cdot\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} + \Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right), \\
\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty} w_i\phi(w_i)dw_i &= -\int_{-\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}}^{\infty}\phi'(w_i)dw_i \\
&= \phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)}\right).
\end{aligned}$$

**M-Step**

In the maximisation step, we derive the first partial derivatives of the following conditional expectation $\Psi$ w.r.t. $\mu, \eta, \theta$ respectively. Then we find the maximum by setting these derivatives equal to 0. We write $\Psi(\mu,\eta,\theta \mid \hat{\mu}^{(v)}, \hat{\eta}^{(v)}, \hat{\theta}^{(v)})$ in detail here

$$\begin{aligned}
\Psi(\mu,\eta,\theta \mid \hat{\mu}^{(v)}, \hat{\eta}^{(v)}, \hat{\theta}^{(v)}) &= n\log\left(\frac{1}{\pi}\right) - n\eta - \frac{1}{2}\sum_{i=1}^{n} b_i^2 - \frac{1}{2}\sum_{i=1}^{n}\Big\{1 + \Big[\alpha(\hat{\theta}^{(v)})b_i^{(v)} - 2\alpha(\theta)b_i\Big] \\
&\quad \times\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} + \Big[\alpha(\hat{\theta}^{(v)})\cdot b_i^{(v)} - \alpha(\theta)\cdot b_i\Big]^2\Big\}, \\
b_i &= \frac{y_i-\mu}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}, \qquad b_i^{(v)} = \frac{y_i-\hat{\mu}^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\eta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}, \\
\sigma(\eta) &= e^{\eta}, \qquad \sigma(\hat{\eta}^{(v)}) = e^{\hat{\eta}^{(v)}}, \\
\delta(\theta) &= 1 - \frac{2}{1+e^{2\theta}}, \qquad \delta(\hat{\theta}^{(v)}) = 1 - \frac{2}{1+e^{2\hat{\theta}^{(v)}}}, \\
\alpha(\theta) &= \frac{1}{2}\left(e^{\theta} - e^{-\theta}\right), \qquad \alpha(\hat{\theta}^{(v)}) = \frac{1}{2}\left(e^{\hat{\theta}^{(v)}} - e^{-\hat{\theta}^{(v)}}\right).
\end{aligned}$$

To update the estimates of $\mu$, $\eta$ and $\theta$, we need to solve the following simultaneous equations

$$\frac{\partial\Psi}{\partial\mu} = 0, \qquad \frac{\partial\Psi}{\partial\eta} = 0, \qquad \frac{\partial\Psi}{\partial\theta} = 0.$$

Before calculating the derivatives on $\Psi$, we calculate in advance that

$$
\begin{aligned}
\frac{\partial b_i}{\partial \mu} &= -1/\sigma(\eta) = e^{-\eta}, \\
\frac{\partial b_i}{\partial \eta} &= -e^{-\eta}(y_i - \mu), \qquad \frac{\partial^2 b_i}{\partial \eta^2} = e^{-\eta}(y_i - \mu), \\
\frac{\partial b_i}{\partial \theta} &= \frac{\partial \delta(\theta)}{\partial \theta}\sqrt{2/\pi} = \frac{4e^{2\theta}}{(1+e^{2\theta})^2}\sqrt{2/\pi}, \\
\frac{\partial^2 b_i}{\partial \theta^2} &= \frac{\partial^2 \delta(\theta)}{\partial \theta^2}\sqrt{2/\pi} = \frac{8e^{2\theta}(1-e^{2\theta})}{(1+e^{2\theta})^3}\sqrt{2/\pi}, \\
\alpha(\theta) &= \frac{\delta(\theta)}{\sqrt{1-\delta(\theta)^2}}, \qquad \delta(\theta) = \frac{\alpha(\theta)}{\sqrt{1+\alpha(\theta)^2}}, \\
\frac{\partial \alpha(\theta)}{\partial \delta(\theta)} &= \frac{1}{\sqrt{1-\delta(\theta)^2}} - \frac{1}{2}\frac{-2\delta(\theta)^2}{(1-\delta(\theta)^2)^{3/2}} = \frac{1}{(1-\delta(\theta)^2)^{3/2}}, \\
\frac{\partial^2 \alpha(\theta)}{\partial \delta(\theta)^2} &= -\frac{3}{2}\frac{-2\delta(\theta)}{(1-\delta(\theta)^2)^{5/2}} = \frac{3\delta(\theta)}{(1-\delta(\theta)^2)^{5/2}}, \\
\frac{\partial \alpha(\theta)}{\partial \theta} &= \frac{1}{2}(e^{\theta} + e^{-\theta}), \qquad \frac{\partial^2 \alpha(\theta)}{\partial \theta^2} = \frac{1}{2}(e^{\theta} - e^{-\theta}) = \alpha(\theta), \\
\frac{\partial \delta(\theta)}{\partial \theta} &= \frac{4e^{2\theta}}{(1+e^{2\theta})^2} = 4\left(\frac{1}{1+e^{2\theta}} - \frac{1}{(1+e^{2\theta})^2}\right), \\
\frac{\partial^2 \delta(\theta)}{\partial \theta^2} &= 4\left(-\frac{2e^{2\theta}}{(1+e^{2\theta})^2} + \frac{4e^{2\theta}}{(1+e^{2\theta})^3}\right) = \frac{8e^{2\theta}(1-e^{2\theta})}{(1+e^{2\theta})^3}, \\
\frac{\partial 1/\sigma(\eta)}{\partial \eta} &= -e^{-\eta}.
\end{aligned}
$$

for the conciseness of notation we denote $\Psi(\mu, \eta, \theta \mid \hat{\mu}^{(v)}, \hat{\eta}^{(v)}, \hat{\theta}^{(v)})$ by using $\Psi$.

Derive the partial derivatives of $\Psi$ as follows:

$$
\begin{aligned}
\frac{\partial \Psi}{\partial \mu} &= -\sum_{i=1}^{n} b_i \frac{\partial b_i}{\partial \mu} + \sum_{i=1}^{n} \alpha(\theta) \frac{\partial b_i}{\partial \mu} \left( \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right) \\
&= \sum_{i=1}^{n} \frac{b_i}{\sigma(\eta)} - \sum_{i=1}^{n} \frac{\alpha(\theta)}{\sigma(\eta)} \left( \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right) \\
\frac{\partial \Psi}{\partial \eta} &= -n - \sum_{i=1}^{n} b_i \frac{\partial b_i}{\partial \eta} + \sum_{i=1}^{n} \alpha(\theta) \frac{\partial b_i}{\partial \eta} \left( \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right) \\
&= -n + e^{-\eta} \cdot \sum_{i=1}^{n} b_i \left(y_i - \mu\right) - \alpha(\theta) e^{-\eta} \cdot \sum_{i=1}^{n} \left(y_i - \mu\right) \\
&\quad \times \left[ \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right] \\
\frac{\partial \Psi}{\partial \theta} &= -\sum_{i=1}^{n} b_i \frac{\partial b_i}{\partial \theta} + \sum_{i=1}^{n} \left( \frac{\partial \alpha(\theta)}{\partial \theta} b_i + \alpha(\theta) \cdot \frac{\partial b_i}{\partial \theta} \right) \\
&\quad \times \left[ \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right] \\
&= -\frac{4e^{2\theta}}{(1+e^{2\theta})^2} \sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} \left[ \frac{1}{2} (e^{\theta} + e^{-\theta}) b_i + \alpha(\theta) \cdot \frac{4e^{2\theta}}{(1+e^{2\theta})^2} \sqrt{\frac{2}{\pi}} \right] \\
&\quad \times \left[ \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) b_i^{(v)} - \alpha(\theta) b_i \right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2\Psi}{\partial\theta^2} &= -\sum_{i=1}^{n}\left[\left(\frac{\partial b_i}{\partial\theta}\right)^2 + b_i\frac{\partial^2 b_i}{\partial\theta^2}\right] + \sum_{i=1}^{n}\left[\frac{\partial^2\alpha(\theta)}{\partial\theta^2}b_i + 2\frac{\partial\alpha(\theta)}{\partial\theta}\cdot\frac{\partial b_i}{\partial\theta} + \alpha(\theta)\cdot\frac{\partial^2 b_i}{\partial\theta^2}\right] \\
&\quad \times\left[\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})b_i^{(v)} - \alpha(\theta)b_i\right] - \left[\frac{\partial\alpha(\theta)}{\partial\theta}b_i + \alpha(\theta)\cdot\frac{\partial b_i}{\partial\theta}\right]^2 \\
&= -\sum_{i=1}^{n}\left\{\left(\frac{\partial b_i}{\partial\theta}\right)^2 + b_i\cdot\frac{\partial^2 b_i}{\partial\theta^2} + \left[\frac{\partial\alpha(\theta)}{\partial\theta}\cdot b_i + \alpha(\theta)\cdot\frac{\partial b_i}{\partial\theta}\right]^2\right\} \\
&\quad + \sum_{i=1}^{n}\left[\frac{\partial^2\alpha(\theta)}{\partial\theta^2}\cdot b_i + 2\frac{\partial\alpha(\theta)}{\partial\theta}\cdot\frac{\partial b_i}{\partial\theta} + \alpha(\theta)\cdot\frac{\partial^2 b_i}{\partial\theta^2}\right] \\
&\quad \times\left[\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})\cdot b_i^{(v)} - \alpha(\theta)\cdot b_i\right] \\
&= -\frac{32n\left(1+\alpha(\theta)^2\right)e^{4\theta}}{\pi\left(e^{2\theta}+1\right)^4} - \sqrt{\frac{2}{\pi}}\cdot\left[\frac{8e^{2\theta}\left(1-e^{2\theta}\right)}{\left(1+e^{2\theta}\right)^3} + \frac{2\alpha(\theta)\cdot e^{\theta}}{1+e^{2\theta}} + \frac{2\alpha(\theta)\cdot e^{\theta}}{1+e^{2\theta}}\right]\sum_{i=1}^{n} b_i \\
&\quad -\frac{\left(e^{2\theta}+1\right)^2}{4e^{2\theta}}\cdot\sum_{i=1}^{n} b_i^2 + \sum_{i=1}^{n}\left[\alpha(\theta)\cdot b_i - \sqrt{\frac{8}{\pi}}\frac{e^{\theta}}{\left(e^{2\theta}+1\right)^3}\left(e^{4\theta}-6e^{2\theta}+1\right)\right] \\
&\quad \times\left[\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})b_i^{(v)} - \alpha(\theta)b_i\right] \\
&= -\frac{8n\cdot e^{2\theta}}{\pi\left(1+e^{2\theta}\right)^2} - \sqrt{\frac{8}{\pi}}\delta(\theta)^3\sum_{i=1}^{n} b_i - \frac{1}{4}\left(e^{\theta}+e^{-\theta}\right)^2\sum_{i=1}^{n} b_i^2 + \sum_{i=1}^{n}\left[\alpha(\theta)\cdot b_i\right. \\
&\quad \left. -\sqrt{\frac{8}{\pi}}\cdot\frac{e^{\theta}\left(e^{4\theta}-6e^{2\theta}+1\right)}{\left(e^{2\theta}+1\right)^3}\right]\left[\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})\cdot b_i^{(v)} - \alpha(\theta)\cdot b_i\right].
\end{aligned}
$$

To solve these simultaneous equations, we adopt following iterative algorithm. Firstly, fixing $(\eta,\theta) = (\hat\eta^{(v)},\hat\theta^{(v)})$, we update $\mu$. It follows from $\frac{\partial\Psi}{\partial\mu} = 0$ that

$$
\mu = \bar y + \sigma(\eta)\cdot\delta(\theta)\cdot\sqrt{\frac{2}{\pi}} - \frac{\sigma(\eta)}{n}\cdot\frac{\alpha(\theta)}{1+\alpha(\theta)^2}\sum_{i=1}^{n}\left(\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})b_i^{(v)}\right).
$$

So, given $\theta = \hat\theta^{(v)}$ and $\eta = \hat\eta^{(v)}$, we can update $\mu$ by

$$
\begin{aligned}
\hat\mu^{(v+1)} &= \bar y + \sigma(\hat\eta^{(v)})\cdot\delta(\hat\theta^{(v)})\cdot\sqrt{\frac{2}{\pi}} - \frac{\sigma(\hat\eta^{(v)})}{n}\cdot\frac{\alpha(\hat\theta^{(v)})}{1+\alpha(\hat\theta^{(v)})^2} \\
&\quad \times\sum_{i=1}^{n}\left\{\frac{\phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat\theta^{(v)})b_i^{(v)}\right)} + \alpha(\hat\theta^{(v)})b_i^{(v)}\right\}.
\end{aligned}
$$

Secondly, fixing $(\mu, \theta) = (\hat{\mu}^{(v+1)}, \hat{\theta}^{(v)})$, we update $\eta$. It follows from $\frac{\partial \Psi}{\partial \eta} = 0$ that

$$\begin{aligned} e^{\eta} &= \left(1+\alpha(\theta)^2\right) \cdot \frac{1}{n}\sum_{i=1}^{n} b_i(y_i-\mu) - \alpha(\theta) \cdot \sum_{i=1}^{n}(y_i-\mu)\left(\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)})b_i^{(v)}\right) \\ &= \frac{1+\alpha(\theta)^2}{n}\sum_{i=1}^{n} e^{-\eta}(y_i-\mu)^2 + \left(1+\alpha(\theta)^2\right)\delta(\theta)\cdot\sqrt{2/\pi}(\bar{y}-\mu) \\ &\quad -\alpha(\theta)\cdot\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)\left(\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)})b_i^{(v)}\right) \end{aligned}$$

which implies that

$$e^{2\eta} - T(\mu, \theta, \boldsymbol{b}^{(v)}) \cdot e^{\eta} - \frac{1+\alpha(\theta)^2}{n}\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

with

$$\begin{aligned} T(\mu, \theta, \boldsymbol{b}^{(v)}) &= \left(1+\alpha(\theta)^2\right)\cdot\delta(\theta)\cdot\sqrt{\frac{2}{\pi}}(\bar{y}-\mu) - \frac{\alpha(\theta)}{n}\sum_{i=1}^{n}(y_i-\mu) \\ &\quad \times\left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)})b_i^{(v)}\right]. \end{aligned}$$

Solving the above quadratic equation, we obtain

$$\sigma(\eta) = e^{\eta} = \frac{1}{2}T(\mu, \theta, \boldsymbol{b}^{(v)}) + \sqrt{\frac{1}{4}T(\mu, \theta, \boldsymbol{b}^{(v)})^2 + \frac{1+\alpha(\theta)^2}{n}\sum_{i=1}^{n}(y_i-\mu)^2}.$$

So, we update $\eta$ (namely $\sigma = e^{\eta}$) via

$$\begin{aligned} \hat{\sigma}^{(v+1)} &= \sigma(\hat{\eta}^{(v+1)}) \\ &= e^{\hat{\eta}^{(v+1)}} \\ &= \frac{1}{2}T(\hat{\mu}^{(v+1)}, \hat{\theta}^{(v)}, \boldsymbol{b}^{(v)}) + \sqrt{\frac{1}{4}T(\hat{\mu}^{(v+1)}, \hat{\theta}^{(v)}, \boldsymbol{b}^{(v)})^2 + \frac{1+\alpha(\hat{\theta}^{(v)})^2}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}^{(v+1)})^2} \end{aligned}$$

where

$$\begin{aligned} T(\hat{\mu}^{(v+1)}, \hat{\theta}^{(v)}, \boldsymbol{b}^{(v)}) &= \left(1+\alpha(\hat{\theta}^{(v)})^2\right)\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}(\bar{y}-\hat{\mu}^{(v+1)}) \\ &\quad -\frac{\alpha(\hat{\theta}^{(v)})}{n}\sum_{i=1}^{n}(y_i-\hat{\mu}^{(v+1)})\left(\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} + \alpha(\hat{\theta}^{(v)})b_i^{(v)}\right). \end{aligned}$$

Finally, fixing $(\mu, \eta) = (\hat{\mu}^{(v+1)}, \hat{\eta}^{(v+1)})$, we update $\theta$ by using the following **Newton-Raphson**

**method**:

$$f(\theta) = \frac{\partial \Psi}{\partial \theta}, \qquad f'(\theta) = \frac{\partial^2 \Psi}{\partial \theta^2}.$$

To solve $f(\theta) = 0$, we iterate

$$\theta_{m+1} = \theta_m - \frac{f(\theta_m)}{f'(\theta_m)}$$

until $|\theta_{m+1} - \theta_m| < t$, where $t$ is the tolerance controlling the accuracy (default value of $t$ is $10^{-8}$).

### Initialisation

It is well known that the success of EM algorithm depends largely on good initial parameter values. The multimodality on likelihood could cause local optimum. To increase the chance of finding the global maximum or a decent local maximum, most of the time it is recommended to use multiple initial points (McLachlan and Peel, 2000). But it will overly increase the computation time when a prohibitively large number was applied. **Method of moments** has been widely used for the estimation of skew normal distribution. And the moments can be linked to Pearson's skewness index to help us express the shape parameter via the moments. Here this Pearson's skewness index $\gamma_1$ can be linked to second and third moments, after rearrangement, it can be used to set an initial value for $\theta$.

$$\gamma_1 = \frac{E\left[(Y-\mu)^3\right]}{\mathrm{var}(Y)^{3/2}} = \frac{4-\pi}{2}\frac{\delta^3(2/\pi)^{3/2}}{(1-2\delta^2/\pi)^{3/2}},$$

$$\delta = \sqrt{\frac{\pi}{2}}\frac{(2\gamma_1/(4-\pi))^{1/3}}{\sqrt{1+(2\gamma_1/(4-\pi))^{2/3}}}, \qquad \text{when } |\gamma_1| \leq 0.9952,$$

$$\theta = 0.5\log\left(\frac{1+\delta}{1-\delta}\right), \qquad \theta \in \mathbb{R}, \qquad \delta \in (-1,1).$$

Then our initialisation starts. Given observations $y_i \quad (i = 1, ..., n)$, we initialise the estimation as follows:

$$\begin{aligned}
\mu^{(0)} &= \bar{y} = \frac{1}{n}\sum y_i, \qquad S^2 = \frac{1}{n-1}\sum (y_i - \bar{y})^2, \qquad \gamma_1^{(0)} = \frac{\sum_{i=1}^n (y_i - \bar{y})^3 / n}{S^3}, \\
\delta^{(0)} &= \sqrt{\frac{\pi}{2}}\frac{\left(2\gamma_1^{(0)}/(4-\pi)\right)^{1/3}}{\sqrt{1+\left(2\gamma_1^{(0)}/(4-\pi)\right)^{2/3}}}, \qquad \text{when } \left|\gamma_1^{(0)}\right| \leq 0.9952, \\
\theta^{(0)} &= 0.5\log\left(\frac{1-\delta^{(0)}}{1+\delta^{(0)}}\right), \qquad \sigma^{(0)2} = \frac{S^2}{1-2\delta^{(0)2}/\pi}, \qquad \eta^{(0)} = \log\left(S/\sqrt{1-2\delta^{(0)2}/\pi}\right).
\end{aligned}$$

**Stopping rule**

In each update, we need to verify whether the incomplete data likelihood is increasing. When $\alpha(\hat{\theta}^{(v)})b_i^{(v)} < -37.519$, $\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)$ will be treated as exact 0 in R because values smaller than $10^{-308}$ (approximately) will be returned as 0 in R, then $\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)/\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)$ **overflows** in R. We need some way to avoid that problem when writing our programs. The following approximation can help us.

$$\frac{\phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right)} \approx \frac{\left|\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right|}{1-\left|\alpha(\hat{\theta}^{(v)})b_i^{(v)}\right|^{-2}}, \qquad \text{when } \alpha(\hat{\theta}^{(v)})b_i^{(v)} < -37.519.$$

Then the EM algorithm iteration alternates between **E-step** and **M-step** until:

$$\left|\frac{l_{inc}(\hat{\mu}^{(v+1)},\hat{\eta}^{(v+1)},\hat{\theta}^{(v+1)} \mid \mathbf{y}) - l_{inc}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)} \mid \mathbf{y})}{l_{inc}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)} \mid \mathbf{y})}\right| < \varepsilon$$

where $\varepsilon$ is the tolerance ($10^{-8}$ is default value). And $l_{inc}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)} \mid \mathbf{y})$ is the **incomplete-data log-likelihood** function calculated after $v$-th iteration:

$$\begin{aligned} l_{inc}(\hat{\mu}^{(v)},\hat{\eta}^{(v)},\hat{\theta}^{(v)} \mid \mathbf{y}) &= \sum_{i=1}^{n} \log\left\{\frac{2}{\sigma(\hat{\eta}^{(v)})}\phi\left(\frac{y_i-\hat{\mu}^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right. \\ &\quad \left.\times\Phi\left(\alpha(\hat{\theta}^{(v)})\left(\frac{y_i-\hat{\mu}^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)\right\}. \end{aligned}$$

This modelling has also been extended to **skew normal regression** and **penalised skew normal regression** (for high-dimensional problems). Owing to some similarities to this part, both of these two additional parts with simulation studies are attached in **Appendix C**.

## 4.3 $\alpha$-Penalised Estimation

As mentioned in the introduction of this chapter, there is peculiar behaviour for the likelihood function when the true parameter $\alpha$ is near 0. We explain it by **Fisher information matrix** of $l_{inc}(\mu,\sigma,\alpha \mid \mathbf{y})$ at $\alpha = 0$, which can be written as

$$\mathcal{I}\left((\mu,\sigma,\alpha=0)^T\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 & 0 \\ 0 & \frac{2n}{\sigma^2} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{4.3.1}$$

The **orthogonal feature** of $\mu$ and $\sigma$ could help the inference as mentioned in Section 4.1. As we can see, for (4.3.1) the determinant is 0 here. This singularity of expected information

causes trouble for the estimation at $\alpha = 0$, making it hard to estimate $\alpha$ when the true value is close to 0. Thus, before the start of the aforementioned EM algorithm, we use D'Agostino's K-squared test (D'Agostino, 1970) to test the null hypothesis that $\alpha = 0$. For the $\alpha = 0$ case, we just need to fix $\alpha = 0$ at the start of our EM algorithm (no $\theta^{(v)}$ updating anymore) and carry on other updatings. Otherwise, we follow the aforementioned EM algorithm to the end.

A more natural way to overcome it is introduced here. We can bring in an $\alpha$-penalised term $\lambda\alpha^2$ to our **incomplete-data log-likelihood** function, that is

$$\begin{aligned} l^{\star}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) &= l_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) - \lambda\alpha^2 \\ &= \frac{n}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2\alpha^2}\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n}\log\left[\Phi(A_i)\right] - \lambda\alpha^2 \end{aligned} \tag{4.3.2}$$

where $\delta = \alpha/\sqrt{1+\alpha^2}$ and $A_i = \alpha\left(\frac{y_i-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)$. The tuning constant $\lambda$ is selected from a grid of candidates by a 10-fold cross-validation step (see Section 4.3.1). Then the Fisher information function at $\alpha = 0$ for $l^{\star}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y})$ is

$$\mathcal{I}\left((\mu, \sigma, \alpha = 0)^T\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 & 0 \\ 0 & \frac{2n}{\sigma^2} & 0 \\ 0 & 0 & 2\lambda \end{pmatrix}.$$

The determinant of the Fisher information matrix is not 0 anymore, which means the likelihood-based inference can be obtained smoothly. From this modification, we can reveal some interesting findings in Section 4.3.2 below. Based on these findings, some intriguing theorems could be obtained in Section 4.3.3 below. The proofs of these theorems are attached in **Appendix B**. The asymptotic analysis of this $\alpha$-penalised skew normal model can help us have a better understanding of the estimation results.

### 4.3.1 Cross-Validation for Selecting $\lambda$

The k-fold cross-validation has been introduced in Section 2.6.2. Here we use it ($k = 10$) to select the tuning constant $\lambda$ for our $\alpha$-penalised skew normal modelling. Divide the given observations $\mathbf{y}$ into 10 groups $(G_1, G_2, \ldots, G_{10})$ and validate by computing $l^{\star}_{inc}(\mu, \sigma, \alpha \mid G_j)$ with estimates $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ from the estimation of training set $G_{-j}$ (log-likelihood $l^{\star}_{inc}(\mu, \sigma, \alpha \mid G_{-j})$), and repeat it for 10 times with $j = 1, 2, \ldots 10$ respectively.

Here $l^{\star}_{inc}(\mu,\sigma,\alpha \mid G_j)$ and $l^{\star}_{inc}(\mu,\sigma,\alpha \mid G_{-j})$ are

$$\begin{aligned}
l^{\star}_{inc}(\mu,\sigma,\alpha \mid G_j) &= \frac{n_j}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2\alpha^2}\sum_{i=1}^{n_j} A_i^2 + \sum_{i=1}^{n_j}\log\left[\Phi(A_i)\right] - \lambda\alpha^2, \\
A_i &= \alpha\left(\frac{G_{j,i}-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right), \\
n_j &: \text{ cardinal number of } G_j; \\
l^{\star}_{inc}(\mu,\sigma,\alpha \mid G_{-j}) &= \frac{n-n_j}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2\alpha^2}\sum_{i=1}^{n-n_j} A_i^{\star 2} + \sum_{i=1}^{n-n_j}\log\left[\Phi(A_i^{\star})\right] - \lambda\alpha^2, \\
A_i^{\star} &= \alpha\left(\frac{G_{-j,i}-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right), \\
n &: \text{ cardinal number of } \boldsymbol{y}.
\end{aligned}$$

Generally the candidate grid $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ is determined through some pilot trials to compromise between the accuracy of $\lambda$ and computation speed of the computer. Algorithm 4.1 uses the pseudocode to show the whole process where the 'update' is realised through EM algorithm as we introduced in Section 4.2. But we need to pay attention that as a result of introducing quadratic penalty term the updating for $\theta$ has changed to

$$f(\theta) = \frac{\partial\Psi}{\partial\theta} - \frac{\lambda}{2}\left(e^{2\theta} - e^{-2\theta}\right), \qquad f'(\theta) = \frac{\partial^2\Psi}{\partial\theta^2} - \lambda\left(e^{2\theta} + e^{-2\theta}\right).$$

Here, the introduction of $\lambda\alpha^2$ term will only affect the updating for $\theta$.

### 4.3.2 Asymptotic Properties of Cross-Validated $\lambda_{CV}$

We want to investigate the behaviour of $\lambda_{CV}$ from cross-validation under different underlying scenarios. We carried out some simulations to find the asymptotic behaviours about $\lambda_{CV}$. For these simulations, we sampled with the underlying true parameter values $(\tilde{\mu}=0, \tilde{\sigma}=1, \tilde{\alpha}=0,2,3,4)$ respectively, and performed $\lambda_{CV}$ selection through 10-fold cross-validation under different sample sizes $n$s for each $\tilde{\alpha}$ setting. Then these simulations were replicated many times to produce the box plots and trend charts. From these graphs, we found some interesting asymptotic properties about this $\lambda_{CV}$. After that, we used some inequalities to verify these properties empirically. The following simulations are specified as four cases to investigate $\lambda_{CV}$'s behaviours under true $\tilde{\alpha}=0$ and true $\tilde{\alpha}\neq 0$.

*Case* 1. Underlying true parameter settings are $(\tilde{\mu}=0, \tilde{\sigma}=1, \tilde{\alpha}=0)$, for different sample size $n$, analyse the behaviour of $\lambda_{CV}/n$.

*Case* 2. Underlying true parameter settings are $(\tilde{\mu}=0, \tilde{\sigma}=1, \tilde{\alpha}=3)$, for different sample size $n$, analyse the behaviour of $\lambda_{CV}/\sqrt{n}$.

**Algorithm 4.1** The 10-fold cross-validation process to select $\lambda_{CV}$

Input: Data points $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, a grid of $\lambda$ candidates $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$.
Output: The log likelihood table with $(A_1, A_2, \ldots)$ for $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$, and the best $\lambda$ from them.
  Divide $\mathbf{y}$ into 10 groups $(G_1, G_2, \ldots, G_{10})$
  for i in $1, 2, \ldots$
   for j in 1:10
    $\boldsymbol{G}_{-j} = (G_1, G_2, \ldots, G_{j-1}, G_{j+1}, \ldots, G_{10})$
    moment estimates $\mu_0, \sigma_0, \alpha_0$ as initial values
     while $\left(\left|l^{\star}_{inc}\left(\hat{\mu}^{(v)}, \hat{\sigma}^{(v)}, \hat{\alpha}^{(v)} \mid \boldsymbol{G}_{-j}\right) - l^{\star}_{inc}\left(\hat{\mu}^{(v-1)}, \hat{\sigma}^{(v-1)}, \hat{\alpha}^{(v-1)} \mid \boldsymbol{G}_{-j}\right)\right| > 10^{-8}$ and iteration_times$\leq 500$)
      update $\hat{\mu}^{(v-1)}$ to $\hat{\mu}^{(v)}$ with $\hat{\mu}^{(v-1)}, \hat{\sigma}^{(v-1)}, \hat{\alpha}^{(v-1)}, \boldsymbol{G}_{-j}$
      update $\hat{\sigma}^{(v-1)}$ to $\hat{\sigma}^{(v)}$ through $\eta$ with $\hat{\mu}^{(v)}, \hat{\sigma}^{(v-1)}, \hat{\alpha}^{(v-1)}, \boldsymbol{G}_{-j}$
      update $\hat{\alpha}^{(v-1)}$ to $\hat{\alpha}^{(v)}$ through $\theta$ with $\hat{\mu}^{(v)}, \hat{\sigma}^{(v)}, \hat{\alpha}^{(v-1)}, \boldsymbol{G}_{-j}$ and $\lambda_i$
      calculate incomplete log-likelihood $l^{\star}_{inc}\left(\hat{\mu}^{(v)}, \hat{\sigma}^{(v)}, \hat{\alpha}^{(v)} \mid \boldsymbol{G}_{-j}\right)$
      $v = v + 1$
     end while
     calculate validation log-likelihood $l_1 = l^{\star}_{inc}(\hat{\mu}, \hat{\sigma}, \hat{\alpha} \mid G_j)$
   end for
   calculate $A_i$ (the average of $l_1, l_2, \ldots, l_{10}$)
  end for
  Choose the maximum $A_{max}$ from $A_1, A_2, \ldots$
  return the best $\lambda_{CV}$ from $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$

*Case* 3. Underlying true parameter settings are $(\tilde{\mu} = 0, \tilde{\sigma} = 1, \tilde{\alpha} = 2)$, for different sample size $n$, analyse the behaviour of $\lambda_{CV}/\sqrt{n}$.

*Case* 4. Underlying true parameter settings are $(\tilde{\mu} = 0, \tilde{\sigma} = 1, \tilde{\alpha} = 4)$, for different sample size $n$, analyse the behaviour of $\lambda_{CV}/\sqrt{n}$.

For each case we set $n = 50, 100, 200, 300, 400, 500, 600, 1000$ respectively, and for each sample size $n$ we performed 20 simulations which means we re-sampled and replicated the whole cross-validation process for 20 times.

Figure 4.3.1 displays the boxplots and trend chart about $\lambda_{CV}/n$ for **Case 1** as $n$ increases. The trend chart indicates that under true $\tilde{\alpha} = 0$ the average of statistic $\lambda_{CV}/n$ manifests a trend of converging to a constant $c_0$ (about 0.0035 here) and variance of $\lambda_{CV}/n$ tends to 0 with the increase of sample size $n$.

Figure 4.3.2 displays the boxplots and trend chart about $\lambda_{CV}/\sqrt{n}$ for **Case 2** as $n$ increases. The trend chart indicates that under true $\tilde{\alpha} \neq 0$ the average and variance of statistic $\lambda_{CV}/\sqrt{n}$ both manifest a trend of converging to 0 with the increase of sample size $n$. Figure 4.3.3 for **Case 3** and Figure 4.3.4 for **Case 4** can provide more support for this idea.

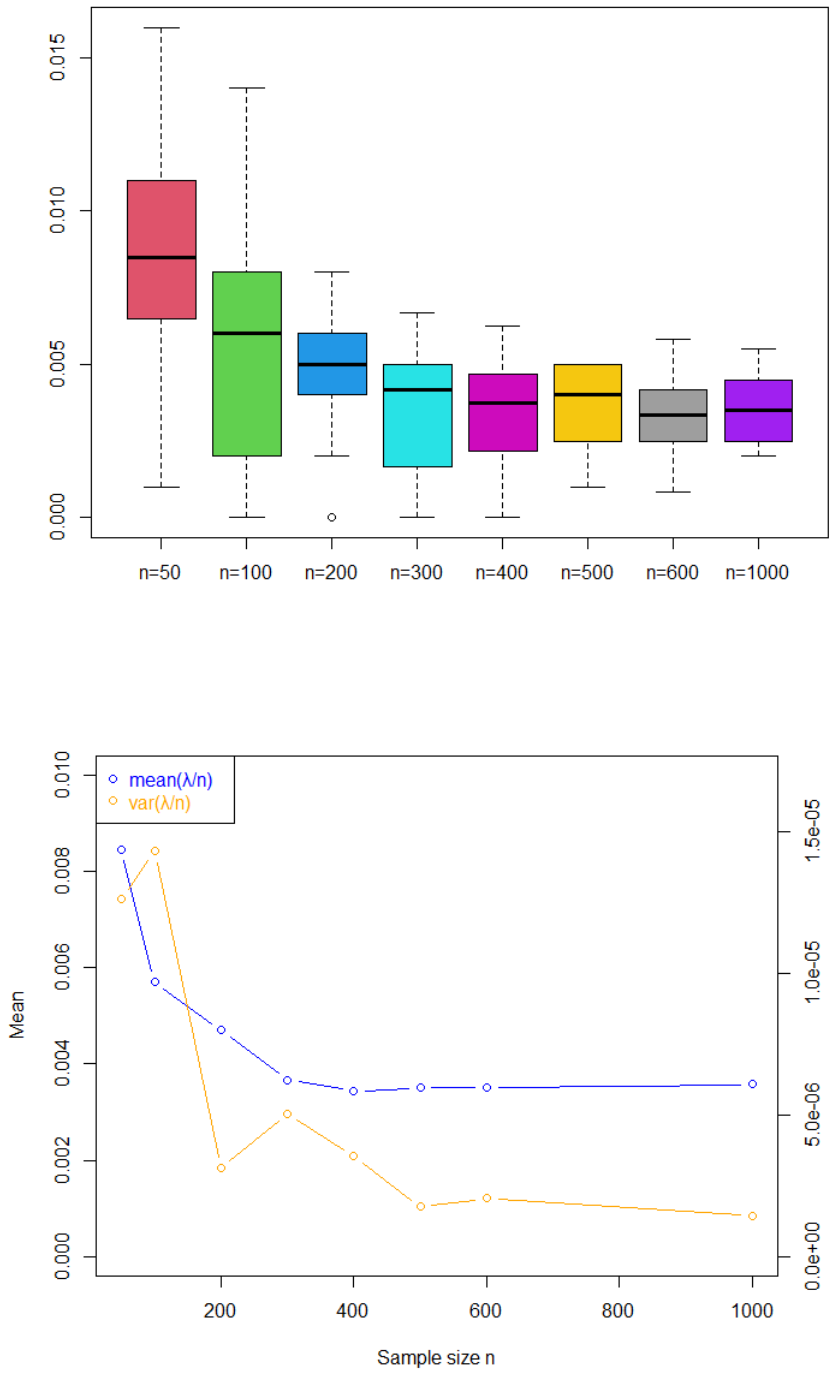
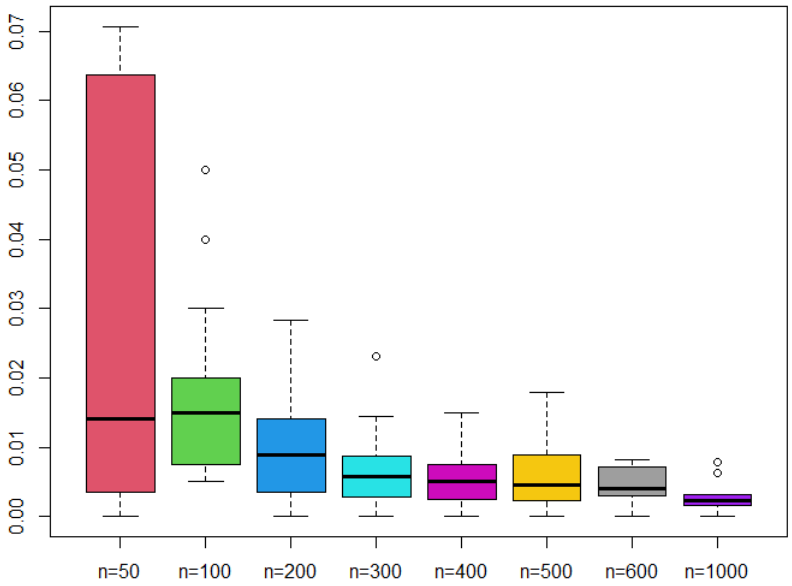



Figure 4.3.1: The box plots for $\lambda_{CV}/n$ and trend chart for average (variance) of $\lambda_{CV}/n$ under **Case 1**. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/n$ as $n$ increases.



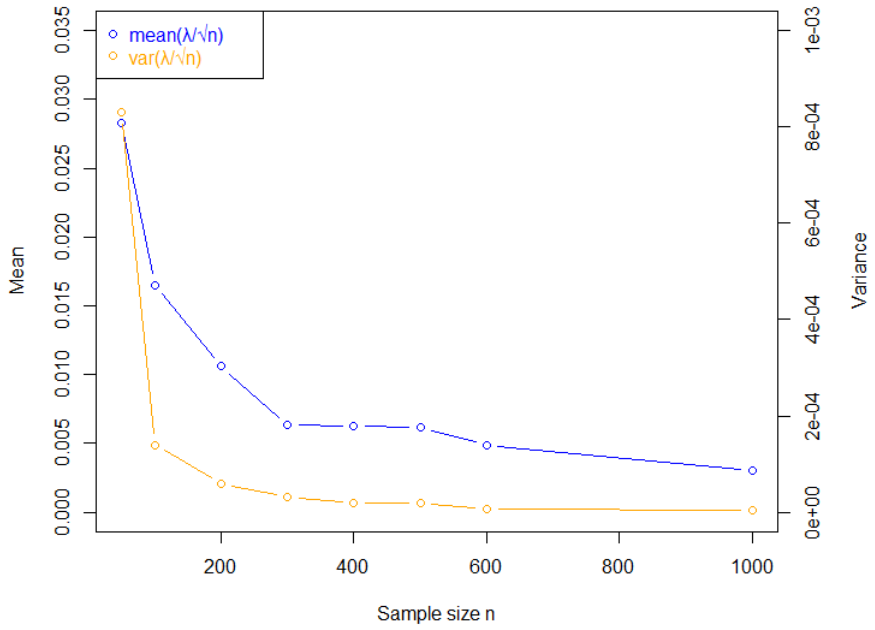




Figure 4.3.2: The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 2. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases.

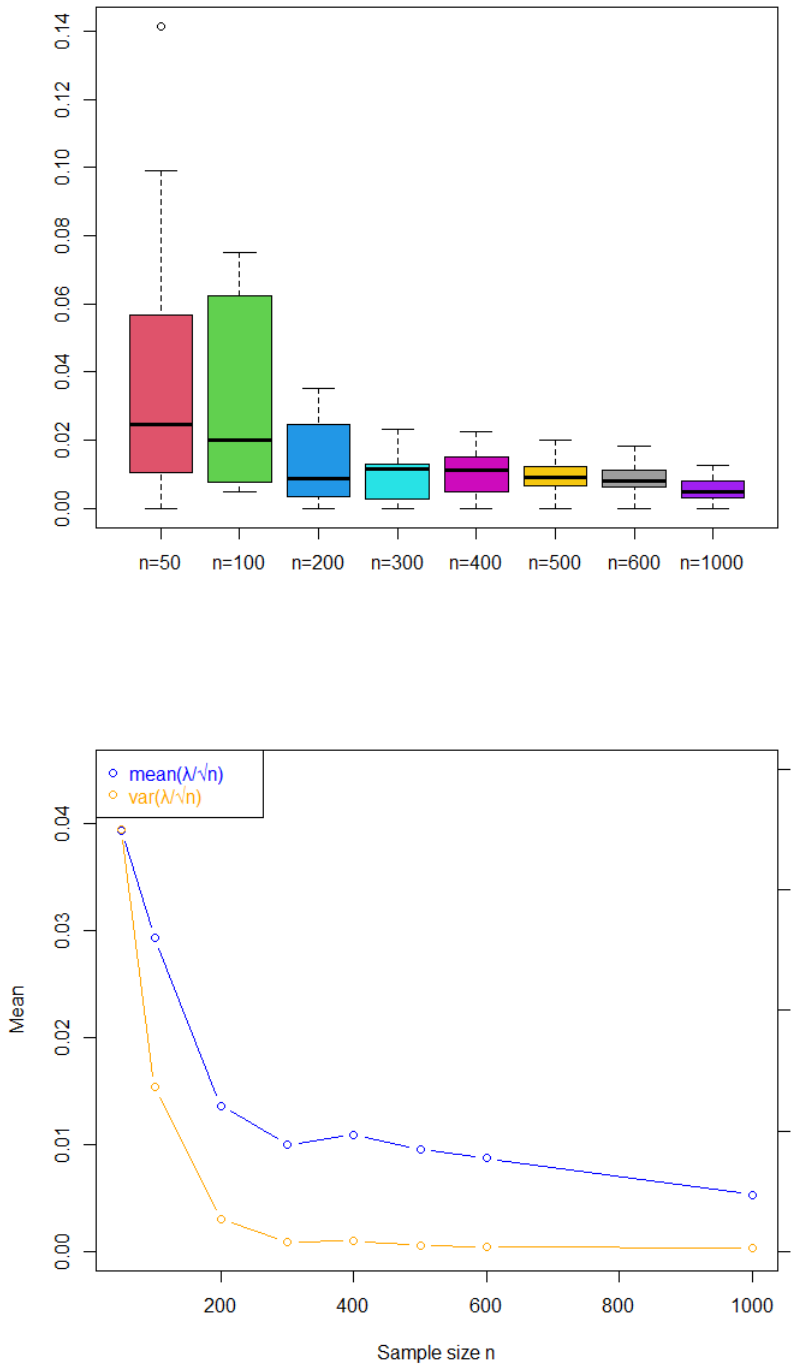



Figure 4.3.3: The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 3. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases.

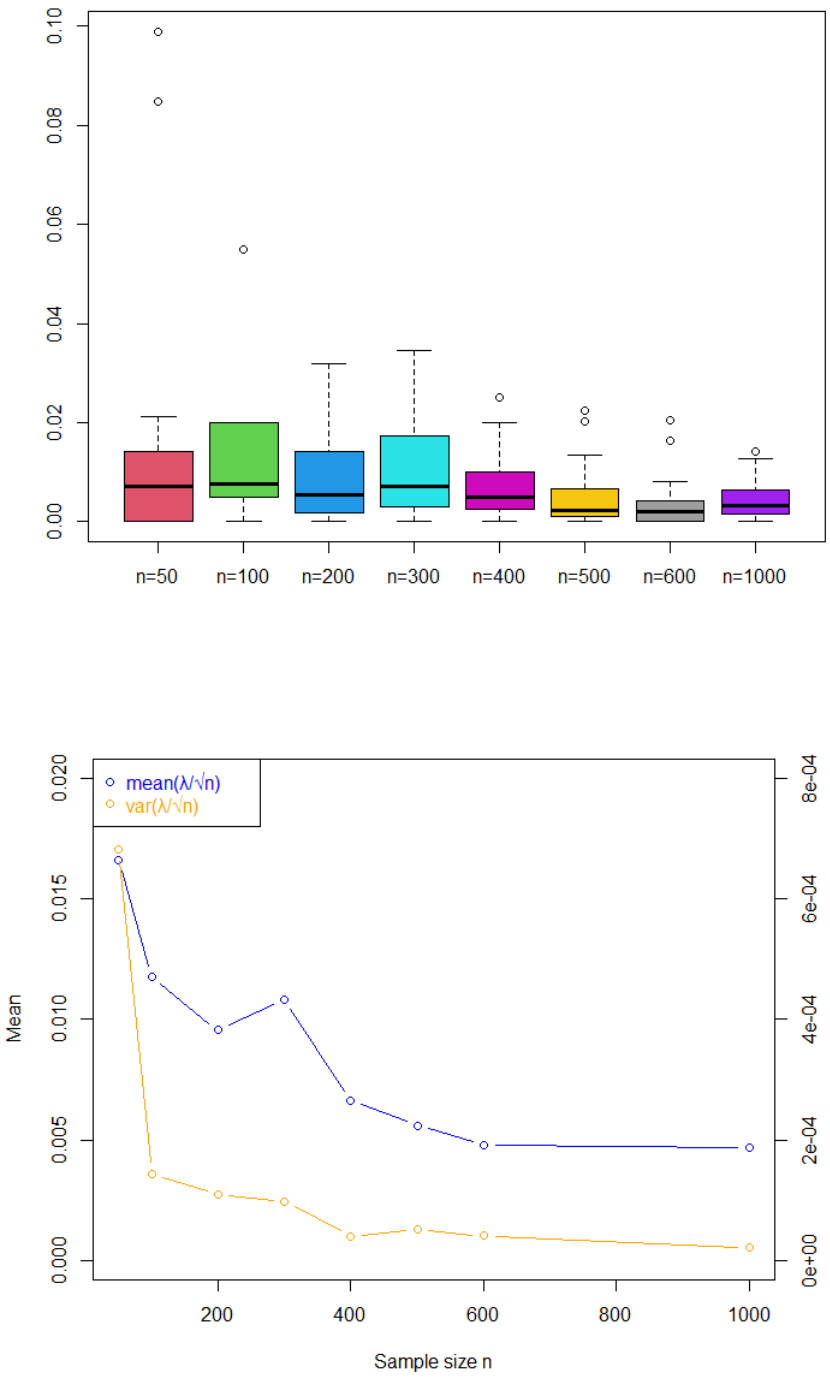



Figure 4.3.4: The box plots for $\lambda_{CV}/\sqrt{n}$ and trend chart for average (variance) of $\lambda_{CV}/\sqrt{n}$ under **Case** 4. Blue line plots different averages and orange line plots different variances for $\lambda_{CV}/\sqrt{n}$ as $n$ increases.

**$\lambda_{CV}$ has the following asymptotic properties. $\frac{\lambda_{CV}}{n} \to c_0 > 0$ in probability when the true value of $\alpha$ is 0, and $\frac{\lambda_{CV}}{\sqrt{n}} \to 0$ in probability when the true value of $\alpha$ is not 0 as $n \to \infty$ respectively.**

Now we use inequalities to verify them. We have

$$\begin{aligned}
\Pr\left(\left|\frac{\lambda_{CV}}{n} - c_0\right| \geq \epsilon\right) &\leq \Pr\left(\left|\frac{\lambda_{CV}}{n} - E\left(\frac{\lambda}{n}\right)\right| + \left|E\left(\frac{\lambda_{CV}}{n}\right) - c_0\right| \geq \epsilon\right) \\
&= \Pr\left(\left|\frac{\lambda_{CV}}{n} - E\left(\frac{\lambda}{n}\right)\right| \geq \epsilon - \left|E\left(\frac{\lambda_{CV}}{n}\right) - c_0\right|\right) \\
&\leq \frac{\mathrm{Var}\left(\frac{\lambda_{CV}}{n}\right)}{\left(\epsilon - \left|E\left(\frac{\lambda_{CV}}{n}\right) - c_0\right|\right)^2} \quad \text{(Chebyshev's inequality)}.
\end{aligned}$$

Judging from Figure 4.3.1, we can use the results that $\lim_{n\to+\infty} E\left(\frac{\lambda_{CV}}{n}\right) = c_0$ and $\lim_{n\to+\infty} \mathrm{Var}\left(\frac{\lambda_{CV}}{n}\right) = 0$. Then using Slutsky's theorem, we can reach that $\lim_{n\to+\infty} \Pr\left(\left|\frac{\lambda_{CV}}{n} - c_0\right| \geq \epsilon\right) = 0$ for any $\epsilon > 0$. Therefore, it holds that $\frac{\lambda_{CV}}{n} \to c_0 > 0$ in probability as $n \to \infty$ under true parameter value $\tilde{\alpha} = 0$.

And we also have known

$$\begin{aligned}
\Pr\left(\left|\frac{\lambda_{CV}}{\sqrt{n}}\right| \geq \varepsilon\right) &= \Pr\left(\left(\frac{\lambda_{CV}}{\sqrt{n}}\right)^2 \geq \varepsilon^2\right) \\
&\leq \frac{E\left[\left(\frac{\lambda_{CV}}{\sqrt{n}}\right)^2\right]}{\varepsilon^2} \quad \text{(Markov's inequality)} \\
&= \frac{\mathrm{Var}\left(\frac{\lambda_{CV}}{\sqrt{n}}\right) + \left[E\left(\frac{\lambda_{CV}}{\sqrt{n}}\right)\right]^2}{\varepsilon^2}.
\end{aligned}$$

Judging from Figure 4.3.2, Figure 4.3.3 and Figure 4.3.4, we use the results that $\lim_{n\to+\infty} E\left(\frac{\lambda_{CV}}{\sqrt{n}}\right) = 0$ and $\lim_{n\to+\infty} \mathrm{Var}\left(\frac{\lambda_{CV}}{\sqrt{n}}\right) = 0$. Again with Slutsky's theorem we can reach that $\lim_{n\to+\infty} \Pr\left(\left|\frac{\lambda_{CV}}{\sqrt{n}}\right| \geq \varepsilon\right) = 0$ for any $\varepsilon > 0$. Therefore, it also holds that $\frac{\lambda_{CV}}{\sqrt{n}} \to 0$ in probability as $n \to \infty$ under true parameter value $\tilde{\alpha} \neq 0$.

### 4.3.3 Asymptotic Distributions of $\alpha-$Penalised MLE

In our EM algorithm, we use the **method of moments estimates** as initial estimates of the parameters. It is known that these method of moments estimates are consistent with rates $\kappa_{\mu n}, \kappa_{\sigma n}, \kappa_{\alpha n}$ respectively. In light of this, we consider the following restricted area that $|\hat{\mu} - \tilde{\mu}| \leq \kappa_{\mu n} \to 0$, $|\hat{\sigma} - \tilde{\sigma}| \leq \kappa_{\sigma n} \to 0$ , $|\hat{\alpha} - \tilde{\alpha}| \leq \kappa_{\alpha n} \to 0$, where $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha})$ are the underlying true parameter values, to improve the performance of our estimates by assuming all good estimates are within this restricted area. In Section 4.3.2, we have found that the $\lambda_{CV}$ selected from our cross-validation has the asymptotic properties $\frac{\lambda_{CV}}{n} \to c_0 > 0$ in probability under

the true parameter value $\tilde{\alpha} = 0$ and $\frac{\lambda_{CV}}{\sqrt{n}} \to 0$ in probability under the true parameter value $\tilde{\alpha} \neq 0$ as $n \to \infty$ respectively. Provided with these assumptions, some asymptotic results can be summarised in the following theorems with detailed proofs attached in **Appendix B**.

In Theorem 1 below, we show the asymptotic distribution when the true parameter value $\tilde{\alpha}$ is 0.

**Theorem 1.** *Let $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ be the restricted maximum likelihood estimate when $(\mu, \sigma, \alpha)$ are restricted to $|\mu - \tilde{\mu}| \leq \kappa_{\mu n} \to 0$, $|\sigma - \tilde{\sigma}| \leq \kappa_{\sigma n} \to 0$ , $|\alpha - 0| \leq \kappa_{\alpha n} \to 0$. Here $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha} = 0)$ are the underlying true parameter values. With $n \to \infty$, assume $\frac{\lambda_{CV}}{n} \to c_0 > 0$ ($c_0$ is a positive constant) in probability. Then the restricted MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ for the $\alpha$-penalised log-likelihood function $l^{\star}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y})$ have the following properties that as $n \to \infty$*

$$\sqrt{n}\left(\frac{\hat{\mu} - \tilde{\mu}}{\tilde{\sigma}}\right) \xrightarrow{d} N(0,1), \qquad \sqrt{n}\left(\frac{\hat{\sigma} - \tilde{\sigma}}{\tilde{\sigma}}\right) \xrightarrow{d} N\left(0, \frac{1}{2}\right),$$

$$\sqrt{n}\left(\frac{\hat{\alpha} - 0}{\tilde{\sigma}}\right) = (\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n}) O_p(1) \to 0.$$

In Theorem 2 below, we show the asymptotic likelihood-ratio when the true parameter value $\tilde{\alpha}$ is 0.

**Theorem 2.** *Let $(\hat{\mu}^*, \hat{\sigma}^*, 0)$ be the maximum likelihood estimate when fixing $\alpha = 0$ and $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ be the restricted maximum likelihood estimate when $(\mu, \sigma, \alpha)$ are restricted to $|\mu - \tilde{\mu}| \leq \kappa_{\mu n} \to 0$, $|\sigma - \tilde{\sigma}| \leq \kappa_{\sigma n} \to 0$ , $|\alpha - 0| \leq \kappa_{\alpha n} \to 0$. Here $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha} = 0)$ are the underlying true parameter values. With $n \to \infty$, assume $\frac{\lambda_{CV}}{n} \to c_0 > 0$ ($c_0$ is a positive constant) in probability. Then the likelihood ratio (mostly used for likelihood-ratio test) has the following property that as $n \to \infty$*

$$l^{\star}_{inc}(\hat{\mu}, \hat{\sigma}, \hat{\alpha} \mid \mathbf{y}) - l^{\star}_{inc}(\hat{\mu}^*, \hat{\sigma}^*, 0 \mid \mathbf{y}) = O_p\left(n^{-1/2}\right) + (\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n}) \cdot O_p(1) \to 0.$$

In Theorem 3 below, we show the asymptotic distribution when the true parameter value $\tilde{\alpha}$ is not 0.

**Theorem 3.** *Let $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ be the restricted maximum likelihood estimate when $(\mu, \sigma, \alpha)$ are restricted to $|\mu - \tilde{\mu}| \leq \kappa_{\mu n} \to 0$, $|\sigma - \tilde{\sigma}| \leq \kappa_{\sigma n} \to 0$ , $|\alpha - \tilde{\alpha}| \leq \kappa_{\alpha n} \to 0$. Here $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha})$ are the underlying true parameter values and $\tilde{\alpha} \neq 0$. With $n \to \infty$, assume $\frac{\lambda_{CV}}{\sqrt{n}} \to 0$ in probability. Then the restricted MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ for the $\alpha$-penalised log-likelihood function $l^{\star}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y})$ have the following properties that as $n \to \infty$*

$$\sqrt{n}(\hat{\mu} - \tilde{\mu}) \xrightarrow{d} N\left(0, \omega_1^2\right), \qquad \sqrt{n}(\hat{\sigma} - \tilde{\sigma}) \xrightarrow{d} N\left(0, \omega_2^2\right), \qquad \sqrt{n}(\hat{\alpha} - \tilde{\alpha}) \xrightarrow{d} N\left(0, \omega_3^2\right),$$

*where*

$$
\begin{aligned}
\omega_1^2 \;=\; & \left\{ -\frac{\left(\pi+3\pi\tilde{\alpha}^2+3\pi\tilde{\alpha}^4+\pi\tilde{\alpha}^6+\tilde{\alpha}^6\right)}{0.25\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^6}+\frac{4\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)^2\left[\pi+(2\pi-1)\,\tilde{\alpha}^2+(\pi-3)\,\tilde{\alpha}^4\right]}{\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^5}\right. \\
& \times E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[\pi+(3\pi-2)\,\tilde{\alpha}^2+3\,(\pi-2)\,\tilde{\alpha}^4+(\pi-3)\,\tilde{\alpha}^6\right]E_{21} \\
& +\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}\cdot\left[\pi+(3\pi-4)\,\tilde{\alpha}^2+(3\pi-8)\,\tilde{\alpha}^4+(\pi-3)\,\tilde{\alpha}^6\right]E_{22} \\
& \left.+\frac{2\tilde{\alpha}^4\left(2+\tilde{\alpha}^2\right)^2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}\left(E_{20}E_{22}-E_{21}^2\right)\right\}\left\{\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+(4-\pi)\,\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{0.25\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\right. \\
& +\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+(13\pi-4)\,\tilde{\alpha}^2+(22\pi-20)\,\tilde{\alpha}^4+(18\pi-34)\,\tilde{\alpha}^6+(7\pi-20)\,\tilde{\alpha}^8\right. \\
& \left.+(\pi-4)\,\tilde{\alpha}^{10}\right]E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+(7\pi-4)\,\tilde{\alpha}^2+(9\pi-12)\,\tilde{\alpha}^4+(5\pi-8)\,\tilde{\alpha}^6\right. \\
& \left.+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21}+\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+(3\pi-4)\,\tilde{\alpha}^2+(3\pi-8)\,\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22} \\
& +\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\cdot\left[-\pi-3\pi\tilde{\alpha}^2-(2+3\pi)\,\tilde{\alpha}^4-(2+\pi)\,\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}} \\
& \left.\times\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right)\cdot E_{20}E_{21}+\frac{2\tilde{\alpha}^2\left[\pi+3\pi\tilde{\alpha}^2+(3\pi-4)\,\tilde{\alpha}^4+(\pi-2)\,\tilde{\alpha}^6\right]}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}E_{20}E_{22}\right\}^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\omega_2^2 \;=\; & \left\{ -\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}+\frac{2\tilde{\alpha}^2\left(3+4\tilde{\alpha}^2+\tilde{\alpha}^4\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}E_{20}+\frac{2\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{3/2}}E_{21}+\frac{1}{\tilde{\sigma}^2}E_{22}\right. \\
& \left.+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{20}E_{22}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{21}^2\right\}\left\{\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+(4-\pi)\,\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{0.25\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\right. \\
& +\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+(13\pi-4)\,\tilde{\alpha}^2+(22\pi-20)\,\tilde{\alpha}^4+(18\pi-34)\,\tilde{\alpha}^6+(7\pi-20)\,\tilde{\alpha}^8\right. \\
& \left.+(\pi-4)\,\tilde{\alpha}^{10}\right]E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+(7\pi-4)\,\tilde{\alpha}^2+(9\pi-12)\,\tilde{\alpha}^4+(5\pi-8)\,\tilde{\alpha}^6\right. \\
& \left.+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21}+\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+(3\pi-4)\,\tilde{\alpha}^2+(3\pi-8)\,\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22} \\
& +\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[-\pi-3\pi\tilde{\alpha}^2-(2+3\pi)\,\tilde{\alpha}^4-(2+\pi)\,\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}} \\
& \left.\times\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right)E_{20}E_{21}+\frac{2\tilde{\alpha}^2\left[\pi+3\pi\tilde{\alpha}^2+(3\pi-4)\,\tilde{\alpha}^4+(\pi-2)\,\tilde{\alpha}^6\right]}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}E_{20}E_{22}\right\}^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\omega_3^2 \;=\; & \left\{ \frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+(3\pi-1)\,\tilde{\alpha}^2+(3\pi-4)\,\tilde{\alpha}^4+(\pi-4)\,\tilde{\alpha}^6\right]+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^2}\right. \\
& \left.\times\left[\pi+(2\pi-1)\,\tilde{\alpha}^2+(\pi-3)\,\tilde{\alpha}^4\right]E_{20}-\frac{2\sqrt{2/\pi}\tilde{\alpha}^5}{\tilde{\sigma}^4}E_{21}+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^4}E_{22}+\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{20}E_{22}-\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{21}^2\right\} \\
& \left\{ 4\cdot\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+(4-\pi)\,\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}+\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+(13\pi-4)\right.\right. \\
& \left.\times\tilde{\alpha}^2+(22\pi-20)\,\tilde{\alpha}^4+(18\pi-34)\,\tilde{\alpha}^6+(7\pi-20)\,\tilde{\alpha}^8+(\pi-4)\,\tilde{\alpha}^{10}\right]E_{20} \\
& +\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+(7\pi-4)\,\tilde{\alpha}^2+(9\pi-12)\,\tilde{\alpha}^4+(5\pi-8)\,\tilde{\alpha}^6+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21} \\
& +\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+(3\pi-4)\,\tilde{\alpha}^2+(3\pi-8)\,\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22}+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3} \\
& \times\left[-\pi-3\pi\tilde{\alpha}^2-(2+3\pi)\,\tilde{\alpha}^4-(2+\pi)\,\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right) \\
& \left.\left.\times E_{20}E_{21}+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+3\pi\tilde{\alpha}^2+(3\pi-4)\,\tilde{\alpha}^4+(\pi-2)\,\tilde{\alpha}^6\right]E_{20}E_{22}\right\}^{-1}, \qquad (4.3.3)
\end{aligned}
$$

$$
E_{20}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right],\quad E_{21}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right],\quad E_{22}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right].
$$

We have obtained the asymptotic distributions for the MLE of skew normal distribution under true parameter value $\tilde{\alpha}\neq 0$. To our best knowledge, it has not been found in the literature about the final calculation results on the variances of these asymptotic distributions. These variance values can help the numerical calculation to find the confidence interval for the MLE under different known skew normal distribution settings such as the 95% confidence interval $\left(\tilde{\alpha}-\frac{1.96\omega_3}{\sqrt{n}},\tilde{\alpha}+\frac{1.96\omega_3}{\sqrt{n}}\right)$ for $\alpha$. For example, when the underlying true parameter values are $(\tilde{\mu}=0,\ \tilde{\sigma}=1,\ \tilde{\alpha}=2)$, the 95% confidence interval for shape parameter $\alpha$ is (0.704, 3.296).

## 4.4 Simulation Studies

In the following simulation studies, first of all, we showed a general estimation accuracy for our location-scale modelling. Then with the further introduction of the quadratic penalty term, we showed how the estimation accuracy is dramatically improved when the underlying true $\tilde{\alpha}$ is 0. After that, we compared the quadratic penalty with another penalty system from the literature to show the superiority of our method. Finally, with a comprehensive box plot collection, we showed the performances of our method under a wide range of scenarios.

### 4.4.1 The Simulation Study for Location-Scale Skew Normal

In this simulation, through *rsn* function from package *sn* in R, $n=300$ observations were sampled under $\left(\xi=\mu-\sigma\delta\sqrt{\frac{2}{\pi}},\ \omega=\sigma,\ \alpha=\alpha\right)$ where $\mu=2,\ \sigma=0.5,\ \alpha=1.5,\ \delta=\alpha/\sqrt{1+\alpha^2}=$

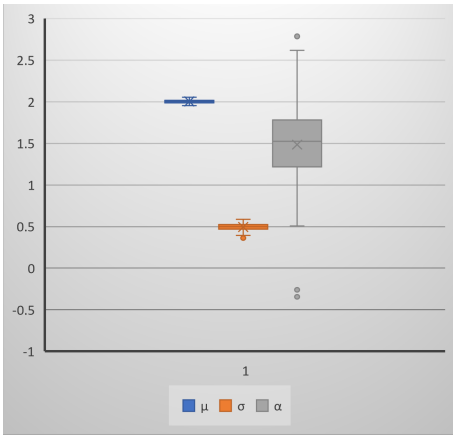



Figure 4.4.1: The box plots for estimated location parameter $\mu$, scale parameter $\sigma$ and shape parameter $\alpha$ after 100 times of resampling and estimation with our location-scale skew normal modelling under true parameter values $\tilde{\mu} = 2,\ \tilde{\sigma} = 0.5,\ \tilde{\alpha} = 1.5$ with sample size $n = 300$. The averages of these parameter estimates (marked with crosses in the boxes) are $\bar{\mu} = 2.003,\ \bar{\sigma} = 0.495,\ \bar{\alpha} = 1.485$.

0.832. Here $\xi$ is the location parameter after centralisation. The centralisation is realised by

$$\begin{aligned} Y &= \xi + \sigma Z \\ &= \mu - \sigma\delta\sqrt{2/\pi} + \sigma Z \\ &= \mu + \sigma\left(Z - \delta\sqrt{2/\pi}\right) \end{aligned}$$

where $Z \sim \mathrm{SN}(0, 1, \alpha)$ follows the standard skew normal distribution. And the expectation of $Z$ is $\delta\sqrt{2/\pi}$. After the centralisation, the density function becomes

$$\begin{aligned} g\left(y; \mu, \sigma, \alpha\right) &= \frac{2}{\sigma}\phi\left(\frac{y-\xi}{\sigma}\right)\Phi\left(\alpha\left(\frac{y-\xi}{\sigma}\right)\right) \\ &= \frac{2}{\sigma}\phi\left(\frac{y-\left(\mu-\sigma\delta\sqrt{2/\pi}\right)}{\sigma}\right)\Phi\left(\alpha\left(\frac{y-\left(\mu-\sigma\delta\sqrt{2/\pi}\right)}{\sigma}\right)\right) \\ &= \frac{2}{\sigma}\phi\left(\frac{y-\mu}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y-\mu}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\right). \end{aligned}$$

Under the underlying true parameter settings, the simulation data were resampled 100 times and each time the estimation process with aforementioned methodology was performed. The box plot regarding these 100 results is shown in Figure 4.4.1.

### 4.4.2 The Simulation Studies for $\alpha$-Penalised Skew Normal

To address the peculiar estimation behaviour when underlying true parameter value $\tilde{\alpha} = 0$, we introduced the $\alpha$-penalised skew normal likelihood function on $n$ observations in Section 4.3:

$$\begin{aligned} l^{\star}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) &= l_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) - \lambda\alpha^2 \\ &= \frac{n}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2\alpha^2}\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n}\log\left[\Phi(A_i)\right] - \lambda\alpha^2, \end{aligned} \tag{4.4.1}$$

where $A_i = \alpha\left(\frac{y_i-\mu}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)$.

Azzalini (2014) used centred parameters (CP) reparameterisation to deal with the singularity problem when true parameter value $\tilde{\alpha} = 0$ and did a good asymptotic analysis on this reparameterisation system but lack of simulation support for $\tilde{\alpha}$ under different value levels. Moreover, a penalised log-likelihood for skew normal was considered by Azzalini and Arellano-Valle (2012), where $l_p(\theta) = l(\theta) - Q(\theta)$ with penalty $Q(\theta) = c_1\log\left(1 + c_2\alpha^2\right)$ was put forward to avoid the divergent behaviour of estimate $\hat{\alpha}$ for a formulation applicable to a wide situations. To check the performance of this penalty term, we applied it to our location-scale skew normal likelihood function:

$$\begin{aligned} l^{\dagger}_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) &= l_{inc}(\mu, \sigma, \alpha \mid \mathbf{y}) - c_1\log\left(1 + c_2\alpha^2\right) \\ &= \frac{n}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2\alpha^2}\sum_{i=1}^{n} A_i^2 + \sum_{i=1}^{n}\log\left[\Phi(A_i)\right] - c_1\log\left(1 + c_2\alpha^2\right). \end{aligned} \tag{4.4.2}$$

But in this penalty system $c_1$ and $c_2$ were fixed as constants ($c_1 \approx 0.875913$, $c_2 \approx 0.856250$). Different from that, we choose $\lambda_{CV}$ by $10-$folded cross-validation which is much more flexible with considering the possible underlying true value $\tilde{\alpha}$. As we know, when the true value $\tilde{\alpha}$ is close to 0, the tuning constant $\lambda$ should be larger compared with the case where the true one is away from 0. To observe the ability of $\lambda\alpha^2$ on penalising $\alpha$, we used the aforementioned EM algorithm to estimate (4.4.1), the simulation result is illustrated through box plots in Figure 4.4.2.

Arellano-Valle and Azzalini (2008) explained that it is the unusual feature of non-quadratic shape of the log-likelihood function at the stationary point $\alpha = 0$ that makes it hard to estimate shape parameter $\alpha$ when the underlying real one is 0. In Figure 4.4.2, we have shown that with this penalty term $\lambda\alpha^2$, shape parameter $\alpha$ can be estimated properly via our EM algorithm when it is 0 in the underlying skew normal distribution. For comparison, the estimation result from the model (4.4.1) without penalty ($\lambda = 0$) is also put in Figure 4.4.2.

To check the performance when $\alpha$ is away from 0, we set the underlying true $\tilde{\mu} = 0$, $\tilde{\sigma} = 1$

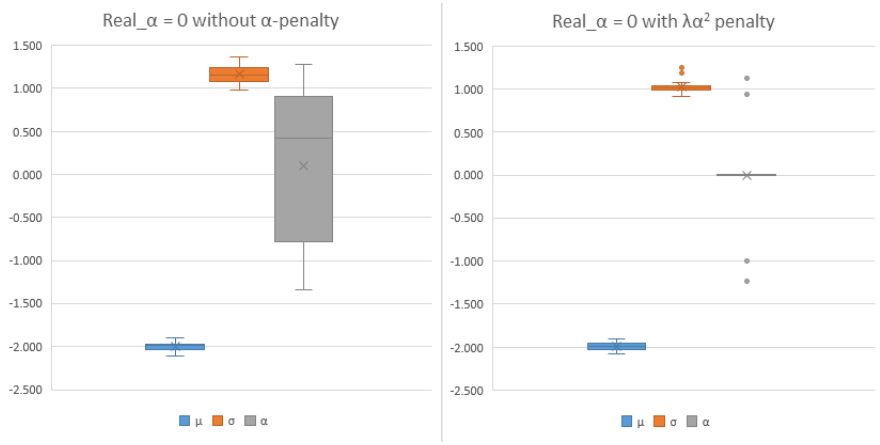



Figure 4.4.2: Box plots for the simulation estimation result of skew normal data sampled from $SN(-2, 1, 0)$, based on sample size $n = 400$, re-sample and replicate 50 times. The left one did not use penalty ($\lambda = 0$). The right one used $\lambda\alpha^2$ penalty and each time $\lambda_{CV}$ is selected from a grid of candidates by 10−folded cross validation.

and $\tilde{\alpha} = 1, 3, 5$ under $n = 50, 100, 200$ respectively. And for each setting, we also re-sample and replicate the trials 50 times, and then calculate the mean bias, median bias and standard deviation for it. To compare with it, we also performed our EM algorithm for (4.4.2). The summary under our $\lambda\alpha^2$ penalty (left block) and the $c_1 \log\left(1 + c_2\alpha^2\right)$ penalty (right block) for the estimation of each parameter is put in Table 4.1. From this table, we can see that the skew normal density function with $c_1 \log\left(1 + c_2\alpha^2\right)$ penalty showed a much higher standard deviation in some cases when estimating $\alpha$. The bias for $\hat{\alpha}$ under $c_1 \log\left(1 + c_2\alpha^2\right)$ penalty is also uncompetitive compared to our result. And the corresponding results under our $\lambda\alpha^2$ penalty for these estimates under all 9 different settings ($\tilde{\alpha} = 1, 3, 5$ and $n = 50, 100, 200$ ) are illustrated with box plots in Figure 4.4.3.

## 4.5 Anti-Cancer Drug Data Application

For $\mathbf{Y}_{397\times187}$ the anti-cancer drug data from Chapter 3, we are curious about the inherent skewness of the data for each drug. These data consist of responses (measured by half maximal inhibitory concentration) to different cell lines. There are 187 anti-cancer drugs in total for the dataset $\mathbf{Y}_{397\times187}$. We applied the $\alpha$-penalised skew normal model onto each anti-cancer drug, then use aforementioned EM algorithm to find the estimates $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$. Figure 4.5.1 shows the scatter plot and histogram for all 187 $\hat{\alpha}$s. The detailed results for each drug can refer to **Appendix C** Figure 6.2.4. When $\alpha = 1.5$, we can calculate the Fisher's moment coefficient of skewness $\gamma_1 \approx 0.3$. There are 131 out of 187 drugs featuring the IC50 data with evident skewnesses that $|\gamma_1| > 0.3$.

Take 84th drug 'Pelitinib' as an example to perform a bootstrapping for it. The estimation

| $\tilde{\alpha}$ | $n$ | Quantity | under $\hat{\mu}$ | $\lambda\alpha^2$ $\hat{\sigma}$ | penalty $\hat{\alpha}$ | under $\hat{\mu}$ | $c_1 \log\left(1+c_2\alpha^2\right)$ $\hat{\sigma}$ | penalty $\hat{\alpha}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 50 | Mean bias | -0.0121 | -0.0499 | -0.6868 | -0.0085 | -0.0761 | -0.2569 |
| | | Median bias | -0.0171 | -0.1009 | -0.9951 | -0.0128 | -0.1280 | -0.9979 |
| | | Std.dev. | 0.1083 | 0.2187 | 1.4872 | 0.1301 | 0.1960 | 3.9374 |
| | 100 | Mean bias | -0.0032 | -0.0432 | -0.4939 | 0.0288 | -0.0776 | -0.5650 |
| | | Median bias | -0.0022 | -0.0516 | -0.5718 | 0.0224 | -0.1415 | -0.9956 |
| | | Std.dev. | 0.0740 | 0.1593 | 0.9795 | 0.0909 | 0.1379 | 0.8608 |
| | 200 | Mean bias | 0.0068 | -0.0481 | -0.3837 | -0.0028 | -0.1029 | -0.7150 |
| | | Median bias | 0.0032 | -0.0529 | -0.3137 | -0.0037 | -0.1440 | -0.9910 |
| | | Std.dev. | 0.0514 | 0.1312 | 0.8040 | 0.0492 | 0.1082 | 0.6986 |
| 3 | 50 | Mean bias | -0.0058 | -0.0765 | 0.2319 | -0.0229 | -0.1429 | 1.3152 |
| | | Median bias | -0.0082 | -0.0696 | -0.5578 | -0.0448 | -0.0961 | -0.8977 |
| | | Std.dev. | 0.0766 | 0.1595 | 2.4054 | 0.0956 | 0.2364 | 7.3542 |
| | 100 | Mean bias | -0.0105 | -0.0558 | -0.2522 | -0.0160 | -0.0710 | 1.3614 |
| | | Median bias | 0.0026 | -0.0458 | -0.3852 | -0.0112 | -0.0714 | -0.5395 |
| | | Std.dev. | 0.0689 | 0.1238 | 1.1349 | 0.0619 | 0.1746 | 6.8711 |
| | 200 | Mean bias | 0.0041 | -0.0116 | -0.2255 | -0.0035 | -0.0224 | 1.3475 |
| | | Median bias | -0.0038 | -0.0127 | -0.4833 | 0.0039 | -0.0225 | -0.3582 |
| | | Std.dev. | 0.0532 | 0.0801 | 0.9482 | 0.0434 | 0.1426 | 6.1749 |
| 5 | 50 | Mean bias | -0.0240 | -0.0727 | -0.5477 | -0.0185 | -0.0614 | 4.0295 |
| | | Median bias | -0.0212 | -0.0417 | -0.4806 | -0.0138 | -0.0538 | -0.8666 |
| | | Std.dev. | 0.1020 | 0.1567 | 2.2759 | 0.1022 | 0.2268 | 10.1609 |
| | 100 | Mean bias | 0.0002 | -0.0511 | -0.4073 | -0.0096 | -0.0030 | 4.4940 |
| | | Median bias | -0.0120 | -0.0227 | -0.9159 | -0.0058 | -0.0011 | -0.4098 |
| | | Std.dev. | 0.0713 | 0.1110 | 2.1078 | 0.0665 | 0.1500 | 9.7276 |
| | 200 | Mean bias | 0.0007 | -0.0265 | -0.0150 | -0.0046 | -0.0020 | 1.9516 |
| | | Median bias | -0.0007 | -0.0310 | -0.4611 | -0.0127 | -0.0120 | -0.6123 |
| | | Std.dev. | 0.0388 | 0.0713 | 1.8044 | 0.0485 | 0.0950 | 6.9836 |

Table 4.1: The summary for skew normal models with our $\lambda\alpha^2$ penalty (left block) and $c_1 \log\left(1+c_2\alpha^2\right)$ penalty (right block) respectively when they are estimated with our EM algorithm under distribution $SN(0,1,\alpha)$ when true parameter $\tilde{\alpha}=1,3,5$, based on sample size $n=50,100,200$. Each entry is based on 50 replicative trials.

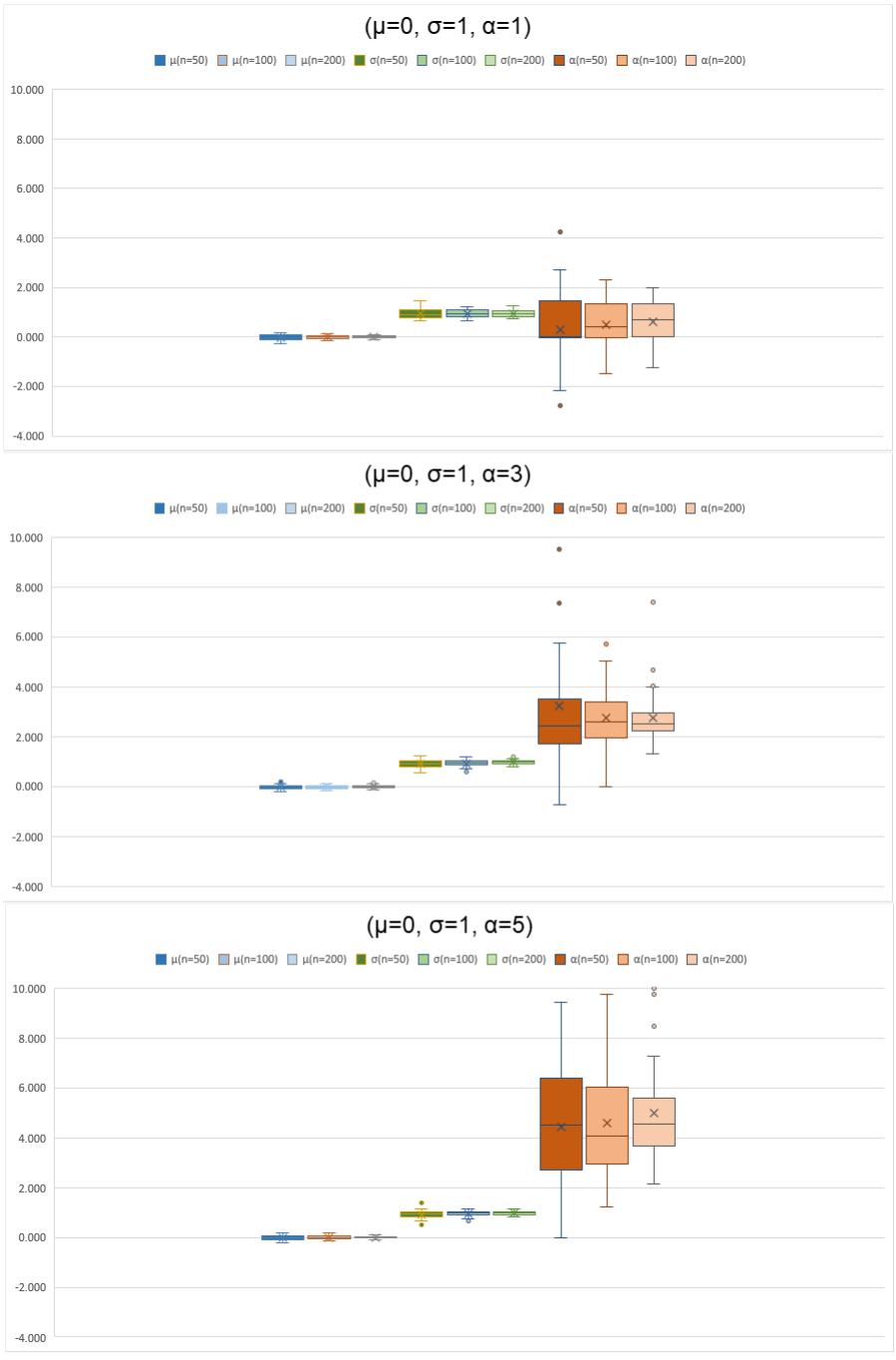
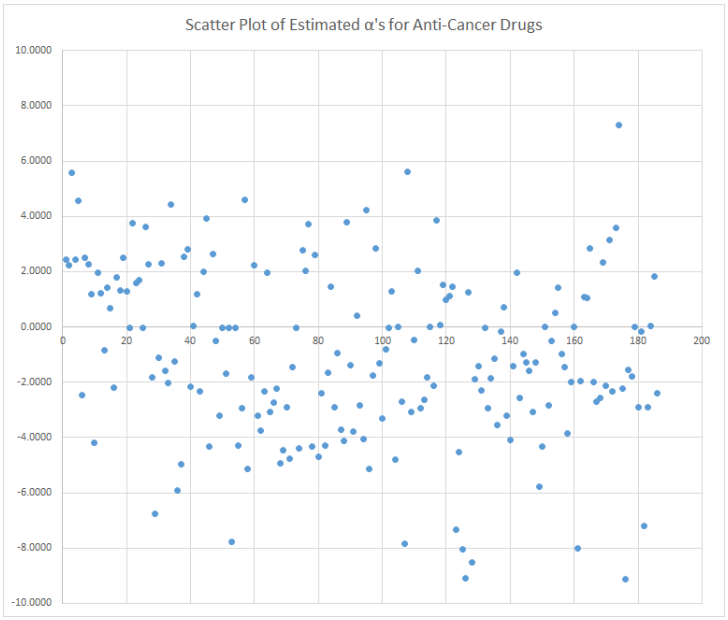



Figure 4.4.3: The box plots correspond to our $\lambda\alpha^2$ penalty result area in Table 4.1. There are three different settings for skew normal distribution skewness parameter $\alpha$ (they are $1, 3, 5$), and three different sample size $n$ (they are $50, 100, 200$) for each $\alpha$.



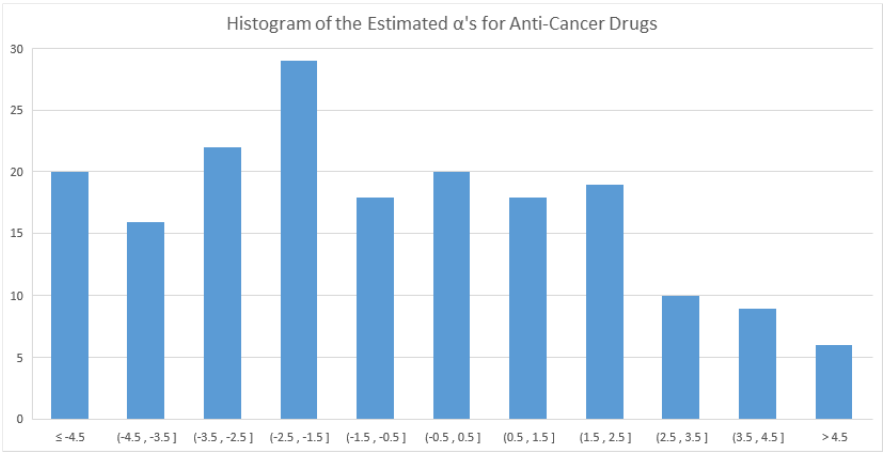




Figure 4.5.1: The scatter plot and histogram of $\hat{\alpha}$s for 187 anti-cancer drugs. The scatter plot shows the value of $\hat{\alpha}$ for each drug from 1st drug to 186th drug. For the convenience of the layout, the drug 'Sepantronium.bromide' is removed from the scatter plot, as it has an oversized $\hat{\alpha} = 27.8853$. Refer to **Appendix C** Figure 6.2.4 for details.

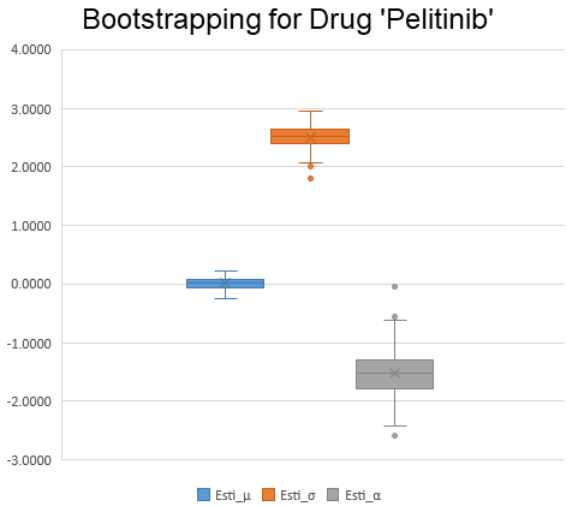



Figure 4.5.2: The box plots for bootstrapping of drug 'Pelitinib'. Resample $n = 397$ samples from $SN\,(0.00742,\ 2.56255,\ -1.64613)$ and perform estimations for 200 times to produce these box plots.

result for 'Pelitinib' is $(\hat{\mu} = 0.00742,\ \hat{\sigma} = 2.56255,\ \hat{\alpha} = -1.64613)$ (as shown in **Appendix C** Figure 6.2.4). Repeatedly sample $n = 397$ observations from the skew normal distribution $SN\,(\mu = 0.00742,\ \sigma = 2.56255,\ \alpha = -1.64613)$ for 200 times and each time estimate it with our $\alpha$-penalised modelling and its EM algorithm. The box plot for these 200 trial estimation results is shown in Figure 4.5.2. For $\hat{\alpha}$, the interval including 95% estimation results is $(-2.256,\ -0.735)$ among these 200 trials. And $\hat{\alpha} = 0$ (skewness does not exist) is not included in this interval. If we calculate the 95% confidence interval with $\omega_3$ formula (4.3.3) and $\left(\tilde{\alpha} - \frac{1.96\omega_3}{\sqrt{n}},\ \tilde{\alpha} + \frac{1.96\omega_3}{\sqrt{n}}\right)$ by assuming $\tilde{\alpha} = -1.64613$, then we can reach $(-2.550,\ -0.742)$ which is close to our bootstrapping result $(-2.256,\ -0.735)$.

# Chapter 5

# Estimating Skew Normal Mixtures

## 5.1 Four-Piece Distribution Family

In R package *mclust* (Fraley and Raftery, 2002), a kind of model-based clustering method is introduced, where equal variance and unequal variances are specified as two different models in the estimation process. Inspired by this, a kind of model-based mixture skew normal clustering method is put forward in this chapter. The scale parameter $\sigma$ and shape (skewness) parameter $\alpha$ are discussed separately regarding whether $\sigma$s are equal and $\alpha$s are equal for all components. Then four-piece distribution family is introduced here, where four different models are considered in the density function construction. They are ESVA (equal $\sigma$ and variable $\alpha$), VSEA (variable $\sigma$ and equal $\alpha$), ESEA (equal $\sigma$ and equal $\alpha$) and VSVA (variable $\sigma$ and variable $\alpha$). After estimation with these four different models, the best model is determined as the one with the best Bayesian Information Criterion value.

### 5.1.1 Density and Likelihood

Suppose that we have $n$ observations $(y_1, y_2, \ldots, y_n)$ on random variable $Y$ which comes from a finite mixture model that is the $K-$component mixture of skew normal densities. We also suppose that any information about the origin (group), which each observation comes from, is unknown to us. Then the **mixture density functions** for the four-piece distribution family

can be given by

$$g\left(y_i;\boldsymbol{\Theta}\right)=\begin{cases}\sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma}\phi\left(\frac{y_i-\mu_k}{\sigma}+\delta_k\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha_k\left(\frac{y_i-\mu_k}{\sigma}+\delta_k\sqrt{\frac{2}{\pi}}\right)\right), & \text{ESVA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma_k}\phi\left(\frac{y_i-\mu_k}{\sigma_k}+\delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y_i-\mu_k}{\sigma_k}+\delta\sqrt{\frac{2}{\pi}}\right)\right), & \text{VSEA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma}\phi\left(\frac{y_i-\mu_k}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y_i-\mu_k}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\right), & \text{ESEA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma_k}\phi\left(\frac{y_i-\mu_k}{\sigma_k}+\delta_k\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha_k\left(\frac{y_i-\mu_k}{\sigma_k}+\delta_k\sqrt{\frac{2}{\pi}}\right)\right), & \text{VSVA}\end{cases}$$

$$\delta=\alpha/\sqrt{1+\alpha^2},\qquad \delta_k=\alpha_k/\sqrt{1+\alpha_k^2},$$

where $\boldsymbol{\pi}_{1:K}=(\pi_1,\ldots,\pi_K)$ are the mixing proportions with constraints $\pi_i\geq 0$ for $i=1,2,\ldots,K$ and $\sum\limits_{i=1}^{K}\pi_i=1$. And $\boldsymbol{\Theta}=(\Theta_1,\ldots,\Theta_K)$ is the parameter set with $\Theta_k=(\pi_k,\mu_k,\sigma_k,\alpha_k)$ (for VSVA) being specific parameters for $k$th component where $\Theta_k=(\pi_k,\mu_k,\sigma,\alpha_k)$ (for ESVA), $\Theta_k=(\pi_k,\mu_k,\sigma_k,\alpha)$ (for VSEA) and $\Theta_k=(\pi_k,\mu_k,\sigma,\alpha)$ (for ESEA) are also included in it.

Once again, to facilitate our next EM algorithm calculation, we introduce the reparametrisation as Chapter 4 to our location-scale skew normal mixture model. Then the density functions for the four-piece distribution family can be written as

$$g\left(y_i;\boldsymbol{\Theta}^{\star}\right)=\begin{cases}\sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma(\eta)}\phi\left(\frac{y_i-\mu_k}{\sigma(\eta)}+\delta(\theta_k)\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\theta_k)\left(\frac{y_i-\mu_k}{\sigma(\eta)}+\delta(\theta_k)\sqrt{\frac{2}{\pi}}\right)\right), & \text{ESVA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma(\eta_k)}\phi\left(\frac{y_i-\mu_k}{\sigma(\eta_k)}+\delta(\theta)\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\theta)\left(\frac{y_i-\mu_k}{\sigma(\eta_k)}+\delta(\theta)\sqrt{\frac{2}{\pi}}\right)\right), & \text{VSEA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma(\eta)}\phi\left(\frac{y_i-\mu_k}{\sigma(\eta)}+\delta(\theta)\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\theta)\left(\frac{y_i-\mu_k}{\sigma(\eta)}+\delta(\theta)\sqrt{\frac{2}{\pi}}\right)\right), & \text{ESEA}\\ \sum\limits_{k=1}^{K}\pi_k\cdot\frac{2}{\sigma(\eta_k)}\phi\left(\frac{y_i-\mu_k}{\sigma(\eta_k)}+\delta(\theta_k)\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\theta_k)\left(\frac{y_i-\mu_k}{\sigma(\eta_k)}+\delta(\theta_k)\sqrt{\frac{2}{\pi}}\right)\right), & \text{VSVA}\end{cases}$$

where

$$\begin{aligned}\boldsymbol{\Theta}^{\star} &= \left(\Theta_1^{\star},\ldots,\Theta_K^{\star}\right),\\ \Theta_k^{\star} &= \left(\pi_k,\mu_k,\eta,\theta_k\right), \qquad \text{ESVA}\\ \Theta_k^{\star} &= \left(\pi_k,\mu_k,\eta_k,\theta\right), \qquad \text{VSEA}\\ \Theta_k^{\star} &= \left(\pi_k,\mu_k,\eta,\theta\right), \qquad \text{ESEA}\\ \Theta_k^{\star} &= \left(\pi_k,\mu_k,\eta_k,\theta_k\right), \qquad \text{VSVA}\\ \sigma(\eta) &= e^{\eta}, \qquad \sigma(\eta_k)=e^{\eta_k},\\ \delta(\theta) &= 1-\frac{2}{1+e^{2\theta}}, \qquad \delta(\theta_k)=1-\frac{2}{1+e^{2\theta_k}},\\ \alpha(\theta) &= \frac{1}{2}\left(e^{\theta}-e^{-\theta}\right), \qquad \alpha(\theta_k)=\frac{1}{2}\left(e^{\theta_k}-e^{-\theta_k}\right).\end{aligned}$$

Then we get the **incomplete-data log-likelihood** functions:

$$l_{inc}\left(\boldsymbol{\Theta}^{\star} \mid \mathbf{y}\right) = \begin{cases} \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma(\eta)} \cdot \phi\left(\frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}}\right)\right. \\ \left. \times \Phi\left(\alpha(\theta_k) \cdot \left(\frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}}\right)\right)\right), & \text{ESVA} \\ \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma(\eta_k)} \cdot \phi\left(\frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}}\right)\right. \\ \left. \times \Phi\left(\alpha(\theta) \cdot \left(\frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}}\right)\right)\right), & \text{VSEA} \\ \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma(\eta)} \cdot \phi\left(\frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}}\right)\right. \\ \left. \times \Phi\left(\alpha(\theta) \cdot \left(\frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}}\right)\right)\right), & \text{ESEA} \\ \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma(\eta_k)} \cdot \phi\left(\frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}}\right)\right. \\ \left. \times \Phi\left(\alpha(\theta_k) \cdot \left(\frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}}\right)\right)\right), & \text{VSVA.} \end{cases}$$

Similar as Chapter 4, for each component we augment $\mathbf{y}$ by introducing $\mathbf{w} = (w_{ik})_{1\leq i\leq n, 1\leq k\leq K}$, and we also introduce a set of latent component indicators $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ with $i = 1, 2, \ldots, n$. We define it with binary values

$$Z_{ik} \quad = \quad \begin{cases} 1 & \text{if } y_i \text{ belongs to component } k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we form the complete dataset $(\mathbf{y}, \mathbf{w}, \mathbf{Z}) = (y_i, \mathbf{w}_i, \mathbf{Z_i})_{1\leq i\leq n}$. The **complete-data log-likelihood** functions for the four-piece distribution family can be obtained now:

$$\begin{aligned} & l_{com}(\boldsymbol{\Theta}^{\star} \mid \mathbf{y}, \mathbf{w}, \mathbf{Z}) \\ = & \begin{cases} \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left\{\log \pi_k + \log \frac{1}{\sigma(\eta)\cdot\pi} - \frac{1}{2}b_{ik}^2 - \frac{1}{2}\left(w_{ik} - \alpha(\theta_k)\cdot b_{ik}\right)^2\right\}, & \text{ESVA} \\ \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left\{\log \pi_k + \log \frac{1}{\sigma(\eta_k)\cdot\pi} - \frac{1}{2}b_{ik}^2 - \frac{1}{2}\left(w_{ik} - \alpha(\theta)\cdot b_{ik}\right)^2\right\}, & \text{VSEA} \\ \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left\{\log \pi_k + \log \frac{1}{\sigma(\eta)\cdot\pi} - \frac{1}{2}b_{ik}^2 - \frac{1}{2}\left(w_{ik} - \alpha(\theta)\cdot b_{ik}\right)^2\right\}, & \text{ESEA} \\ \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left\{\log \pi_k + \log \frac{1}{\sigma(\eta_k)\cdot\pi} - \frac{1}{2}b_{ik}^2 - \frac{1}{2}\left(w_{ik} - \alpha(\theta_k)\cdot b_{ik}\right)^2\right\}, & \text{VSVA.} \end{cases} \end{aligned}$$

where $b_{ik}$ is used for notation conciseness, and it is specified as

$$b_{ik} = \begin{cases} \frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta_k)\cdot\sqrt{\frac{2}{\pi}}, & \text{ESVA} \\ \frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}, & \text{VSEA} \\ \frac{y_i-\mu_k}{\sigma(\eta)} + \delta(\theta)\cdot\sqrt{\frac{2}{\pi}}, & \text{ESEA} \\ \frac{y_i-\mu_k}{\sigma(\eta_k)} + \delta(\theta_k)\cdot\sqrt{\frac{2}{\pi}}, & \text{VSVA.} \end{cases}$$

### 5.1.2 E-Step and M-Step in EM Algorithm

To find the MLE of log-likelihood $l_{inc}(\boldsymbol{\Theta}^\star \mid \mathbf{y})$, we use the EM algorithm introduced in Chapter 4 again to help our calculations. Then the **E Step** starts by finding the expectation for **complete-data log-likelihood** $l_{com}(\boldsymbol{\Theta}^\star \mid \mathbf{y}, \mathbf{w}, \mathbf{Z})$.

Given $\hat{\boldsymbol{\Theta}}^{\star(v)}$ which are the parameter estimates obtained from the $v$-th iteration step, then

$$\Psi(\boldsymbol{\Theta}^\star \mid \hat{\boldsymbol{\Theta}}^{\star(v)}) = \begin{cases} \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\frac{1}{\sigma(\eta)\cdot\pi} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left\{ 1 + \left[\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)} - 2\alpha(\theta_k)\cdot b_{ik}\right] \frac{\phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)}\right)} \right. \\ \left. + \left[\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)} - \alpha(\theta_k)\cdot b_{ik}\right]^2 \right\}, & \text{ESVA} \\ \\ \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\frac{1}{\sigma(\eta_k)\cdot\pi} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left\{ 1 + \left[\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)} - 2\alpha(\theta)\cdot b_{ik}\right] \frac{\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)}\right)} \right. \\ \left. + \left[\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)} - \alpha(\theta)\cdot b_{ik}\right]^2 \right\}, & \text{VSEA} \\ \\ \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\frac{1}{\sigma(\eta)\cdot\pi} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left\{ 1 + \left[\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)} - 2\alpha(\theta)\cdot b_{ik}\right] \frac{\phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)}\right)} \right. \\ \left. + \left[\alpha(\hat{\theta}^{(v)})\cdot b_{ik}^{(v)} - \alpha(\theta)\cdot b_{ik}\right]^2 \right\}, & \text{ESEA} \\ \\ \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log\frac{1}{\sigma(\eta_k)\cdot\pi} - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left\{ 1 + \left[\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)} - 2\alpha(\theta_k)\cdot b_{ik}\right] \frac{\phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)}\right)} \right. \\ \left. + \left[\alpha(\hat{\theta}_k^{(v)})\cdot b_{ik}^{(v)} - \alpha(\theta_k)\cdot b_{ik}\right]^2 \right\}, & \text{VSVA} \end{cases}$$

where $b_{ik}^{(v)}$ is specified as

$$b_{ik}^{(v)} = \begin{cases} \frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\theta}_k^{(v)}) \cdot \sqrt{\frac{2}{\pi}}, & \text{ESVA} \\ \frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})} + \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}}, & \text{VSEA} \\ \frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})} + \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}}, & \text{ESEA} \\ \frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})} + \delta(\hat{\theta}_k^{(v)}) \cdot \sqrt{\frac{2}{\pi}}, & \text{VSVA.} \end{cases}$$

Then the **expectation** of the latent component indicator is updated by

$$\hat{Z}_{ik}^{(v)} = 1 \times P\left(Z_{ik}=1 \mid y_i, \hat{\boldsymbol{\Theta}}^{\star(v)}\right) + 0 \times P\left(Z_{ik}=0 \mid y_i, \hat{\boldsymbol{\Theta}}^{\star(v)}\right) = E\left[Z_{ik} \mid y_i, \hat{\boldsymbol{\Theta}}^{\star(v)}\right],$$

$$\hat{Z}_{ik}^{(v)} = \begin{cases} \frac{\hat{\pi}_k^{(v)} \phi\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}_k^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}_k^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}{\sum_{t=1}^{K}\hat{\pi}_t^{(v)}\phi\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}_t^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}_t^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}_t^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}, & \text{ESVA} \\ \frac{\hat{\pi}_k^{(v)}\cdot\frac{2}{\sigma(\hat{\eta}_k^{(v)})}\phi\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}{\sum_{t=1}^{K}\hat{\pi}_t^{(v)}\cdot\frac{2}{\sigma(\hat{\eta}_t^{(v)})}\phi\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}_t^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}_t^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}, & \text{VSEA} \\ \frac{\hat{\pi}_k^{(v)} \phi\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}{\sum_{t=1}^{K}\hat{\pi}_t^{(v)}\phi\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}^{(v)})}+\delta(\hat{\theta}^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}, & \text{ESEA} \\ \frac{\hat{\pi}_k^{(v)} \phi\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})}+\delta(\hat{\theta}_k^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}_k^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_k^{(v)}}{\sigma(\hat{\eta}_k^{(v)})}+\delta(\hat{\theta}_k^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}{\sum_{t=1}^{K}\hat{\pi}_t^{(v)}\phi\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}_t^{(v)})}+\delta(\hat{\theta}_t^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha(\hat{\theta}_t^{(v)})\cdot\left(\frac{y_i-\hat{\mu}_t^{(v)}}{\sigma(\hat{\eta}_t^{(v)})}+\delta(\hat{\theta}_t^{(v)})\cdot\sqrt{\frac{2}{\pi}}\right)\right)}, & \text{VSVA.} \end{cases}$$

In the **M step** we calculate the derivatives of $\Psi$ concerning $\Theta_k^\star$ block by block to update each parameter. Firstly, by using the Lagrange multiplier on $\Psi$, we update the block of $\pi_k$ by

$$\hat{\pi}_k^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}.$$

Secondly, we update $\mu_k$ by setting $\frac{\partial\Psi}{\partial\mu_k}=0$, with known $\hat{\eta}_k^{(v)}$ (or $\hat{\eta}^{(v)}$), $\hat{\theta}_k^{(v)}$ (or $\hat{\theta}^{(v)}$) from last

iteration, then we can update $\mu_k$ by

$$\hat{\mu}_k^{(v+1)} = \begin{cases} \frac{\sum_{i=1}^n \hat{Z}_{ik}^{(v)} y_i}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} + \sigma(\hat{\eta}^{(v)}) \cdot \delta(\hat{\theta}_k^{(v)}) \cdot \sqrt{\frac{2}{\pi}} - \frac{\sigma(\hat{\eta}^{(v)})}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} \cdot \frac{\alpha(\hat{\theta}_k^{(v)})}{1+\alpha(\hat{\theta}_k^{(v)})^2} \\ \times \sum_{i=1}^n \hat{Z}_{ik}^{(v)} \left[ \frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)} \right], & \text{ESVA} \\ \\ \frac{\sum_{i=1}^n \hat{Z}_{ik}^{(v)} y_i}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} + \sigma(\hat{\eta}_k^{(v)}) \cdot \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}} - \frac{\sigma(\hat{\eta}_k^{(v)})}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} \cdot \frac{\alpha(\hat{\theta}^{(v)})}{1+\alpha(\hat{\theta}^{(v)})^2} \\ \times \sum_{i=1}^n \hat{Z}_{ik}^{(v)} \left[ \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)} \right], & \text{VSEA} \\ \\ \frac{\sum_{i=1}^n \hat{Z}_{ik}^{(v)} y_i}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} + \sigma(\hat{\eta}^{(v)}) \cdot \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}} - \frac{\sigma(\hat{\eta}^{(v)})}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} \cdot \frac{\alpha(\hat{\theta}^{(v)})}{1+\alpha(\hat{\theta}^{(v)})^2} \\ \times \sum_{i=1}^n \hat{Z}_{ik}^{(v)} \left[ \frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)} \right], & \text{ESEA} \\ \\ \frac{\sum_{i=1}^n \hat{Z}_{ik}^{(v)} y_i}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} + \sigma(\hat{\eta}_k^{(v)}) \cdot \delta(\hat{\theta}_k^{(v)}) \cdot \sqrt{\frac{2}{\pi}} - \frac{\sigma(\hat{\eta}_k^{(v)})}{\sum_{i=1}^n \hat{Z}_{ik}^{(v)}} \cdot \frac{\alpha(\hat{\theta}_k^{(v)})}{1+\alpha(\hat{\theta}_k^{(v)})^2} \\ \times \sum_{i=1}^n \hat{Z}_{ik}^{(v)} \left[ \frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)} \right], & \text{VSVA.} \end{cases}$$

Thirdly, to update $\eta_k$ (or $\eta$) we set $\partial\Psi/\partial\eta_k = 0$ (or $\partial\Psi/\partial\eta = 0$), then we can have

**ESVA** with known $\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\boldsymbol{\theta}}^{(v)}\right)$ from previous step and last iteration, we update $\sigma = \exp(\eta)$ by solving the quadratic equation w.r.t. $e^{\eta}$, then

$$\begin{aligned} \hat{\sigma}^{(v+1)} &= \sigma(\hat{\eta}^{(v+1)}) \\ &= e^{\hat{\eta}^{(v+1)}} \\ &= \frac{1}{2n} T_2\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\boldsymbol{\theta}}^{(v)}\right) + \frac{1}{n} \\ &\quad \times \sqrt{\frac{1}{4} T_2\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\boldsymbol{\theta}}^{(v)}\right)^2 + n \sum_{i=1}^n \sum_{k=1}^K \left(1 + \alpha(\hat{\theta}_k^{(v)})^2\right) \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right)^2} \end{aligned}$$

in which

$$\begin{aligned} T_2\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\boldsymbol{\theta}}^{(v)}\right) &= \sqrt{\frac{2}{\pi}} \sum_{i=1}^n \sum_{k=1}^K \left(1 + \alpha(\hat{\theta}_k^{(v)})^2\right) \delta(\hat{\theta}_k^{(v)}) \cdot \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \\ &\quad - \sum_{i=1}^n \sum_{k=1}^K \alpha(\hat{\theta}_k^{(v)}) \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \cdot \left[ \frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)} \right]; \end{aligned}$$

**VSEA** with known $\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}^{(v)}\right)$ from previous step and last iteration, we update $\sigma_k = \exp(\eta_k)$ by solving a quadratic equation w.r.t. $e^{\eta_k}$, then

$$\begin{aligned}
\hat{\sigma}_k^{(v+1)} &= \sigma(\hat{\eta}_k^{(v+1)}) \\
&= e^{\hat{\eta}_k^{(v+1)}} \\
&= \frac{1}{2\sum_{i=1}^{n} \hat{Z}_{ik}^{(v)}} T_3\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}^{(v)}\right) + \frac{1}{\sum_{i=1}^{n} \hat{Z}_{ik}^{(v)}} \\
&\quad \times \sqrt{\frac{1}{4} T_3\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}^{(v)}\right)^2 + \left(1 + \alpha(\hat{\theta}^{(v)})^2\right) \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \cdot \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right)^2}
\end{aligned}$$

in which

$$\begin{aligned}
T_3\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}^{(v)}\right) &= \left(1 + \alpha(\hat{\theta}^{(v)})^2\right) \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}} \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \\
&\quad - \alpha(\hat{\theta}^{(v)}) \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \cdot \left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right];
\end{aligned}$$

**ESEA**

with known $\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\theta}^{(v)}\right)$ from previous step and last iteration, we update $\sigma = \exp(\eta)$ by solving a quadratic equation w.r.t. $e^{\eta}$, then

$$\begin{aligned}
\hat{\sigma}^{(v+1)} &= \sigma(\hat{\eta}^{(v+1)}) \\
&= e^{\hat{\eta}^{(v+1)}} \\
&= \frac{1}{2n} T_4\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\theta}^{(v)}\right) + \frac{1}{n} \\
&\quad \times \sqrt{\frac{1}{4} T_4\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\theta}^{(v)}\right)^2 + n\left(1 + \alpha(\hat{\theta}^{(v)})^2\right) \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right)^2}
\end{aligned}$$

in which

$$\begin{aligned}
T_4\left(\hat{\boldsymbol{\mu}}^{(v+1)}, \hat{\theta}^{(v)}\right) &= \left(1 + \alpha(\hat{\theta}^{(v)})^2\right) \delta(\hat{\theta}^{(v)}) \cdot \sqrt{\frac{2}{\pi}} \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \\
&\quad - \alpha(\hat{\theta}^{(v)}) \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left(y_i - \hat{\mu}_k^{(v+1)}\right) \cdot \left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right];
\end{aligned}$$

**VSVA**

with known $\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}_k^{(v)}\right)$ from previous step and last iteration, we update $\sigma_k = \exp(\eta_k)$ by

solving a quadratic equation w.r.t. $e^{\eta_k}$, then

$$\begin{aligned}
\hat{\sigma}_k^{(v+1)} &= \sigma(\hat{\eta}_k^{(v+1)}) \\
&= e^{\hat{\eta}_k^{(v+1)}} \\
&= \frac{1}{2\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}} T_5\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}_k^{(v)}\right) + \frac{1}{\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}} \\
&\quad \times \sqrt{\frac{1}{4}T_5\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}_k^{(v)}\right)^2 + \left(1+\alpha(\hat{\theta}_k^{(v)})^2\right)\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} \cdot \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \hat{\mu}_k^{(v+1)}\right)^2}
\end{aligned}$$

in which

$$\begin{aligned}
T_5\left(\hat{\mu}_k^{(v+1)}, \hat{\theta}_k^{(v)}\right) &= \left(1+\alpha(\hat{\theta}_k^{(v)})^2\right)\delta(\hat{\theta}_k^{(v)}) \cdot \sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \hat{\mu}_k^{(v+1)}\right) \\
&\quad -\alpha(\hat{\theta}_k^{(v)})\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \hat{\mu}_k^{(v+1)}\right) \cdot \left[\frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right].
\end{aligned}$$

Finally to update parameter $\theta_k$ (or $\theta$), we use **Newton-Raphson method**, so calculate the first- and second- partial derivative of $\Psi$ w.r.t. $\theta_k$ (or $\theta$). We calculate some pre-knowledge:

$$\frac{\partial b_{ik}}{\partial \theta_k} = \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2}\sqrt{\frac{2}{\pi}} \qquad \frac{\partial \alpha(\theta_k)}{\partial \theta_k} = \frac{1}{2}\left(e^{\theta_k} + e^{-\theta_k}\right).$$

Then we can have

**ESVA**

$$\begin{aligned}
\frac{\partial \Psi}{\partial \theta_k} &= -\sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} b_{ik} + \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left[\frac{1}{2}\left(e^{\theta_k} + e^{-\theta_k}\right) b_{ik}\right. \\
&\quad \left. +\alpha(\theta_k)\sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2}\right]\left[\frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)} - \alpha(\theta_k) \cdot b_{ik}\right], \\
\frac{\partial^2 \Psi}{\partial \theta_k^2} &= -\frac{8e^{2\theta_k}}{\pi\left(1+e^{2\theta_k}\right)^2}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} - 2\sqrt{\frac{2}{\pi}} \cdot \delta(\theta_k)^3 \cdot \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} b_{ik} - \frac{1}{4}\left(e^{\theta_k} + e^{-\theta_k}\right)^2 \\
&\quad \times \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} b_{ik}^2 + \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left[\alpha(\theta_k) b_{ik} - 2\sqrt{\frac{2}{\pi}} \cdot \frac{e^{\theta_k}\left(e^{4\theta_k} - 6e^{2\theta_k} + 1\right)}{\left(e^{2\theta_k}+1\right)^3}\right] \\
&\quad \times \left[\frac{\phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}_k^{(v)}) \cdot b_{ik}^{(v)} - \alpha(\theta_k) \cdot b_{ik}\right], \\
b_{ik} &= \frac{y_i - \mu_k}{\sigma(\eta)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}};
\end{aligned}$$

**VSEA**

$$\begin{aligned}
\frac{\partial \Psi}{\partial \theta} &= -\sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta} \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}}{\left(e^{2\theta}+1\right)^2} + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left[\frac{1}{2}\left(e^{\theta}+e^{-\theta}\right) b_{ik} + \alpha(\theta)\sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta}}{\left(e^{2\theta}+1\right)^2}\right] \\
&\quad \times \left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)} - \alpha(\theta) \cdot b_{ik}\right], \\
\frac{\partial^2 \Psi}{\partial \theta^2} &= -\frac{8e^{2\theta} n}{\pi\left(1+e^{2\theta}\right)^2} - 2\sqrt{\frac{2}{\pi}} \cdot \delta(\theta)^3 \cdot \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik} - \frac{1}{4}\left(e^{\theta}+e^{-\theta}\right)^2 \\
&\quad \times \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left[\alpha(\theta) b_{ik} - 2\sqrt{\frac{2}{\pi}} \cdot \frac{e^{\theta}\left(e^{4\theta}-6e^{2\theta}+1\right)}{\left(e^{2\theta}+1\right)^3}\right] \\
&\quad \times \left[\frac{\phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)}{\Phi\left(\alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)}\right)} + \alpha(\hat{\theta}^{(v)}) \cdot b_{ik}^{(v)} - \alpha(\theta) \cdot b_{ik}\right], \\
b_{ik} &= \frac{y_i - \mu_k}{\sigma(\eta_k)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}};
\end{aligned}$$

**ESEA** $\partial\Psi/\partial\theta$ and $\partial^2\Psi/\partial\theta^2$ are the same as **VSEA** in forms but $b_{ik}$ has a different formula

$$b_{ik} = \frac{y_i - \mu_k}{\sigma(\eta)} + \delta(\theta) \cdot \sqrt{\frac{2}{\pi}};$$

**VSVA** $\partial\Psi/\partial\theta$ and $\partial^2\Psi/\partial\theta^2$ are the same as **ESVA** in forms but $b_{ik}$ has a different formula

$$b_{ik} = \frac{y_i - \mu_k}{\sigma(\eta_k)} + \delta(\theta_k) \cdot \sqrt{\frac{2}{\pi}}.$$

Then $\partial\Psi/\partial\theta_k = 0$ (or $\partial\Psi/\partial\theta = 0$) can be solved by successively better approximations, which has been introduced in Section 4.2.2.

### 5.1.3 Model Selection

To compare these four models in the four-piece distribution family after estimation, the Bayesian Information Criterion (BIC) is calculated for each model (the best one yields the smallest BIC

here).

$$\text{BIC} = \begin{cases} \left[3K - \sum_{k=1}^{K} \mathbb{I}(\alpha_k = 0)\right] \log(n) - 2l_{inc}(\boldsymbol{\Theta}^\star \mid \mathbf{y}), & \text{ESVA} \\ [3K - \mathbb{I}(\alpha = 0)] \log(n) - 2l_{inc}(\boldsymbol{\Theta}^\star \mid \mathbf{y}), & \text{VSEA} \\ [2K + 1 - \mathbb{I}(\alpha = 0)] \log(n) - 2l_{inc}(\boldsymbol{\Theta}^\star \mid \mathbf{y}), & \text{ESEA} \\ \left[4K - 1 - \sum_{k=1}^{K} \mathbb{I}(\alpha_k = 0)\right] \log(n) - 2l_{inc}(\boldsymbol{\Theta}^\star \mid \mathbf{y}), & \text{VSVA} \end{cases}$$

where $\mathbb{I}(x)$ is a binary function to show whether $x$ is true or not. $\mathbb{I}(x)$ is 1 when $x$ is true, otherwise 0.

This four-piece distribution family has also been extended to mixture regressions in my thesis. For the time reason, the model (VSVA as an example) and detailed EM algorithm steps are attached in **Appendix D** for your interest.

## 5.2 Simulation Studies

To observe the performances of our four-piece distribution family, we use Bayesian information criterion (Schwarz, 1978) (see Chapter 2.5.1) to check whether the model has been improved from *mclust* (Fraley and Raftery, 2002), BIC is a good criterion for model selection among different models. In our skew normal mixture model, for a $K-$component mixture dataset, at most $K$ more parameters are added in the modelling to describe the skewness of each component compared with *mclust* where skewness parameters $\alpha$s are all assumed as 0. Our four-piece distribution family uses the *mclust* results as initialisation and carries out D'Agostino's K-squared test (D'Agostino, 1970) for each initial component (cluster) to check the existence of skewness before carrying on the EM algorithm in our four-piece distribution family. In Table 5.1 for each setting (from setting I to setting XVI), we sampled 100 times and each time modelled it with our four-piece distribution family. Meanwhile, the BIC value calculated from *mclust* was kept to compare with our four-piece distribution family's BICs. The Im_R (Improvement Rate) is the percentage where at least one of our four-piece distribution family's BICs was superior to *mclust*'s BIC. In Table 5.1, $K$ is the underlying component number, $n$ is the sample size, $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha})$ are the underlying true skew normal parameters, $\boldsymbol{\pi}$ is the mixing proportions which add up to 1 for each setting, and 'Model' indicates the underlying true model out of four different models in the four-piece distribution family regarding the values of $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$. As we can see from Table 5.1, there is at least a quarter (25% from setting VIII and setting IX) chance that the model selection for the mixture dataset is improved with the help of our four-piece distribution family. Even in setting XIII where real underlying $\alpha$s are all 0, there is

| No. | $n$ | $\boldsymbol{\pi}$ | $\boldsymbol{\mu}$ | $\boldsymbol{\sigma}$ | $\boldsymbol{\alpha}$ | Model | Im_R |
|---|---|---|---|---|---|---|---|
| I | 400 | $(25\%, 35\%, 40\%)$ | $(0, -3, 2)$ | $(0.5, 1, 1.5)$ | $(-2, 0, 2.5)$ | VSVA | 44% |
| II | 400 | $(25\%, 35\%, 40\%)$ | $(0, -3, 2)$ | $(0.5, 1, 1.5)$ | $(0, 3, 0)$ | VSVA | 89% |
| III | 400 | $(25\%, 35\%, 40\%)$ | $(0, -3, 2)$ | $(0.5, 1, 1.5)$ | $(2.5, 2.5, 2.5)$ | VSEA | 80% |
| IV | 400 | $(25\%, 35\%, 40\%)$ | $(0, -3, 2)$ | $(1, 1, 1)$ | $(1.5, 3, 2)$ | ESVA | 86% |
| V | 400 | $(25\%, 35\%, 40\%)$ | $(0, -3, 2)$ | $(1, 1, 1)$ | $(2.5, 2.5, 2.5)$ | ESEA | 81% |
| VI | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(-2, 2.5)$ | VSVA | 52% |
| VII | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(0, 2.5)$ | VSVA | 51% |
| VIII | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(1, 1)$ | $(0, 2.5)$ | ESVA | 25% |
| IX | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(1, 1)$ | $(-2, 2.5)$ | ESVA | 25% |
| X | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(2.5, 2.5)$ | VSEA | 70% |
| XI | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(1, 1)$ | $(2.5, 2.5)$ | ESEA | 70% |
| XII | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(0.5, 0.5)$ | VSEA | 46% |
| XIII | 300 | $(40\%, 60\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(0, 0)$ | VSEA | 38% |
| XIV | 300 | $(40\%, 60\%)$ | $(0, 3)$ | $(1, 1)$ | $(-2, 2.5)$ | ESVA | 66% |
| XV | 300 | $(40\%, 60\%)$ | $(0, 3)$ | $(1, 1)$ | $(0, 2.5)$ | ESVA | 64% |
| XVI | 300 | $(40\%, 60\%)$ | $(0, 3)$ | $(0.5, 1)$ | $(0.5, 0.5)$ | VSEA | 89% |
| XVII | 300 | $(40\%, 60\%)$ | $(0, 3)$ | $(1, 1)$ | $(2.5, 2.5)$ | ESEA | 92% |
| XVIII | 300 | $(20\%, 80\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(2.5, 2.5)$ | VSEA | 81% |
| XIX | 300 | $(20\%, 80\%)$ | $(0, 2)$ | $(0.5, 1)$ | $(0, 2.5)$ | VSVA | 72% |
| XX | 300 | $(20\%, 80\%)$ | $(0, 2)$ | $(1, 1)$ | $(0, 2.5)$ | ESVA | 53% |
| XXI | 300 | $(20\%, 80\%)$ | $(0, 2)$ | $(1, 1)$ | $(2.5, 2.5)$ | ESEA | 84% |

Table 5.1: The Improvement Rates (the percentage where at least one of our four-piece distribution family's BICs was superior to *mclust*'s BIC) under 21 different settings where 4 possible mixture model types and different skewness degrees were simulated. Repeated each scenario 100 times by resampling and replicating the estimation process to produce each Im_R value.

still a good chance of 38% to improve the modelling with our method considering *mclust* has assumed $\boldsymbol{\alpha} = (0, 0)$ beforehand. It comes from the benefit that one fewer $\sigma$ is used to depict the dataset in our ESEA model, then BIC value is improved by taking one fewer parameter into the penalty. Therefore, judging from the results in Table 5.1, our four-piece distribution family is a good complement to the existing R package *mclust* to adapt to any possible skewed mixture data.

To take a closer look at the estimation accuracy of our four-piece distribution family under different scenarios, we calculated the Mean Absolute Error (MAE) for each estimate value and Rand Index (Rand, 1971) which is a measure to quantify the agreement between two classifications based on the class label of each object. As shown in Table 5.2, we simulated some scenarios from Table 5.1 (marked with Roman numerals) and compared them with the results from *mclust* (the ones with dash notation). As we can see, the estimation accuracy could be improved vastly for the scenarios with both skews being evident. For scenarios with an evident skew in one component, some statistics could be improved. In scenarios where skews were not obvious in both components, our method could not exceed *mclust* because of the nature of *mclust* which assumes the inexistence of skewnesses for all components.

| No. | Results from four-piece distribution | | | | Results from package *mclust* | | | |
|---|---|---|---|---|---|---|---|---|
| | RI | $MAE_{\boldsymbol{\mu}}$ | $MAE_{\boldsymbol{\sigma}}$ | $MAE_{\boldsymbol{\alpha}}$ | RI' | $MAE'_{\boldsymbol{\mu}}$ | $MAE'_{\boldsymbol{\sigma}}$ | $MAE'_{\boldsymbol{\alpha}}$ |
| VI | 92.6% | 0.06 | 0.14 | 1.42 | 95.5% | 0.06 | 0.24 | 2.31 |
| VIII | 51.0% | 0.31 | 0.13 | 2.13 | 49.9% | 0.28 | 0.23 | 1.71 |
| X | 90.0% | 0.06 | 0.12 | 1.22 | 86.8% | 0.09 | 0.23 | 2.50 |
| XI | 73.3% | 0.11 | 0.13 | 1.03 | 67.4% | 0.17 | 0.31 | 2.50 |
| XIV | 97.9% | 0.05 | 0.10 | 1.29 | 98.2% | 0.05 | 0.30 | 2.30 |
| XVI | 93.5% | 0.06 | 0.08 | 0.61 | 93.5% | 0.05 | 0.07 | 0.50 |
| XVIII | 87.3% | 0.09 | 0.12 | 1.26 | 85.9% | 0.10 | 0.28 | 2.50 |
| XX | 66.8% | 0.17 | 0.15 | 1.57 | 63.7% | 0.17 | 0.29 | 2.12 |

Table 5.2: Comparison of the estimation performances between our four-piece distribution family and *mclust*. RI is Rand Index. $MAE_{\boldsymbol{\mu}}$, $MAE_{\boldsymbol{\sigma}}$ and $MAE_{\boldsymbol{\alpha}}$ are the Mean Absolute Errors for the estimated parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\alpha}$. The left-side four statistics (RI, $MAE_{\boldsymbol{\mu}}$, $MAE_{\boldsymbol{\sigma}}$, $MAE_{\boldsymbol{\alpha}}$) are the results from our four-piece distribution family. The right-side four dashed statistics (RI', $MAE'_{\boldsymbol{\mu}}$, $MAE'_{\boldsymbol{\sigma}}$, $MAE'_{\boldsymbol{\alpha}}$) are the results from R package *mclust*. Repeated each scenario 50 times by resampling and replicating the estimation process, then the average values were recorded in this table.

# Chapter 6

# Conclusions and Future Works

## 6.1 Conclusions

In this thesis, first of all, a brand new penalised likelihood fusion model is established to handle the multivariate high-dimensional mixture regression problem. In Chapter 3, we examine the group structure of our drug-and-gene dataset by adopting LASSO (Tibshirani,1996) onto each drug individually and using Jaccard Index to find similarities between them. Then inspired by the likelihood fusion model (Oftadeh, 2017), we introduce our penalised likelihood fusion model to adapt to the high-dimensional property of the gene expression data. With the Expectation-Maximisation algorithm (Dempster et al., 1977), the parameter estimation of this model is realised, at the same time, the penalty term and the number of components are selected through the Bayesian information criterion method (Schwarz, 1978). In the following drug data application, after dimension-reduction steps (individual LASSO and marginal analysis), the whole penalised likelihood fusion model is applied to the drug-and-gene dataset to reach the result that 227 anti-cancer drugs are divided into 9 groups with significantly correlated genes discovered for each group. The likelihood comparisons from cross-validation can manifest the superiority of our model over *mclust* (Fraley and Raftery, 2002) in clustering anti-cancer drugs when gene expression data are provided.

Next, we take missingness into consideration for the aforementioned penalised likelihood fusion model by introducing subscripts to exclude missing data in each anti-cancer drug. Meanwhile, the LASSO-penalty ($l_1-$norm) is generalised to the $l_q-$norm (Hastie et al., 2015) to obtain a more flexible result. With a warm start, the relative error also decreases in the simulation trials compared with $l_1-$norm case and without-a-warm-start case. Then our $l_q-$norm penalised likelihood fusion model considering missingness is constructed with $q = 0.5$ as the preferable choice for our drug and gene data application. In the numerical results, this time 187

anti-cancer drugs are divided into 5 groups and the related genes are also marked with weights for each group. Compared with the previous numerical result, sample size $n$ increases from 111 to 397, which benefits from our subscript (to avoid missingness) upgrade. Moreover, fewer drugs and more genes are put into our modelling with $l_1-$norm upgrading to $l_{0.5}-$norm (and a warm start) to improve the accuracy (conclude from simulations). In summary, Figure 3.2.5 illustrates our final result by relating the anti-cancer drug IC50 sensitivity data to gene expression data. After that, we propose another method (different from the former BIC method) based on a matrix imputation method primePCA (Zhu, Wang and Samworth, 2019) to determine the number of components $K$. This new method can be used in not only complete-case (data without missingness) but also incomplete-case (data with missingness) problems.

From Chapter 4 we commit to the inherent skewness of our drug data. With our new location-scale model and the introduction of a quadratic $\alpha-$penalty, we use the EM algorithm to estimate the parameters in this model successfully. The next asymptotic analysis reveals the behaviour of our estimates under certain conditions for two scenarios 'underlying true $\alpha$ is 0 or not'. With our efforts, the asymptotic distributions including detailed variance formulae for the MLE estimates of the skew normal distribution are discovered. The following bootstrapping on one certain drug shows, to some degree, the validity of our variance calculations by deriving the confidence interval for the drug data skewness. The simulation studies also show the superiority of our method over the constant penalty from Azzalini and Arellano-Valle (2012). In the next anti-cancer drug data applications, our modelling has detected the inherent skewness within most drug data, where the data from 131 out of 187 anti-cancer drugs manifest Fisher's moment coefficients of skewnesses larger than 0.3. In Chapter 5, inspired by *mclust* (Fraley and Raftery, 2002), we put forward the four-piece distribution family to look at skew normal mixture problems. With the aforementioned EM algorithm, the MLE estimation is accomplished after the model selection process motivated by optimising the BIC value. The following simulation studies verify that our four-piece distribution family is a good complement to the existing R package *mclust* when confronting mixture data with potential skewness features.

## 6.2 Future Works

- When building our $l_q-$norm penalised likelihood fusion model considering missingness, we only compared the performance between $q = 1$ and $q = 0.5$, the choice of $q$ could actually be any number within $(0, 1]$. A proper grid could be set such as $0.1, 0.2, \ldots 1$ to select a more convincing $q$ based on cross-validation or other criteria.

- When dealing with the missing data in our applications, we excluded them by introducing

subscripts to avoid the discussion about the actual underlying values there. We have left the space for finding a good matrix imputation method after which we could also apply our $l_q-$norm penalised likelihood fusion model to the completed dataset. Actually we have tried Soft Impute (Mazumder, Hastie and Tibshirani, 2010) and primePCA (Zhu, Wang and Samworth, 2019) to complete the matrix imputation problem separately in the first step. But the following applications in our $l_q-$norm penalised likelihood fusion modelling showed inferior performances compared with our original method using subscripts. Thus, whether there exists a good imputation method for incomplete data before applying them to our methodology is still in the air.

- For our location-scale skew normal model with a quadratic $\alpha-$penalty in Chapter 4, we have not thrown it into the high-dimensional situation, since the introduction of one more penalty term could result in a much more complex procedure of selecting these two tuning parameters. There is also a big space left to build a multivariate version of the location-scale skew normal model with the quadratic $\alpha-$penalty which could even be used to analyse our complete drug-and-gene dataset when extending to regression and high-dimensional version. If we continue linking with Chapter 5 and also introduce subscripts to avoid the missingness in the given dataset, after some great efforts, finally $l_q-$norm penalised likelihood fusion model considering missingness and skewness could be accomplished, where our entire anti-cancer drug and gene expression dataset could be analysed with more considerations. But it could bring about the choices of multiple penalties and sparsities, which are two more big topics to research. I hope our research could arouse interest among researchers to move forward in this drug-and-gene statistical field.

- In this thesis, we discussed about the classification of drugs with our modelling. But it is still of interest for medical researchers to classify patients' cancer cells at the same time. Then it will become the modelling and clustering on our data $\mathbf{X}$. There are some papers relating to mixture modelling with concomitant variables, see Huang and Yao (2012), Huang et al. (2013), and Huang et al. (2018).

# Appendix A

## Cyclic Coordinate Descent Calculation

Let $q = 1$ and (3.2.6) can also be expressed as

$$\hat{\boldsymbol{\beta}_k}^{(v+1)} = \arg\min_{\boldsymbol{\beta}_k} f = \arg\min_{\boldsymbol{\beta}_k} \sum_{j=1}^{J} \left\{ \frac{\tau_{jk}^{(v)}}{2\sigma_k} \left\| \mathbf{y}_{(j)} - \bar{y}_{(j)} 1 - \mathbf{X}_{(j)} \boldsymbol{\beta}_k \right\|_2^2 + \tau_{jk}^{(v)} \cdot n_{(j)} \cdot \lambda_1 \left\| \boldsymbol{\beta}_k \right\|_1 \right\},$$

which can be seen as a summation of $J$ LASSOs sharing a common coefficient vector $\boldsymbol{\beta}_k$. Our cyclic coordinate descent starts by deriving the partial derivative w.r.t. each coordinate direction:

$$\begin{aligned} \beta_{ki}^{(n+1)} &= \arg\min_{y \in \mathbb{R}} f\left(\beta_{k1}^{(n+1)}, \ldots, \beta_{k(i-1)}^{(n+1)}, y, \beta_{k(i+1)}^{(n)}, \ldots, \beta_{kp}^{(n)}\right) \\ \frac{\partial f}{\partial \beta_{ki}} &= \beta_{ki} \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \mathbf{X}_{(j)i} + \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \mathbf{X}_{(j)(-i)} \boldsymbol{\beta}_{k(-i)} \\ &\quad - \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\right) + \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)} \cdot \lambda_1 \mathrm{sgn}(\beta_{ki}) \\ &= 0 \end{aligned} \tag{6.2.1}$$

for $i \in \{1, 2, \ldots, p\}$. When each variable $\beta_{ki}$ of $\boldsymbol{\beta}_k$ is updated once, round $n+1$ is finished. Here $\mathbf{X}_{(j)(-i)}$ means the sub-matrix $\mathbf{X}_{(j)}$ without $i$-th column. As we have standardised $\mathbf{X}_{(j)}$ by each column, so we have $\frac{1}{n_{(j)}} \mathbf{X}_{(j)i}^T \mathbf{X}_{(j)i} = 1$. Then, for convenience, we can define:

$$\begin{aligned} \mathcal{A} &= \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \mathbf{X}_{(j)i} = \sum_{j=1}^{J} \frac{n_{(j)} \tau_{jk}^{(v)}}{\sigma_k} \\ \mathcal{B} &= \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \mathbf{X}_{(j)(-i)} \boldsymbol{\beta}_{k(-i)} - \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \mathbf{X}_{(j)i}^T \left(\mathbf{y}_{(j)} - \bar{y}_{(j)}\right) \\ \mathcal{C} &= \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n_{(j)} \cdot \lambda_1. \end{aligned}$$

Equation (6.2.1) can be written concisely as $\boldsymbol{\beta}_{ki} \cdot \mathcal{A} + \mathcal{B} + \mathcal{C} \cdot \operatorname{sgn}(\boldsymbol{\beta}_{ki}) = 0$. This is a soft thresholding. After rearranging, we have

$$\hat{\boldsymbol{\beta}}_{ki}^{(n+1)} = \begin{cases} \frac{-\mathcal{B}-\mathcal{C}}{\mathcal{A}}, & \mathcal{B} < -\mathcal{C} \\ 0, & |\mathcal{B}| < \mathcal{C} \\ \frac{-\mathcal{B}+\mathcal{C}}{\mathcal{A}}, & \mathcal{B} > \mathcal{C}. \end{cases}$$

Continue rounds of updates cyclically until

$$\frac{\left| f\left(\boldsymbol{\beta}_k^{(n+1)}\right) - f\left(\boldsymbol{\beta}_k^{(n)}\right) \right|}{f\left(\boldsymbol{\beta}_k^{(n)}\right)} \leq \varepsilon$$

here $\boldsymbol{\beta}_k^{(t)}$ is the $\boldsymbol{\beta}_k$ after $t$-th round of updating, thus $\boldsymbol{\beta}_k^{(t)} = \left(\beta_{k1}^{(t)}, \beta_{k2}^{(t)}, \ldots, \beta_{kp}^{(t)}\right)$.

# EM Algorithm for Penalised Likelihood Fusion Complete Case (Without Missingness)

## E-step

Using the Bayes theorem considering the current information (the $v$-th parameter estimates) and the given data, the hidden variable indicator $z_{jk}$ can have a proportional expectation with weights $\pi_k^{(v)}$:

$$\begin{aligned}
& \tau_{jk}^{(v)} \\
= \; & \Pr(z_{jk} = 1 \mid \mathbf{y}_j, \boldsymbol{\Phi}^{(v)}) \\
= \; & \mathrm{E}(z_{jk} \mid \mathbf{y}_j, \boldsymbol{\Phi}^{(v)}) \\
= \; & \frac{\pi_k^{(v)} f_k(\mathbf{y}_j \mid \alpha_j, \mathbf{X}\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2})}{\sum_{k=1}^{K} \pi_k^{(v)} f_k(\mathbf{y}_j \mid \alpha_j, \mathbf{X}\boldsymbol{\beta}_k^{(v)}, \sigma_k^{(v)2})} \\
= \; & \frac{\frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n/2}} \exp\left\{ -\frac{\left(\mathbf{y}_j - \alpha_j \mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k^{(v)}\right)^T \left(\mathbf{y}_j - \alpha_j \mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k^{(v)}\right)}{2\sigma_k^{(v)2}} \right\} \times \exp\left( -\frac{n\lambda \left\| \boldsymbol{\beta}_k^{(v)} \right\|_1}{\sigma_k^{(v)}} \right) \frac{1}{\sigma_k^{(v)2}}}{\sum_{k=1}^{K} \frac{\pi_k^{(v)}}{\left(2\pi\sigma_k^{(v)2}\right)^{n/2}} \exp\left\{ -\frac{\left(\mathbf{y}_j - \alpha_j \mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k^{(v)}\right)^T \left(\mathbf{y}_j - \alpha_j \mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k^{(v)}\right)}{2\sigma_k^{(v)2}} \right\} \times \exp\left( -\frac{n\lambda \left\| \boldsymbol{\beta}_k^{(v)} \right\|_1}{\sigma_k^{(v)}} \right) \frac{1}{\sigma_k^{(v)2}}}.
\end{aligned}$$

After replacing the $z_{jk}$ in (3.1.2) with $\tau_{jk}^{(v)}$, we reach the expectation function

$$\begin{aligned} Q &= Q(\mathbf{\Phi} \mid \mathbf{\Phi}^{(v)}) \\ &= E_{\mathbf{Z}|\mathbf{Y},\mathbf{X},\mathbf{\Phi}^{(v)}}\left[\log\left(pL_{com}(\mathbf{\Phi})\right)\right] \\ &= \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left(\pi_k\right) - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left((2\pi)^{n/2}\sigma_k^{n+2}\right) \\ &\quad - \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}\cdot\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)}{2\sigma_k^2} - \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\frac{n\lambda\left\|\boldsymbol{\beta}_k\right\|_1}{\sigma_k}. \end{aligned}$$

## M-step

To update the estimate of $\mathbf{\Phi}$, we maximise $Q(\mathbf{\Phi} \mid \mathbf{\Phi}^{(v)})$ with respect to $\mathbf{\Phi}$ block by block. Starting from the intercept $\alpha_j$, we let

$$\begin{aligned} \frac{\partial Q}{\partial \alpha_j} &= \frac{\partial \sum_{j=1}^{J}\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}\cdot\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_j-\alpha_j\mathbf{1}-\mathbf{X}\boldsymbol{\beta}_k\right)}{2\sigma_k^2}}{\partial\alpha_j} \\ &= -\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\tau_{jk}^{(v)}}{\sigma_k^2}\left(y_{ij}-\alpha_j-\mathbf{x}_i^T\boldsymbol{\beta}_k\right) \\ &= -\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}}{\sigma_k^2}\left(\sum_{i=1}^{n}y_{ij}-n\alpha_j-\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\beta}_k\right) \\ &= 0. \end{aligned}$$

If we standardise $\mathbf{X}$ by column-wise way before the EM algorithm, we can have $\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\beta}_k = 0$, hence $\sum_{k=1}^{K}\frac{\tau_{jk}^{(v)}}{\sigma_k^2}\cdot\left(\sum_{i=1}^{n}y_{ij}-n\alpha_j\right)=0$. Then

$$\alpha_j = \frac{\sum_{i=1}^{n}y_{ij}}{n} = \bar{y_j},$$

where $\bar{y_j}$ is the mean of every $j$th column. Therefore, the calculation $\mathbf{y}_j-\alpha_j\mathbf{1}$ is a centralisation operation to each column of $\mathbf{Y}$, and $\alpha_j$ just depends on the original data matrix $\mathbf{Y}$, which is independent with other parameters such as $\tau_{jk},\sigma_k,\boldsymbol{\beta}_k,\pi_k,\lambda$. Accordingly, We don't need to update $\alpha_j$ in the iteration of the EM algorithm. We replace it with $\bar{y_j}$ in the next steps.

To update $\pi_k$ with the constraint $\sum_{k=1}^{K}\pi_k = 1$ and $\pi_k \geq 0$, we turn to Lagrange multiplier for help, thus the corresponding Lagrangian function is

$$L(\pi_k,\Lambda) = \sum_{j=1}^{J}\sum_{k=1}^{K}\tau_{jk}^{(v)}\cdot\log\left(\pi_k\right) - \Lambda\left(\sum_{k=1}^{K}\pi_k-1\right),$$

and to maximise this function, we differentiate it with respect to $\pi_k$ and equal 0

$$\frac{\partial L(\pi_k, \Lambda)}{\partial \pi_k} = \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\pi_k} - \Lambda = 0,$$

after rearrangement, it gives

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)}}{\Lambda}. \tag{6.2.2}$$

Substitute the $\pi_k$ with our estimate $\hat{\pi}_k^{(v+1)}$, we get

$$\sum_{k=1}^{K} \pi_k = \sum_{k=1}^{K} \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)}}{\Lambda} = 1.$$

Hence $\Lambda = \sum_{k=1}^{K}\sum_{j=1}^{J} \tau_{jk}^{(v)}$ . From the definition of $\tau_{jk}$, we have $\sum_{k=1}^{K} \tau_{jk}^{(v)} = 1$. Thus $\Lambda = \sum_{j=1}^{J} 1 = J$. Finally we update the (6.2.2) in the $v$-th EM iteration to reach the $(v+1)$-th estimate of $\pi_k$:

$$\hat{\pi}_k^{(v+1)} = \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)}}{J}.$$

To update the block of $\boldsymbol{\beta}_k$, we can apply

$$\begin{aligned}
\frac{\partial Q}{\partial \beta_k} &= \frac{\partial \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\tau_{jk}^{(v)} \cdot (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{2\sigma_k^2}}{\partial \boldsymbol{\beta}_k} + \frac{\partial \sum_{j=1}^{J}\sum_{k=1}^{K} \tau_{jk}^{(v)} \cdot \frac{n\lambda \|\boldsymbol{\beta}_k\|_1}{\sigma_k}}{\partial \boldsymbol{\beta}_k} \\
&= \sum_{j=1}^{J} \frac{-2\tau_{jk}^{(v)} \cdot \mathbf{X}^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{2\sigma_k^2} + \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)} \cdot n\lambda \cdot \operatorname{sgn}(\boldsymbol{\beta_k})}{\sigma_k} \\
&= -\frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \mathbf{X}^T (\mathbf{y}_j - \bar{y_j}\mathbf{1})}{\sigma_k^2} + \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_k}{\sigma_k^2} + \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n\lambda \cdot \operatorname{sgn}(\boldsymbol{\beta_k})}{\sigma_k} \\
&= 0
\end{aligned}$$

Thus

$$\sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \cdot \mathbf{X}^T (\mathbf{y}_j - \bar{y_j}\mathbf{1}) - \sum_{j=1}^{J} \frac{\tau_{jk}^{(v)}}{\sigma_k} \cdot \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_k - \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n\lambda \cdot \operatorname{sgn}(\boldsymbol{\beta_k}) = 0$$

We can write it as

$$\mathbf{X}^{*T}\mathbf{Y}^* - \mathbf{X}^{*T}\mathbf{X}^*\boldsymbol{\beta}_k - \lambda \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n \cdot \operatorname{sgn}(\boldsymbol{\beta_k}) = 0, \tag{6.2.3}$$

where $\operatorname{sgn}(\boldsymbol{\beta_k}) = (\operatorname{sgn}(\beta_{k1}), \ldots, \operatorname{sgn}(\beta_{kp}))^T$ is a vector of sign functions with length $p$, and $\mathbf{X}^*$,

$\mathbf{Y}^*$ for given $k$ are defined as

$$\mathbf{X}^* = \text{diag}\left(\frac{\tau_{1k}^{(v)}}{\sigma_k}, \frac{\tau_{2k}^{(v)}}{\sigma_k}, \ldots, \frac{\tau_{Jk}^{(v)}}{\sigma_k}\right)^{1/2} \begin{pmatrix} \mathbf{X} \\ \mathbf{X} \\ \vdots \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}} \cdot \mathbf{X} \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}} \cdot \mathbf{X} \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}} \cdot \mathbf{X} \end{pmatrix}, \tag{6.2.4}$$

$$\mathbf{Y}^* = \text{diag}\left(\frac{\tau_{1k}^{(v)}}{\sigma_k}, \frac{\tau_{2k}^{(v)}}{\sigma_k}, \ldots, \frac{\tau_{Jk}^{(v)}}{\sigma_k}\right)^{1/2} \mathbf{Y} = \begin{pmatrix} \sqrt{\frac{\tau_{1k}^{(v)}}{\sigma_k}} \cdot (\mathbf{y}_1 - \bar{y_1}\mathbf{1}) \\ \sqrt{\frac{\tau_{2k}^{(v)}}{\sigma_k}} \cdot (\mathbf{y}_2 - \bar{y_2}\mathbf{1}) \\ \vdots \\ \sqrt{\frac{\tau_{Jk}^{(v)}}{\sigma_k}} \cdot (\mathbf{y}_J - \bar{y_J}\mathbf{1}) \end{pmatrix}, \tag{6.2.5}$$

where for every entry of $\mathbf{X}^*$, $\mathbf{X}$ is a $n \times p$ matrix, so $\mathbf{X}^*$ is a $nJ \times p$ matrix. And for every element of $\mathbf{Y}^*$, $\mathbf{y}_j$ is a $n \times 1$ vector, so $\mathbf{Y}^*$ is a $nJ \times 1$ vector.

To solve the equation (6.2.3), it suffices to solve the following LASSO problem

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k^{(v+1)} &= \arg\min_{\beta_k} \left\{ \frac{1}{2} \sum_{j=1}^{J} \left(\mathbf{y}_j^* - \mathbf{x}_j^* \boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_j^* - \mathbf{x}_j^* \boldsymbol{\beta}_k\right) + \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n \cdot \lambda \left\|\boldsymbol{\beta}_k\right\|_1 \right\} \\ &= \arg\min_{\beta_k} \left\{ \frac{1}{2} \sum_{j=1}^{J} \sum_{i=1}^{n} \left(y_{ij}^* - \mathbf{x}_{ij}^* \boldsymbol{\beta}_k\right)^2 + \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n \cdot \lambda \left\|\boldsymbol{\beta}_k\right\|_1 \right\} \\ &= \arg\min_{\beta_k} \left\{ \frac{1}{2} \left\|\mathbf{Y}^* - \mathbf{X}^* \boldsymbol{\beta}_k\right\|_2^2 + \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot n \cdot \lambda \left\|\boldsymbol{\beta}_k\right\|_1 \right\}. \end{aligned}$$

As explained above, and $\boldsymbol{\beta}_k$ is a $p \times 1$ vector, therefore, it is a LASSO problem with $nJ$ observations and $p$ covariates.

To update $\sigma_k$ block, we set

$$\frac{\partial Q}{\partial \sigma_k^2} = 0,$$

which can yield

$$\begin{aligned} 0 = & \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \frac{n+2}{2} - \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \frac{1}{2} \sigma_k^{-2} \cdot \left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k\right)^T \left(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k\right) \\ & - \sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot \frac{1}{2} \sigma_k^{-1} \cdot \left(n\lambda \left\|\boldsymbol{\beta}_k\right\|_1\right). \end{aligned}$$

Rearrange the equation, we get

$$0 = (n+2)\,\sigma_k^2 - (n\lambda\,\|\boldsymbol{\beta}_k\|_1 + n\kappa_0)\,\sigma_k - \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{\sum_{j=1}^{J} \tau_{jk}^{(v)}}.$$

Thus we can update the block of $\sigma_k$ by

$$\hat{\sigma}_k^{(v+1)} = \frac{n\lambda\,\|\boldsymbol{\beta}_k\|_1}{2\,(n+2)} + \sqrt{\left(\frac{n\lambda\,\|\boldsymbol{\beta}_k\|_1}{2\,(n+2)}\right)^2 + \frac{\sum_{j=1}^{J} \tau_{jk}^{(v)} \cdot (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{(n+2)\sum_{j=1}^{J} \tau_{jk}^{(v)}}}.$$

The EM algorithm alternates between the above **E-step** and **M-step** until convergence. The stopping rule that we impose to confirm the convergence is based on the relative change in the log-likelihood. We say that the convergence is attained when the following tolerance holds

$$\left|\frac{pl_{inc}(\boldsymbol{\Phi}^{(v+1)}) - pl_{inc}(\boldsymbol{\Phi}^{(v)})}{pl_{inc}(\boldsymbol{\Phi}^{(v)})}\right| \leq \epsilon,$$

where $\epsilon$ is a reasonably small value (default value of $\epsilon$ is $10^{-5}$) and $pl_{inc}(\boldsymbol{\Phi})$ is calculated by

$$pl_{inc}(\boldsymbol{\Phi}) = \sum_{j=1}^{J} \log\left\{\sum_{k=1}^{K} \frac{\pi_k}{(2\pi\sigma_k^2)^{n/2}} \exp\left[-\frac{(\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)^T (\mathbf{y}_j - \bar{y_j}\mathbf{1} - \mathbf{X}\boldsymbol{\beta}_k)}{2\sigma_k^2}\right] \times \exp\left(-\frac{n\lambda\,\|\boldsymbol{\beta}_k\|_1}{\sigma_k}\right)\frac{1}{\sigma_k^2}\right\}.$$

## Real Data Results from $q = 1$ Case

The part shows the table mentioned in Section 3.2.5. Although $q = 1$ case has a strong power of clustering, it still has a few differences on the detailed clustering results compared to the final results in our main text. As shown in Table 6.1, Table 6.2 and Table 6.3, in total, 14 anti-cancer drugs are clustered into different groups (i.e. PAC.1, TAK.715, BX.912, AS605240, Enzastaurin, GSK429286A, Quizartinib, CP724714, KIN001.266, OSI.930, SB52334, Methotrexate, Navitoclax, MK.2206), which can result in the changes of Group Size, $\hat{\boldsymbol{\pi}}_{1:K}$ and $\hat{\boldsymbol{\sigma}}$. And more genes are selected out for each group, which manifests a less sparse coefficient result for $\hat{\boldsymbol{\beta}}$.

Table 6.1: From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 1)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| 1 | LFM.A13, SGC0946, Ruxolitinib, FMK, STF.62247, Tivozanib, Zibotentan, VNLG.124, T0901317, Selisistat, Veliparib, Lenalidomide, Vismodegib, SL0101, IOX2, UNC1215, Temozolomide, PFI.3, GW.2580, XMD14.99, KIN001.236, FR.180204, KIN001.270, XMD15.27, Tamoxifen, VX.702, GSK1904529A. | 27 | 166 | 0.145 | 0.661 |
| 2 | NSC.87877, Midostaurin, JNK.9L, PF.562271, FTI.277, OSU.03012, Shikonin, Embelin, FH535, Dacinostat, Bexarotene, TW.37, Luminespib, Linifanib, Bryostatin.1, Rucaparib, AS601245, Alectinib, WHI.P97, CP724714, Cetuximab, PF.4708671, AICA.Ribonucleotide, Vinblastine, Cisplatin, Docetaxel, Tretinoin, Gefitinib, Vorinostat, Axitinib, GW441756, Lestaurtinib, Motesanib, KU.55933, BX795, NU7441, Doramapimod, JNK.Inhibitor.VIII, PD173074, ZM447439, RO.3306, Serdemetan, Dactolisib, AZD8055, CCT007093, EHT.1864, CCT.018159, SB505124, | 52 | 144 | 0.278 | 0.973 |

Table 6.2: From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 2)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| | Piperlongumine, PFI.1, YK.4.279, XAV939. | | | | |
| 3 | CP466722, BMS.345541, TL.2.105, Idelalisib, Cabozantinib, JW.7.24.1, NPK76.II.72.1, NG.25, TL.1.85, Tubastatin.A, I.BET.762, BIX02189, KIN001.244, Masitinib, KIN001.260, PIK.93, MPS.1.IN.1, TPCA.1, NVP.BHG712, PAC.1, Fedratinib, Foretinib, Y.39983, YM201636, TAK.715, GSK429286A, Quizartinib, OSI.930, QL.XI.92, XMD13.2, KIN001.266, SB52334. | 32 | 212 | 0.171 | 1.057 |
| 4 | Doxorubicin, Etoposide, Gemcitabine, Vinorelbine, Mitomycin.C, QS11, Ponatinib, HG6.64.1, JQ12, GSK650394, DMOG, BAY.61.3606, IPA.3, Thapsigargin, Obatoclax.Mesylate, BMS.754807, Linsitinib, Bleomycin, Phenformin, Pazopanib, Epothilone.B, Tipifarnib, SB590885, Enzastaurin, VX.11e, NSC.207895, MK.2206, Amuvatinib, Cytarabine, Nilotinib, CI.1040, Temsirolimus, Bosutinib, AZD7762, Tanespimycin, Elesclomol, Nutlin.3a, Palbociclib, PD0325901, X.5Z..7.Oxozeaenol, Talazoparib, rTRAIL, | 46 | 132 | 0.246 | 1.553 |

Table 6.3: From $q = 1$ case, we find $(\lambda_1 = 0.025, K = 5)$. The clustering outcome and group characters from this point are listed. (Part 3)

| Group Label | Drug Names | Group Size | Number of Related Genes | $\hat{\boldsymbol{\pi}}_{1:K}$ | $\hat{\boldsymbol{\sigma}}$ |
|---|---|---|---|---|---|
| | ICL1100013, SN.38, Trametinib, Dabrafenib, Bleomycin..50.uM. | | | | |
| 5 | X5.Fluorouracil, AR.42, Ispinesib.Mesylate, AT.7519, GSK1070916, Daporinad, QL.XII.47, ZSTK474, WZ3105, Genentech.Cpd.10, Sepantronium.bromide, CUDC.101, Belinostat, CAY10603, Pelitinib, Omipalisib, OSI.027, CX.5461, PHA.793887, PI.103, GSK690693, SNX.2112, QL.X.138, THZ.2.49, THZ.2.102.1, BX.912, AS605240, Enzastaurin, Navitoclax, Methotrexate. | 30 | 216 | 0.160 | 1.574 |

# Appendix B

## Proof of Theorem 1

*Proof.* From mean value theorem $(\bar{\mu}, \bar{\sigma}, \bar{\alpha})$ is the parameter values between the MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ and underlying real parameter values $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha} = 0)$. Use multivariate Taylor series expansion of Lagrange form at MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$:

$$
\begin{aligned}
\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}_{3\times 1} &= \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix}_{3\times 1} |_{\mu=\hat{\mu},\sigma=\hat{\sigma},\alpha=\hat{\alpha}} \\
&= \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix}_{3\times 1} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \\
&\quad + \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha^2} \end{pmatrix}_{3\times 3} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}_{3\times 1},
\end{aligned}
$$

where

$$
\begin{pmatrix} \bar{\mu} \\ \bar{\sigma} \\ \bar{\alpha} \end{pmatrix} = \begin{pmatrix} \tilde{\mu} \\ \tilde{\sigma} \\ 0 \end{pmatrix} + t \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}, \qquad t \in (0,1)\,.
$$

To make it simple, we use block matrix to simplify it.

$$
-\begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}
$$

where

$$C_{11} = \begin{pmatrix} \frac{\partial^2 l^\star_{inc}}{\partial \mu^2} & \frac{\partial^2 l^\star_{inc}}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l^\star_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^\star_{inc}}{\partial \sigma^2} \end{pmatrix}, \qquad C_{12} = \begin{pmatrix} \frac{\partial^2 l^\star_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^\star_{inc}}{\partial \sigma \partial \alpha} \end{pmatrix}, \qquad C_{21} = \begin{pmatrix} \frac{\partial^2 l^\star_{inc}}{\partial \alpha \partial \mu} & \frac{\partial^2 l^\star_{inc}}{\partial \alpha \partial \sigma} \end{pmatrix}, \qquad C_{22} = \frac{\partial^2 l^\star_{inc}}{\partial \alpha^2}.$$

Then we denote it as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} .$$

The inverse of it is

$$C^{-1} = \begin{pmatrix} C^-_{11} & C^-_{12} \\ C^-_{21} & C^-_{22} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}$$

where

$$\begin{aligned} C^-_{11} &= C_{11}^{-1} + C_{11}^{-1} C_{12} \left(C_{22} - C_{21} C_{11}^{-1} C_{12}\right)^{-1} C_{21} C_{11}^{-1} \\ C^-_{12} &= -C_{11}^{-1} C_{12} \left(C_{22} - C_{21} C_{11}^{-1} C_{12}\right)^{-1} \\ C^-_{21} &= -\left(C_{22} - C_{21} C_{11}^{-1} C_{12}\right)^{-1} C_{21} C_{11}^{-1} \\ C^-_{22} &= \left(C_{22} - C_{21} C_{11}^{-1} C_{12}\right)^{-1}. \end{aligned}$$

Therefore, we can solve the equation now.

$$-C^{-1} \begin{pmatrix} \frac{\partial l^\star_{inc}}{\partial \mu} \\ \frac{\partial l^\star_{inc}}{\partial \sigma} \\ \frac{\partial l^\star_{inc}}{\partial \alpha} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} = \begin{pmatrix} \hat{\mu} - \tilde{\mu} \\ \hat{\sigma} - \tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}.$$

That is

$$\begin{cases} -C^-_{11} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l^\star_{inc}}{\partial \mu} \\ \frac{\partial l^\star_{inc}}{\partial \sigma} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} - C^-_{12} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \cdot \frac{\partial l^\star_{inc}}{\partial \alpha} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} = \begin{pmatrix} \hat{\mu} - \tilde{\mu} \\ \hat{\sigma} - \tilde{\sigma} \end{pmatrix} \\ -C^-_{21} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l^\star_{inc}}{\partial \mu} \\ \frac{\partial l^\star_{inc}}{\partial \sigma} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} - C^-_{22} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \cdot \frac{\partial l^\star_{inc}}{\partial \alpha} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} = \hat{\alpha}. \end{cases}$$

As $\frac{\partial l^{\star}_{inc}}{\partial \alpha} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}= 0$, then

$$\begin{cases} -C_{11}^{-} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \end{pmatrix} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}= \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \end{pmatrix} \\ -C_{21}^{-} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \end{pmatrix} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}= \hat{\alpha}. \end{cases}$$

For simplicity, we denote

$$C_{11} = \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} a & c \\ c & d \end{pmatrix}$$

where $\bar{A}_i = \bar{\alpha}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}} + \bar{\delta}\sqrt{\frac{2}{\pi}}\right)$, $\bar{B}_i = \frac{y_i-\bar{\mu}}{\bar{\sigma}} + \bar{\delta}\sqrt{\frac{2}{\pi}} + \frac{\bar{\delta}\sqrt{2/\pi}}{1+\bar{\alpha}^2}$ and $\bar{\delta} = \bar{\alpha}\left(1+\bar{\alpha}^2\right)^{-1/2}$, then for convenience we define

$$\begin{aligned} a \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\ &= -\frac{n}{\bar{\sigma}^2} - \frac{\bar{\alpha}^2}{\bar{\sigma}^2}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i - \frac{\bar{\alpha}^2}{\bar{\sigma}^2}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} \\ c \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\ &= -\frac{1}{\bar{\sigma}^2\bar{\alpha}}\sum_{i=1}^{n}\bar{A}_i - \frac{1}{\bar{\sigma}^3}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right) + \frac{\bar{\alpha}}{\bar{\sigma}^2}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\ &\quad -\frac{\bar{\alpha}^2}{\bar{\sigma}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i - \frac{\bar{\alpha}^2}{\bar{\sigma}^3}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} \\ d \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\ &= \frac{n}{\bar{\sigma}^2} - \frac{2}{\bar{\sigma}^3\bar{\alpha}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\bar{A}_i - \frac{1}{\bar{\sigma}^4}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2 \\ &\quad +\frac{2\bar{\alpha}}{\bar{\sigma}^3}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{\bar{\alpha}^2}{\bar{\sigma}^4}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i \\ &\quad -\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}. \end{aligned}$$

We also denote:

$$C_{12} = C_{21}^{T} = \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \alpha} \end{pmatrix} = \begin{pmatrix} b \\ e \end{pmatrix},$$

where

$$
\begin{aligned}
b\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\mu\partial\alpha}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{n\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}} - \frac{1}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} + \frac{1}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2 \\
&\quad + \frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i + \frac{\bar{\alpha}}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i \\
e\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\sigma\partial\alpha}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right) - \frac{1}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad + \frac{1}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2 + \frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i \\
&\quad + \frac{\bar{\alpha}}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i.
\end{aligned}
$$

And

$$
\begin{aligned}
C_{22}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\alpha^2}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= f\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{3\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}}\sum_{i=1}^{n}\bar{A}_i - \frac{2n}{\pi\left(1+\bar{\alpha}^2\right)^3} + \frac{\sqrt{2/\pi}\left(2-\bar{\alpha}^2\right)}{\left(1+\bar{\alpha}^2\right)^{5/2}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad - \sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\bar{B}_i{}^2 - \sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i{}^2 - 2\lambda.
\end{aligned}
$$

As $(\hat{\mu},\hat{\sigma},\hat{\alpha})$ is restricted to $|\hat{\mu}-\tilde{\mu}|\leq\kappa_{\mu n}\to 0$, $|\hat{\sigma}-\tilde{\sigma}|\leq\kappa_{\sigma n}\to 0$ and $|\hat{\alpha}-0|\leq\kappa_{\alpha n}\to 0$, we can also give that

$$
|\bar{\mu}-\tilde{\mu}|\leq\kappa_{\mu n}, \qquad |\bar{\sigma}-\tilde{\sigma}|\leq\kappa_{\sigma n}, \qquad |\bar{\alpha}|\leq\kappa_{\alpha n}.
$$

Then we also have

$$
|\bar{\mu}|\leq|\tilde{\mu}|+\kappa_{\mu n}, \qquad \tilde{\sigma}-\kappa_{\sigma n}\leq\bar{\sigma}\leq\tilde{\sigma}+\kappa_{\sigma n}, \qquad \frac{1}{\bar{\sigma}^2}\leq\frac{1}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}.
$$

For

$$
\begin{aligned}
\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= \left(-\frac{1}{\bar{\sigma}^2}+\frac{1}{\tilde{\sigma}^2}\right) - \left(\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i - 0\right) \\
&\quad - \left(\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - 0\right)
\end{aligned}
$$

Because

$$\left|-\frac{1}{\bar{\sigma}^2}+\frac{1}{\tilde{\sigma}^2}\right| \quad = \quad |\tilde{\sigma}-\bar{\sigma}|\,\frac{\tilde{\sigma}+\bar{\sigma}}{\tilde{\sigma}^2\bar{\sigma}^2} \leq |\tilde{\sigma}-\bar{\sigma}|\,\frac{\tilde{\sigma}+\tilde{\sigma}+\kappa_{\sigma n}}{\tilde{\sigma}^2\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2} \leq O\left(1\right)\cdot\kappa_{\sigma n}.$$

And for function $\phi\left(x\right)/\Phi\left(x\right)$, when $x<0$, we have

$$\frac{\phi\left(x\right)}{\Phi\left(x\right)} < \frac{1+x^2}{|x|}, \qquad \text{thus, } \frac{\phi\left(x\right)}{|x|\,\Phi\left(x\right)} < \frac{1+x^2}{x^2} < +\infty.$$

We can find a positive constant $\beta_0$, when $x \geq -\beta_0$, as $\phi\left(x\right) \leq \frac{1}{\sqrt{2\pi}}$ and $\Phi\left(x\right) \geq \Phi\left(-\beta_0\right)$, so

$$\sup_{x\geq-\beta_0}\frac{\phi\left(x\right)}{\Phi\left(x\right)} \quad \leq \quad \frac{1}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}.$$

Meanwhile, when $x<-\beta_0$

$$\sup_{x<-\beta_0}\frac{\phi\left(x\right)}{|x|\,\Phi\left(x\right)} \quad \leq \quad 1+\frac{1}{\beta_0^2}.$$

Then

$$\begin{aligned}
\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right| &= \frac{|\bar{\alpha}|}{\sqrt{1+\bar{\alpha}^2}}\cdot\sqrt{\frac{2}{\pi}} \\
&= \frac{|\bar{\alpha}|}{\left(1+\bar{\alpha}'^2\right)^{3/2}}\cdot\sqrt{\frac{2}{\pi}} \\
&\quad \text{(Lagrange form of Taylor series expansion at } 0, \bar{\alpha}'\text{is between 0 and } \bar{\alpha}) \\
&= |\bar{\alpha}|\cdot O\left(1\right) \\
&\leq \kappa_{\alpha n}\cdot O\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right| &= |\bar{\alpha}|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)+\bar{\delta}\sqrt{\frac{2}{\pi}}\right| \\
&\leq |\bar{\alpha}|\cdot\left(\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\right) \\
&= |\bar{\alpha}|\cdot\left(\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}+\tilde{\mu}-\bar{\mu}}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\bar{\sigma}}\right)\right|+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\right) \\
&\leq |\bar{\alpha}|\cdot\left(\frac{\tilde{\sigma}}{\bar{\sigma}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\bar{\sigma}}+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\right) \\
&\leq |\bar{\alpha}|\cdot\left(\frac{\tilde{\sigma}}{\tilde{\sigma}-\kappa_{\sigma n}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\tilde{\sigma}-\kappa_{\sigma n}}+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\right) \\
&\leq |\bar{\alpha}|\cdot\left(\frac{\tilde{\sigma}}{\tilde{\sigma}-\kappa_{\sigma n}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}-\kappa_{\sigma n}}+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\right) \\
&\leq \kappa_{\alpha n}\cdot\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O\left(1\right)+\kappa_{\alpha n}\cdot O\left(1\right)\right] \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}$$

Thus, from the result above, we can derive

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}+\tilde{\mu}-\bar{\mu}}{\bar{\sigma}}\right)\right| \\
&\leq \frac{\tilde{\sigma}}{\bar{\sigma}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\bar{\sigma}} \\
&\leq \frac{\tilde{\sigma}}{\tilde{\sigma}-\kappa_{\sigma n}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}-\kappa_{\sigma n}} \\
&\leq O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)\right| &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right| \\
&= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1) \\
\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}+\tilde{\mu}-\bar{\mu}}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\bar{\sigma}}\right)^2 \\
&= \frac{\tilde{\sigma}^2}{\bar{\sigma}^2}\frac{1}{n}\left[\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2+\frac{2(\tilde{\mu}-\bar{\mu})}{\tilde{\sigma}}\cdot\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)+\frac{n(\tilde{\mu}-\bar{\mu})^2}{\tilde{\sigma}^2}\right] \\
&\leq \frac{\tilde{\sigma}^2}{\bar{\sigma}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2+\frac{2|\tilde{\mu}-\bar{\mu}|\,\tilde{\sigma}}{\bar{\sigma}^2}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
&\quad +\frac{(\tilde{\mu}-\bar{\mu})^2}{\bar{\sigma}^2} \\
&\leq \frac{\tilde{\sigma}^2}{(\tilde{\sigma}-\kappa_{\sigma n})^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2+\frac{2|\tilde{\mu}-\bar{\mu}|\,\tilde{\sigma}}{(\tilde{\sigma}-\kappa_{\sigma n})^2}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
&\quad +\frac{(\tilde{\mu}-\bar{\mu})^2}{(\tilde{\sigma}-\kappa_{\sigma n})^2} \\
&= O_p(1)+|\tilde{\mu}-\bar{\mu}|\cdot O_p\left(n^{-1/2}\right)+(\tilde{\mu}-\bar{\mu})^2\cdot O(1) \\
&= O_p(1)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O(1)=O_p(1) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A_i}^2\right| &= \bar{\alpha}^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^2 \\
&\leq \bar{\alpha}^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2+2\left|\bar{\delta}\right|\sqrt{\frac{2}{\pi}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)+\bar{\delta}^2\cdot\frac{2}{\pi}\right] \\
&\leq \bar{\alpha}^2\left[O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O(1)+\kappa_{\alpha n}^2\cdot O(1)\right] \\
&\leq \kappa_{\alpha n}^2\cdot O_p(1)
\end{aligned}$$

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\cdot\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\left|\bar{A}_i\right|\Phi\left(\bar{A}_i\right)}\bar{A}^2\cdot\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right|\cdot\frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right|\cdot\frac{1}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot 1 \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O\left(1\right)+\kappa_{\alpha n}^2\cdot O\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)
\end{aligned}$$

where $\mathbb{I}\left(x\right)$ is an indicator function, $\mathbb{I}\left(x\right)=1$ when $x$ is true, otherwise $\mathbb{I}\left(x\right)=0$. Thus, $\left|\mathbb{I}\left(x\right)=1\right|\leq 1$.

$$\begin{aligned}
\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i-0\right| &\leq \frac{\bar{\alpha}^2}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| \\
&\leq \frac{\bar{\alpha}^2}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}\cdot\left[\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O\left(1\right)+\kappa_{\alpha n}^2\cdot O\left(1\right)\right] \\
&\leq \kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right) \\
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{{\bar{A}_i}^2\Phi\left(\bar{A}_i\right)^2}{\bar{A}_i}^2\right. \\
&\quad \left.\times\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)^2\cdot\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)^2\cdot 1 \\
&\leq O\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
&= O_p\left(1\right) \\
\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}-0\right| &\leq \frac{\bar{\alpha}^2}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
&\leq \frac{\bar{\alpha}^2}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}\cdot O_p\left(1\right) \\
&= \bar{\alpha}^2\cdot O_p\left(1\right) \\
&= \kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}$$

Now we can find

$$\begin{aligned}
\left|\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &\leq \left|-\frac{1}{\bar{\sigma}^2}+\frac{1}{\tilde{\sigma}^2}\right|+\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| \\
&\quad +\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
&\leq O\left(1\right)\cdot\kappa_{\sigma n}+\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O\left(1\right) \\
&\quad +\kappa_{\alpha n}^4\cdot O\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
&= O_p\left(1\right)\cdot\kappa_{\sigma n}+\kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}$$

And

$$\begin{aligned}
\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= \left(\frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right)-\left(\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right) \\
&\quad +\left(\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2-0\right) \\
&\quad +\left(\frac{\bar{\alpha}}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i-0\right) \\
&\quad +\left(\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i-0\right) \\
\left|\frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right| &= \frac{\sqrt{2/\pi}}{\tilde{\sigma}\bar{\sigma}}\left|\frac{\tilde{\sigma}}{\left(1+\bar{\alpha}^2\right)^{3/2}}-\bar{\sigma}\right| \\
&\leq \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(\tilde{\sigma}-\kappa_{\sigma n}\right)}\left|\frac{\tilde{\sigma}}{\left(1+\bar{\alpha}^2\right)^{3/2}}-\bar{\sigma}\right| \\
&= \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(\tilde{\sigma}-\kappa_{\sigma n}\right)}\left|\tilde{\sigma}\left(\frac{1}{\left(1+\bar{\alpha}^2\right)^{3/2}}-1\right)+\left(\tilde{\sigma}-\bar{\sigma}\right)\right| \\
&\leq \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(\tilde{\sigma}-\kappa_{\sigma n}\right)}\left[\tilde{\sigma}\frac{\left|\left(1+\bar{\alpha}^2\right)^{3/2}-1\right|}{\left(1+\bar{\alpha}^2\right)^{3/2}}+\left|\tilde{\sigma}-\bar{\sigma}\right|\right] \\
&= \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(\tilde{\sigma}-\kappa_{\sigma n}\right)}\left[\frac{3\tilde{\sigma}}{2}\left(1+\bar{\alpha}'^2\right)^{-7/2}\left|1-4\bar{\alpha}'^2\right|\bar{\alpha}^2+\left|\tilde{\sigma}-\bar{\sigma}\right|\right] \\
&\quad \text{(Lagrange form of Taylor series, } \bar{\alpha}'\text{is between 0 and } \bar{\alpha}\text{)} \\
&\leq O\left(1\right)\left(O\left(1\right)\cdot\kappa_{\alpha n}^2+\kappa_{\sigma n}\right) \\
&= O\left(1\right)\cdot\kappa_{\alpha n}^2+O\left(1\right)\cdot\kappa_{\sigma n}.
\end{aligned}$$

From Lagrange form of Taylor series at 0, here $\bar{A}'$ is between 0 and $\bar{A}$, we can have

$$\begin{aligned}
\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} &= \frac{\phi\left(0\right)}{\Phi\left(0\right)} + \left[\frac{\phi\left(x\right)}{\Phi\left(x\right)}\right]' \mid_{x=\bar{A}_i'} \cdot \bar{A}_i \\
&= \frac{\phi\left(0\right)}{\Phi\left(0\right)} - \left[\bar{A}_i' \frac{\phi\left(\bar{A}_i'\right)}{\Phi\left(\bar{A}_i'\right)} + \frac{\phi\left(\bar{A}_i'\right)^2}{\Phi\left(\bar{A}_i'\right)^2}\right] \bar{A}_i \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}+\tilde{\mu}-\bar{\mu}}{\tilde{\sigma}}\cdot\frac{\tilde{\sigma}}{\bar{\sigma}}\right)^3\right| \\
&= \frac{\tilde{\sigma}^3}{\bar{\sigma}^3}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3 + \frac{3\left(\tilde{\mu}-\bar{\mu}\right)^2}{\tilde{\sigma}^2}\cdot\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right. \\
&\quad \left. + \frac{3\left(\tilde{\mu}-\bar{\mu}\right)}{\tilde{\sigma}}\cdot\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 + \frac{n\left(\tilde{\mu}-\bar{\mu}\right)^3}{\tilde{\sigma}^3}\right| \\
&\leq \frac{\tilde{\sigma}^3}{\bar{\sigma}^3}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right| + \frac{3\left(\tilde{\mu}-\bar{\mu}\right)^2\tilde{\sigma}}{\bar{\sigma}^3}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
&\quad + \frac{3\left|\tilde{\mu}-\bar{\mu}\right|\tilde{\sigma}^2}{\bar{\sigma}^3}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 + \frac{\left|\tilde{\mu}-\bar{\mu}\right|^3}{\bar{\sigma}^3} \\
&\leq \frac{\tilde{\sigma}^3}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^3}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right| + \frac{3\left(\tilde{\mu}-\bar{\mu}\right)^2\tilde{\sigma}}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^3}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
&\quad + \frac{3\left|\tilde{\mu}-\bar{\mu}\right|\tilde{\sigma}^2}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^3}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 + \frac{\left|\tilde{\mu}-\bar{\mu}\right|^3}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^3} \\
&\leq O\left(1\right)\cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}^2\cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\cdot O_p\left(1\right) + \kappa_{\mu n}^3\cdot O\left(1\right) \\
&= O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\cdot O_p\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^3\right| &= \left|\bar{\alpha}\right|^3\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^3\right| \\
&= \left|\bar{\alpha}\right|^3\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3 + 3\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2 + \bar{\delta}^2\cdot\frac{6}{\pi}\cdot\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right.\right. \\
&\quad \left.\left. + \bar{\delta}^3\left(\frac{2}{\pi}\right)^{3/2}\right]\right| \\
&\leq \left|\bar{\alpha}\right|^3\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right| + 3\left|\bar{\delta}\right|\sqrt{\frac{2}{\pi}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|\right. \\
&\quad \left. + \bar{\delta}^2\cdot\frac{6}{\pi}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right| + \left|\bar{\delta}\right|^3\left(\frac{2}{\pi}\right)^{3/2}\right] \\
&= \kappa_{\alpha n}^3\cdot\left[O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\cdot O_p\left(1\right) + \kappa_{\alpha n}\cdot O_p\left(1\right)\right. \\
&\quad \left. + \kappa_{\alpha n}^2\left(O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\cdot O\left(1\right)\right) + \kappa_{\alpha n}^3 O\left(1\right)\right] \\
&= \kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right) + \kappa_{\alpha n}^4\cdot O_p\left(1\right)
\end{aligned}$$

$$
\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\left[\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left[\bar{A_i}'\frac{\phi\left(\bar{A_i}'\right)}{\Phi\left(\bar{A_i}'\right)}+\frac{\phi\left(\bar{A_i}'\right)^2}{\Phi\left(\bar{A_i}'\right)^2}\right]\bar{A}_i\right| \\
&= \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A_i}'\right)}{\Phi\left(\bar{A_i}'\right)}\left[\frac{\phi\left(\bar{A_i}'\right)}{\Phi\left(\bar{A_i}'\right)}+\bar{A_i}'\right]\cdot\bar{A}_i\right| \\
&\quad\times\left[\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right] \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A_i}'\right)}{\Phi\left(\bar{A_i}'\right)}\left[\frac{\phi\left(\bar{A_i}'\right)}{\Phi\left(\bar{A_i}'\right)}+\bar{A_i}'\right]\cdot\bar{A}_i\right|\cdot\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right| \\
&\quad+\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A_i}'\right)}{\bar{A_i}'\Phi\left(\bar{A_i}'\right)}\left[\frac{\phi\left(\bar{A_i}'\right)}{\bar{A_i}'\Phi\left(\bar{A_i}'\right)}+1\right]\cdot\left(\bar{A_i}'\right)^2\cdot\bar{A}_i\right| \\
&\quad\times\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \max\left\{\frac{1}{\sqrt{2\pi}\Phi\left(-\beta_0\right)}\left|\frac{1}{\sqrt{2\pi}\Phi\left(-\beta_0\right)}+\bar{A_i}'\right|,\frac{\bar{A_i}'^2}{4}\right\}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right| \\
&\quad\times\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|+\max\left\{\left(1+\frac{1}{\beta_0^2}\right)\left|\left(1+\frac{1}{\beta_0^2}\right)+1\right|,\frac{1}{4}\right\} \\
&\quad\times\frac{1}{n}\left|\sum_{i=1}^{n}\left(\bar{A_i}'\right)^2\cdot\bar{A}_i\right|\cdot\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq O\left(1\right)\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right|\cdot1+O\left(1\right)\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i^3\right|\cdot1 \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
&\quad+\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right) \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\cdot\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\bar{A}_i\Phi\left(\bar{A}_i\right)}\bar{A}_i\right. \\
&\quad\left.\times\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot1 \\
&\leq O\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
&= O_p\left(1\right) \\
\left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right| &= \left|\left(\frac{1}{\bar{\sigma}}-\frac{1}{\tilde{\sigma}}\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}+\frac{1}{\tilde{\sigma}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\sqrt{\frac{2}{\pi}}\right)\right| \\
&\leq \frac{\left|\tilde{\sigma}-\bar{\sigma}\right|}{\tilde{\sigma}\bar{\sigma}}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right|+\frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right| \\
&\leq \kappa_{\sigma n}O_p\left(1\right)+\kappa_{\alpha n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}O_p\left(1\right)+\kappa_{\alpha n}^2O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}{\bar{A}_i}^2\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right){\bar{A}_i}^2}{\Phi\left(\bar{A}_i\right)}\mathbb{I}\left(\bar{A}_i \geq -\beta_0\right) + \frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right){\bar{A}_i}^3}{\bar{A}_i\Phi\left(\bar{A}_i\right)}\mathbb{I}\left(\bar{A}_i < -\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\frac{\left|\mathbb{I}\left(\bar{A}_i \geq -\beta_0\right)\right|}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)} + \frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^3\right|\left(1+\frac{1}{\beta_0^2}\right)\left|\mathbb{I}\left(\bar{A}_i < -\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\right|\cdot\frac{1}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)} + \frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^3\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot 1 \\
&\leq \kappa_{\alpha n}^2\cdot O_p\left(1\right) + \kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right) + \kappa_{\alpha n}^4\cdot O_p\left(1\right) \\
&= \kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
\left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}{\bar{A}_i}^2\right| &\leq \frac{1}{\tilde{\sigma}-\kappa_{\sigma n}}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}{\bar{A}_i}^2\right| \\
&\leq \kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}
$$

From Lagrange form of Taylor series at 0, here $\bar{\alpha}'$ is between 0 and $\bar{\alpha}$, we can have

$$\begin{aligned}
\frac{\bar{\alpha}^2}{\left(1+\bar{\alpha}^2\right)^{3/2}} &= 0+0+\frac{1}{2}\cdot\frac{2-11\bar{\alpha}'^2+2\bar{\alpha}'^4}{\left(1+\bar{\alpha}'^2\right)^{7/2}}\cdot\bar{\alpha}^2 \\
&= \bar{\alpha}^2\cdot O\left(1\right) \\
&\leq \kappa_{\alpha n}^2\cdot O\left(1\right) \\
\left|\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| &\leq \left|\frac{\sqrt{2/\pi}}{\tilde{\sigma}-c_\sigma\frac{1}{\sqrt{n}}}\right|\cdot\frac{\bar{\alpha}^2}{\left(1+\bar{\alpha}^2\right)^{3/2}}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| \\
&\leq O\left(1\right)\kappa_{\alpha n}^2\left[\kappa_{\alpha n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}O_p\left(1\right)+\kappa_{\alpha n}^2O_p\left(1\right)\right] \\
&= \kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\bar{B}_i\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}+\frac{\bar{\delta}\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right)\right| \\
&\leq \frac{\tilde{\sigma}}{\tilde{\sigma}-\kappa_{\sigma n}}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\tilde{\mu}-\bar{\mu}|}{\tilde{\sigma}-\kappa_{\sigma n}} \\
&\quad +\left(\sqrt{\frac{2}{\pi}}+\frac{\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right)|\bar{\delta}| \\
&= O\left(1\right)\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O\left(1\right)+O\left(1\right)\cdot\kappa_{\alpha n}\cdot O\left(1\right) \\
&= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^3\right| &= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\bar{B}_i\right| &= \bar{\alpha}^2\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^2\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}+\frac{\bar{\delta}\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right)\right| \\
&\leq \bar{\alpha}^2\cdot\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^3\right|+\left|\frac{\bar{\delta}\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right|\right. \\
&\quad \left.\times\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^2\right|\right] \\
&\leq \kappa_{\alpha n}^2\cdot\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\right. \\
&\quad \left.+\kappa_{\alpha n}\cdot O\left(1\right)\cdot O_p\left(1\right)\right] \\
&= \kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)
\end{aligned}$$

$$
\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\cdot\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right. \\
&\quad \left.+\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{{\bar{A}_i}^2\Phi\left(\bar{A}_i\right)^2}{\bar{A}_i}^2\bar{B}_i\cdot\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\bar{B}_i\right|\cdot\frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\bar{B}_i\right| \\
&\quad \times\left(1+\frac{1}{\beta_0^2}\right)^2\cdot\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\bar{B}_i\right|\cdot\frac{1}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}{\bar{A}_i}^2\bar{B}_i\right|\left(1+\frac{1}{\beta_0^2}\right)^2\cdot 1 \\
&\leq O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right) \\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right) \\
&= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right) \\
\left|\frac{\bar{\alpha}}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\right| &\leq \frac{|\bar{\alpha}|}{\tilde{\sigma}-\kappa_{\sigma n}}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\right| \\
&\leq \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\left|\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| \\
\leq\ &\left|\frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right|+\left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\right|+\left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}{\bar{A}_i}^2\right| \\
&+\left|\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right|+\left|\frac{\bar{\alpha}}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\right| \\
\leq\ &O\left(1\right)\cdot\kappa_{\alpha n}^2+O\left(1\right)\cdot\kappa_{\sigma n}+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right) \\
&+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right) \\
&+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
=\ &O_p\left(1\right)\cdot\kappa_{\alpha n}^2+O_p\left(1\right)\cdot\kappa_{\sigma n}+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right).
\end{aligned}
$$

Then we calculate $c/n$

$$
\begin{aligned}
&\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \\
=\ &\left[-\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&-\left[\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]+\left[\frac{\bar{\alpha}}{\bar{\sigma}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-0\right] \\
&-\left[\frac{\bar{\alpha}^2}{\bar{\sigma}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i-0\right]-\left[\frac{\bar{\alpha}^2}{\bar{\sigma}^3}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}-0\right].
\end{aligned}
$$

As

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}-\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i\tilde{\sigma}-\bar{\mu}\tilde{\sigma}-y_i\bar{\sigma}+\bar{\sigma}\tilde{\mu}}{\bar{\sigma}\tilde{\sigma}}\right)\right| \\
&= \frac{1}{\bar{\sigma}\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(y_i\tilde{\sigma}-y_i\bar{\sigma}-\bar{\mu}\tilde{\sigma}+\tilde{\sigma}\tilde{\mu}-\tilde{\sigma}\tilde{\mu}+\bar{\sigma}\tilde{\mu}\right)\right| \\
&= \frac{1}{\bar{\sigma}\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[y_i\cdot(\tilde{\sigma}-\bar{\sigma})+\tilde{\sigma}(\tilde{\mu}-\bar{\mu})-\tilde{\mu}\cdot(\tilde{\sigma}-\bar{\sigma})\right]\right| \\
&= \frac{1}{\bar{\sigma}\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[(y_i-\tilde{\mu})\cdot(\tilde{\sigma}-\bar{\sigma})+\tilde{\sigma}(\tilde{\mu}-\bar{\mu})\right]\right| \\
&\leq \frac{1}{(\tilde{\sigma}-\kappa_{\sigma n})}\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|\cdot|\tilde{\sigma}-\bar{\sigma}|+|\tilde{\mu}-\bar{\mu}|\right] \\
&\leq O(1)\cdot\left[O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+\kappa_{\mu n}\right] \\
&= O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p(1)\cdot\kappa_{\mu n}
\end{aligned}$$

$$\begin{aligned}
&\left|-\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
= &\left|-\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)\right. \\
&\left.-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
\leq &\left|\frac{1}{\tilde{\sigma}^2}-\frac{1}{\bar{\sigma}^2}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)\right|+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}-\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\frac{|\bar{\delta}|}{\tilde{\sigma}^2}\sqrt{\frac{2}{\pi}} \\
\leq &O(1)\cdot\kappa_{\sigma n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1)\right]+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p(1)\cdot\kappa_{\mu n} \\
&+\kappa_{\sigma n}\cdot O(1) \\
= &O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p(1)\cdot\kappa_{\mu n}+O_p(1)\cdot\kappa_{\alpha n}
\end{aligned}$$

$$\begin{aligned}
&\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
\leq &\left|\frac{1}{\tilde{\sigma}^2}-\frac{1}{\bar{\sigma}^2}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}-\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
\leq &O(1)\cdot\kappa_{\sigma n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)\right]+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p(1)\cdot\kappa_{\mu n} \\
= &O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p(1)\cdot\kappa_{\mu n}
\end{aligned}$$

$$\begin{aligned}
\left|\frac{\bar{\alpha}}{\bar{\sigma}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-0\right| &\leq |\bar{\alpha}|\cdot\frac{1}{\left(\tilde{\sigma}-c_\sigma\frac{1}{\sqrt{n}}\right)^2}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| \\
&\leq \kappa_{\alpha n}\cdot O\left(1\right)\cdot O_p\left(1\right) \\
&= \kappa_{\alpha n}\cdot O_p\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i\right| &= |\bar{\alpha}|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2+\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right| \\
&\leq |\bar{\alpha}|\cdot\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|+\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\right) \\
&\leq \kappa_{\alpha n}\left[O_p\left(1\right)+O\left(1\right)\cdot\kappa_{\alpha n}\left(O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)\right)\right] \\
&= O_p\left(1\right)\cdot\kappa_{\alpha n}
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\right| \\
= \;& \bar{\alpha}^2\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3+2\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2+\bar{\delta}^2\cdot\frac{2}{\pi}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right| \\
\leq \;& \bar{\alpha}^2\cdot\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right|+2\left|\bar{\delta}\right|\sqrt{\frac{2}{\pi}}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|+\bar{\delta}^2\cdot\frac{2}{\pi}\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\right) \\
\leq \;& \kappa_{\alpha n}^2\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\left(O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)\right)\right] \\
= \;& \kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)
\end{aligned}$$

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| \\
\leq \;& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\left|\bar{A}_i\right|\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
\leq \;& \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i\right|\frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\right|\left(1+\frac{1}{\beta_0^2}\right)\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
\leq \;& \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i\right|\cdot\frac{1}{\sqrt{2\pi}\cdot\Phi\left(-\beta_0\right)}+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)\cdot 1 \\
\leq \;& O_p\left(1\right)\cdot\kappa_{\alpha n}+\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right) \\
= \;& O_p\left(1\right)\cdot\kappa_{\alpha n}
\end{aligned}$$

$$\begin{aligned}
\left|\frac{\bar{\alpha}^2}{\bar{\sigma}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i-0\right| &= \bar{\alpha}^2\cdot\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| \\
&\leq \kappa_{\alpha n}^3\cdot O_p\left(1\right)
\end{aligned}$$

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
= & \left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)^2}{{\bar{A}_i}^2\Phi\left(\bar{A}_i\right)^2}{\bar{A}_i}^2\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
\leq & \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\frac{\left|\mathbb{I}\left(\bar{A}_i\geq-\beta_0\right)\right|}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right){\bar{A}_i}^2\right|\left(1+\frac{1}{\beta_0^2}\right)^2\left|\mathbb{I}\left(\bar{A}_i<-\beta_0\right)\right| \\
\leq & \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\cdot\frac{1}{2\pi\cdot\Phi\left(-\beta_0\right)^2}+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right){\bar{A}_i}^2\right|\cdot\left(1+\frac{1}{\beta_0^2}\right)^2\cdot 1 \\
\leq & O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right) \\
= & O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^3}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}-0\right| & \leq \frac{\bar{\alpha}^2}{\bar{\sigma}^2}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
& \leq \kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^5\cdot O_p\left(1\right).
\end{aligned}
$$

Then we can reach

$$
\begin{aligned}
& \left|\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| \\
\leq & \left|-\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
& +\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|+\left|\frac{\bar{\alpha}}{\bar{\sigma}^2}\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| \\
& +\left|\frac{\bar{\alpha}^2}{\bar{\sigma}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right|+\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^3}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
\leq & O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\mu n}+O_p\left(1\right)\cdot\kappa_{\alpha n}+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n} \\
& +O_p\left(1\right)\cdot\kappa_{\mu n}+\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right) \\
& +\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^5\cdot O_p\left(1\right) \\
= & O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\mu n}+O_p\left(1\right)\cdot\kappa_{\alpha n}.
\end{aligned}
$$

Moreover, we calculate $d/n$

$$\begin{aligned}
&\frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \\
= \quad & \left(\frac{1}{\bar{\sigma}^2}-\frac{1}{\tilde{\sigma}^2}\right) - \left[\frac{2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right) - \frac{2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] \\
& - \left[\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2 - \frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] + \left[\frac{2\bar{\alpha}}{\bar{\sigma}^3}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - 0\right] \\
& - \left[\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i - 0\right] - \left[\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - 0\right]
\end{aligned}$$

$$\begin{aligned}
\left|\frac{1}{\bar{\sigma}^2}-\frac{1}{\tilde{\sigma}^2}\right| &= |\tilde{\sigma}-\bar{\sigma}|\frac{\tilde{\sigma}+\bar{\sigma}}{\tilde{\sigma}^2\bar{\sigma}^2} \leq |\tilde{\sigma}-\bar{\sigma}|\frac{\tilde{\sigma}+\tilde{\sigma}+\kappa_{\sigma n}}{\tilde{\sigma}^2\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2} \leq \kappa_{\sigma n}\cdot O\left(1\right) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i\tilde{\sigma}-\tilde{\sigma}\bar{\mu}+y_i\bar{\sigma}-\bar{\sigma}\tilde{\mu}}{\bar{\sigma}\tilde{\sigma}}\right)\right| \\
&= \frac{1}{\bar{\sigma}\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(y_i\tilde{\sigma}-\tilde{\sigma}\bar{\mu}+\tilde{\sigma}\tilde{\mu}-\tilde{\sigma}\tilde{\mu}+y_i\bar{\sigma}-\bar{\sigma}\tilde{\mu}\right)\right| \\
&= \frac{1}{\bar{\sigma}\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[y_i\left(\tilde{\sigma}+\bar{\sigma}\right)+\tilde{\sigma}\left(\tilde{\mu}-\bar{\mu}\right)-\tilde{\mu}\left(\tilde{\sigma}+\bar{\sigma}\right)\right]\right| \\
&\leq \frac{1}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)\tilde{\sigma}}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\left(\tilde{\sigma}+\bar{\sigma}\right)\right| + \tilde{\sigma}\left|\tilde{\mu}-\bar{\mu}\right|\right] \\
&\leq \frac{1}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)}\left[\left(\tilde{\sigma}+\bar{\sigma}\right)\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| + \left|\tilde{\mu}-\bar{\mu}\right|\right] \\
&\leq O\left(1\right)\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\right] \\
&= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)
\end{aligned}$$

$$
\begin{aligned}
& \frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right| \\
= \; & \frac{1}{\bar{\sigma}^2\tilde{\sigma}^2}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(y_i^2\tilde{\sigma}^2-2y_i\bar{\mu}\tilde{\sigma}^2+\bar{\mu}^2\tilde{\sigma}^2-\bar{\sigma}^2y_i^2+2y_i\tilde{\mu}\bar{\sigma}^2-\bar{\sigma}^2\tilde{\mu}^2\right)\right| \\
= \; & \frac{1}{\bar{\sigma}^2\tilde{\sigma}^2}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[y_i^2\left(\tilde{\sigma}^2-\bar{\sigma}^2\right)-2y_i\bar{\mu}\tilde{\sigma}^2+2y_i\tilde{\mu}\tilde{\sigma}^2-2y_i\tilde{\mu}\tilde{\sigma}^2+\bar{\mu}^2\tilde{\sigma}^2+2y_i\tilde{\mu}\bar{\sigma}^2-\bar{\sigma}^2\tilde{\mu}^2\right]\right| \\
= \; & \frac{1}{\bar{\sigma}^2\tilde{\sigma}^2}\frac{1}{n}\left|\sum_{i=1}^{n}\left[y_i^2\left(\tilde{\sigma}^2-\bar{\sigma}^2\right)+2y_i\tilde{\sigma}^2\left(\tilde{\mu}-\bar{\mu}\right)-2y_i\tilde{\mu}\left(\tilde{\sigma}^2-\bar{\sigma}^2\right)+\bar{\mu}^2\tilde{\sigma}^2-\tilde{\mu}^2\tilde{\sigma}^2+\tilde{\mu}^2\tilde{\sigma}^2-\bar{\sigma}^2\tilde{\mu}^2\right]\right| \\
= \; & \frac{1}{\bar{\sigma}^2}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\left(\tilde{\sigma}^2-\bar{\sigma}^2\right)+\left(\tilde{\mu}-\bar{\mu}\right)\left(y_i-\tilde{\mu}+y_i-\bar{\mu}\right)\right]\right| \\
\leq \; & \frac{1}{\bar{\sigma}^2}\left[\frac{\left|\tilde{\sigma}-\bar{\sigma}\right|\left(\tilde{\sigma}+\bar{\sigma}\right)}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|+\frac{\tilde{\sigma}\left|\tilde{\mu}-\bar{\mu}\right|}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right|\right. \\
& \left.+\frac{\bar{\sigma}\left|\tilde{\mu}-\bar{\mu}\right|}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\right] \\
\leq \; & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O\left(1\right)\right] \\
= \; & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
= \; & \left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2+\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq \; & \left|\frac{1}{\bar{\sigma}^2}-\frac{1}{\tilde{\sigma}^2}\right|\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2+\frac{1}{\tilde{\sigma}^2}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq \; & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right) \\
= \; & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)-\frac{2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq \; & 2\cdot\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right|+\left|\bar{\delta}\right|\cdot\left|\frac{2\sqrt{2/\pi}}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right| \\
\leq \; & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| \\
\leq \; & \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\cdot O\left(1\right)+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i\right|\cdot O\left(1\right) \\
\leq \; & O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
\left|\frac{2\bar{\alpha}}{\bar{\sigma}^3}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| &\leq \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\bar{A}_i\right| &\leq |\bar{\alpha}|\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right|+\left|\bar{\delta}\right|\sqrt{\frac{2}{\pi}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|\right]\\
&\leq \kappa_{\alpha n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\right]\\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
&\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2{\bar{A}_i}^2\right|\\
\leq\ & \bar{\alpha}^2\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^4\right|+\left|\bar{\delta}\right|\sqrt{\frac{8}{\pi}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right|+\bar{\delta}^2\frac{2}{\pi}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|\right]\\
\leq\ & \kappa_{\alpha n}^2\left[O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\right]\\
=\ & \kappa_{\alpha n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
&\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right|\\
\leq\ & \bar{\alpha}^2\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\bar{A}_i\right|\cdot O\left(1\right)+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2{\bar{A}_i}^2\right|\cdot O\left(1\right)\right]\\
\leq\ & \kappa_{\alpha n}^2\left[\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\right]\\
=\ & \kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
&\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right|\\
\leq\ & \bar{\alpha}^2\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|\cdot O\left(1\right)+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2{\bar{A}_i}^2\right|\cdot O\left(1\right)\right]\\
\leq\ & \kappa_{\alpha n}^2\left[O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\right]\\
=\ & \kappa_{\alpha n}^2\cdot O_p\left(1\right).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
& \left|\frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| \\
\leq & \left|\frac{1}{\bar{\sigma}^2}-\frac{1}{\tilde{\sigma}^2}\right|+\left|\frac{2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)-\frac{2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
& +\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right|+\left|\frac{2\bar{\alpha}}{\bar{\sigma}^3}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right| \\
& +\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right|+\left|\frac{\bar{\alpha}^2}{\bar{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)^2\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
\leq & \kappa_{\sigma n}\cdot O\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}O_p\left(n^{-1/2}\right) \\
& +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right) \\
& +\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right) \\
& +\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
= & \kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right) \\
& +\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{1}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{1}{\tilde{\sigma}}\right| \\
= & \left|\frac{1}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{1}{\tilde{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}+\frac{1}{\tilde{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{1}{\tilde{\sigma}}\right| \\
= & \left|\frac{1}{\left(1+\bar{\alpha}^2\right)^{3/2}}\cdot\left(\frac{1}{\bar{\sigma}}-\frac{1}{\tilde{\sigma}}\right)+\frac{1}{\tilde{\sigma}}\cdot\left[\frac{1}{\left(1+\bar{\alpha}^2\right)^{3/2}}-1\right]\right| \\
\leq & O\left(1\right)\cdot\frac{\left|\bar{\sigma}-\tilde{\sigma}\right|}{\bar{\sigma}\tilde{\sigma}}+\frac{1}{\tilde{\sigma}}\cdot O\left(1\right)\cdot\bar{\alpha}^2 \\
& \text{use Lagrange form of Taylor series expansion at } \bar{\alpha}=0 \\
\leq & O\left(1\right)\cdot\kappa_{\sigma n}+O\left(1\right)\cdot\kappa_{\alpha n}^2
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
= & \frac{\sqrt{2/\pi}}{n}\left|\sum_{i=1}^{n}\left[\frac{1}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{1}{\tilde{\sigma}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)+\frac{1}{\tilde{\sigma}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)-\frac{1}{\tilde{\sigma}}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]\right| \\
\leq & \sqrt{2/\pi}\left|\frac{1}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}-\frac{1}{\tilde{\sigma}}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|+\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}-\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
\leq & \left[O\left(1\right)\cdot\kappa_{\sigma n}+O\left(1\right)\cdot\kappa_{\alpha n}^2\right]\cdot\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)\right]+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\mu n} \\
= & O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(n^{-1/2}\right)\cdot\kappa_{\alpha n}^2+O_p\left(1\right)\cdot\kappa_{\mu n}.
\end{aligned}
$$

Here we again use Lagrange form of Taylor series expansion at $\bar{A}_i = 0$ to reach

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\left[\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| &= O\left(1\right)\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i{}^2\right| \leq O_p\left(1\right)\cdot\kappa_{\alpha n}^2 \\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left[\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| &= |\bar{\alpha}|\cdot O_p\left(1\right) \leq O_p\left(1\right)\cdot\kappa_{\alpha n}.
\end{aligned}$$

Then

$$\begin{aligned}
&\frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| \\
= &\frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(0\right)}{\Phi\left(0\right)}+\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(0\right)}{\Phi\left(0\right)}-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| \\
\leq &\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left[\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right|+\frac{\sqrt{2/\pi}}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}-\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
\leq &O_p\left(1\right)\cdot\kappa_{\alpha n}+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\mu n}
\end{aligned}$$

$$\begin{aligned}
&\left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right| \\
= &\frac{1}{n}\left|\sum_{i=1}^{n}\left[\frac{1}{\bar{\sigma}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{1}{\tilde{\sigma}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}+\frac{1}{\tilde{\sigma}}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right.\right. \\
&\left.\left.-\frac{\sqrt{2/\pi}}{\tilde{\sigma}}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]\right| \\
\leq &\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\right|\cdot\frac{|\tilde{\sigma}-\bar{\sigma}|}{\tilde{\sigma}\bar{\sigma}}+\frac{1}{\tilde{\sigma}n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(0\right)}{\Phi\left(0\right)}\right]\right| \\
\leq &\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\right]\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\alpha n}+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n} \\
&+O_p\left(1\right)\cdot\kappa_{\mu n} \\
= &O_p\left(1\right)\cdot\kappa_{\alpha n}+O_p\left(n^{-1/2}\right)\cdot\kappa_{\sigma n}+O_p\left(1\right)\cdot\kappa_{\mu n}
\end{aligned}$$

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^3\right| &\leq |\bar{\alpha}|^3\cdot\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^4\right|+3\sqrt{\frac{2}{\pi}}\left|\bar{\delta}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right|\right.\\
&\quad\left.+\frac{6}{\pi}\bar{\delta}^2\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|+\left(\frac{2}{\pi}\right)^{3/2}\left|\bar{\delta}\right|^3\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\right]\\
&\leq \kappa_{\alpha n}^3\left\{O_p\left(1\right)+\kappa_{\alpha n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}O_p\left(1\right)\right]+\kappa_{\alpha n}^2O_p\left(1\right)\right.\\
&\quad\left.+\kappa_{\alpha n}^3\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}O_p\left(1\right)\right]\right\}\\
&= \kappa_{\alpha n}^3\cdot O_p\left(1\right)\\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2\right| &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\right|\cdot O\left(1\right)+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^3\right|\cdot O\left(1\right)\\
&\leq \kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)\\
&\quad+\kappa_{\alpha n}^3\cdot O_p\left(1\right)\\
&= \kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)
\end{aligned}$$

$$\begin{aligned}
\left|\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right| &\leq \bar{\alpha}^2\left[O\left(1\right)\cdot O_p\left(1\right)\cdot\kappa_{\alpha n}\right]\\
&\leq O_p\left(1\right)\cdot\kappa_{\alpha n}^3\\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{B}_i\right| &= \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\frac{2+\bar{\alpha}^2}{1+\bar{\alpha}^2}\right)\right|\\
&\leq \frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|+\left|\bar{\delta}\right|\cdot\sqrt{\frac{2}{\pi}}\cdot\frac{2+\bar{\alpha}^2}{1+\bar{\alpha}^2}\\
&\quad\times\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\\
&\leq O_p\left(1\right)+\kappa_{\alpha n}\cdot O\left(1\right)\cdot\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)\right]\\
&= O_p\left(1\right)\\
\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\bar{B}_i\right| &= \bar{\alpha}^2\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)^2\right.\\
&\quad\left.\times\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\frac{2+\bar{\alpha}^2}{1+\bar{\alpha}^2}\right)\right|\\
&\leq \bar{\alpha}^2\cdot O_p\left(1\right)\\
&= \kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
\left|\frac{\bar{\alpha}}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i\right| &\leq |\bar{\alpha}|\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{B}_i\right|\cdot O\left(1\right)\right.\\
&\quad\left.+\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\bar{A}_i{}^2\bar{B}_i\right|\cdot O\left(1\right)\right]\\
&\leq \kappa_{\alpha n}\left[O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\right]\\
&= O_p\left(1\right)\cdot\kappa_{\alpha n}
\end{aligned}$$

$$
\begin{aligned}
& \left| \frac{e}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{e}{n} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \right| \\
\leq & \left| \frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{\mu}}{\bar{\sigma}} \right) - \frac{\sqrt{2/\pi}}{\tilde{\sigma}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \right| \\
& + \left| \frac{1}{\bar{\sigma}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{\mu}}{\bar{\sigma}} \right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{\sqrt{2/\pi}}{\tilde{\sigma}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \right| + \left| \frac{1}{\bar{\sigma}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{\mu}}{\bar{\sigma}} \right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} {\bar{A}_i}^2 \right| \\
& + \left| \frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{\mu}}{\bar{\sigma}} \right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i \right| + \left| \frac{\bar{\alpha}}{\bar{\sigma}} \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i - \bar{\mu}}{\bar{\sigma}} \right) \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} \bar{B}_i \right| \\
\leq & O_p\left(n^{-1/2}\right) \cdot \kappa_{\sigma n} + O_p\left(n^{-1/2}\right) \cdot \kappa_{\alpha n}^2 + O_p\left(1\right) \cdot \kappa_{\mu n} + O_p\left(1\right) \cdot \kappa_{\alpha n} + O_p\left(n^{-1/2}\right) \cdot \kappa_{\sigma n} \\
& + O_p\left(1\right) \cdot \kappa_{\mu n} + \kappa_{\alpha n}^2 \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n}^2 \cdot O_p\left(1\right) + \kappa_{\alpha n}^3 \cdot O_p\left(1\right) + O_p\left(1\right) \cdot \kappa_{\alpha n}^3 \\
& + O_p\left(1\right) \cdot \kappa_{\alpha n} \\
= & O_p\left(n^{-1/2}\right) \cdot \kappa_{\sigma n} + O_p\left(1\right) \cdot \kappa_{\mu n} + O_p\left(1\right) \cdot \kappa_{\alpha n}
\end{aligned}
$$

$$
\left| \frac{2}{\pi\left(1+\bar{\alpha}^2\right)^3} - \frac{2}{\pi} \right| = O\left(1\right) \cdot \bar{\alpha}^2 \leq \kappa_{\alpha n}^2 \cdot O\left(1\right)
$$

$$
\begin{aligned}
& \left| \frac{\sqrt{2/\pi}\left(2-\bar{\alpha}^2\right)}{\left(1+\bar{\alpha}^2\right)^{5/2}} \frac{1}{n} \sum_{i=1}^{n} \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{4}{\pi} \right| \\
= & \frac{\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}} \frac{1}{n} \sum_{i=1}^{n} \left[ \left(2-\bar{\alpha}^2\right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - 2\sqrt{\frac{2}{\pi}} \left(1+\bar{\alpha}^2\right)^{5/2} \right] \\
= & \frac{\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}} \frac{1}{n} \sum_{i=1}^{n} \left[ \left(2-\bar{\alpha}^2\right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \left(2-\bar{\alpha}^2\right) \sqrt{\frac{2}{\pi}} + \left(2-\bar{\alpha}^2\right) \sqrt{\frac{2}{\pi}} - 2\sqrt{\frac{2}{\pi}} \left(1+\bar{\alpha}^2\right)^{5/2} \right] \\
\leq & \frac{\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}} \cdot \left[ \left|2-\bar{\alpha}^2\right| \cdot \left| \frac{1}{n} \sum_{i=1}^{n} \left( \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{\phi\left(0\right)}{\Phi\left(0\right)} \right) \right| + \sqrt{\frac{2}{\pi}} \left| \left(2-\bar{\alpha}^2\right) - 2\left(1+\bar{\alpha}^2\right)^{5/2} \right| \right] \\
\leq & \kappa_{\alpha n} \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n} \cdot O_p\left(1\right) + \kappa_{\alpha n}^2 \cdot O_p\left(1\right) + \kappa_{\alpha n}^2 \cdot O\left(1\right) \\
= & \kappa_{\alpha n} \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n} \cdot O_p\left(1\right) + \kappa_{\alpha n}^2 \cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{n} \left| \sum_{i=1}^{n} \left[ \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - \frac{\phi\left(0\right)^2}{\Phi\left(0\right)^2} \right] {\bar{A}_i}^2 \right| &= O\left(1\right) \cdot \frac{1}{n} \left| \sum_{i=1}^{n} {\bar{A}_i}^3 \right| \\
&\leq \kappa_{\alpha n}^3 \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n}^3 \cdot O_p\left(1\right) + \kappa_{\alpha n}^4 \cdot O_p\left(1\right) \\
\frac{1}{n} \left| \sum_{i=1}^{n} \left[ \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - \frac{\phi\left(0\right)^2}{\Phi\left(0\right)^2} \right] \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right)^2 \right| &= \kappa_{\alpha n} \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}\kappa_{\alpha n} \cdot O_p\left(1\right) + \kappa_{\alpha n}^2 \cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right| \\
\leq\ & \frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right| O\left(1\right)+\frac{1}{n}\left|\sum_{i=1}^{n}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\bar{A_i}^2\right| O\left(1\right) \\
\leq\ & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
& +\kappa_{\sigma n}\kappa_{\alpha n}^3\cdot O_p\left(1\right) \\
=\ & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}{\bar{B}_i}^2-\frac{2}{\pi}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
=\ & \frac{1}{n}\left|\sum_{i=1}^{n}\left[\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}{\bar{B}_i}^2-\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2+\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-\frac{\phi\left(0\right)^2}{\Phi\left(0\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right| \\
\leq\ & \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left[{\bar{B}_i}^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right|+\frac{1}{n}\left|\sum_{i=1}^{n}\left[\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}-\frac{\phi\left(0\right)^2}{\Phi\left(0\right)^2}\right]\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq\ & \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left[\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2-\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]\right|+\left|\bar{\delta}\right|\left(2\sqrt{\frac{2}{\pi}}+\frac{2\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right) \\
& \times\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|+\bar{\delta}^2\left(\sqrt{\frac{2}{\pi}}+\frac{\sqrt{2/\pi}}{1+\bar{\alpha}^2}\right)^2\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\right| \\
& +\frac{1}{n}\left|\sum_{i=1}^{n}\left[\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}-\frac{\phi\left(0\right)^2}{\Phi\left(0\right)^2}\right]\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq\ & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O\left(1\right) \\
& \cdot\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(1\right)\right]+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right) \\
& +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right) \\
=\ & \kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right) \\
& +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\bar{B}_i{}^2\right| &= |\bar{\alpha}|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\right)\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}+\bar{\delta}\sqrt{\frac{2}{\pi}}\cdot\frac{2+\bar{\alpha}^2}{1+\bar{\alpha}^2}\right)^2\right| \\
&\leq |\bar{\alpha}|\left[\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^3\right|\cdot O(1)+\left|\bar{\delta}\right|\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)^2\right|\cdot O(1)\right. \\
&\quad \left.+\bar{\delta}^2\cdot\frac{1}{n}\left|\sum_{i=1}^{n}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}}\right)\right|\cdot O(1)+\left|\bar{\delta}\right|^3\cdot O(1)\right] \\
&\leq \kappa_{\alpha n}\left[O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)\right. \\
&\quad \left.+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O(1)+\kappa_{\alpha n}^3\cdot O(1)\right] \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1) \\
\frac{1}{n}\left|\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\bar{B}_i{}^2\right| &\leq \frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i\bar{B}_i{}^2\right|\cdot O(1)+\frac{1}{n}\left|\sum_{i=1}^{n}\bar{A}_i{}^2\bar{B}_i{}^2\right|\cdot O(1) \\
&\leq \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1) \\
&= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)
\end{aligned}
$$

$$
\begin{aligned}
&\left|\frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| \\
\leq\ &\left|\frac{3\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\bar{A}_i\right|+\left|\frac{2}{\pi\left(1+\bar{\alpha}^2\right)^3}-\frac{2}{\pi}\right|+\left|\frac{\sqrt{2/\pi}\left(2-\bar{\alpha}^2\right)}{\left(1+\bar{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{4}{\pi}\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\bar{B}_i{}^2\right|+\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i{}^2-\frac{2}{\pi}\cdot\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right| \\
\leq\ &\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right) \\
&+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1) \\
&+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right) \\
&+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1) \\
=\ &\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right) \\
&+\kappa_{\mu n}^2\cdot O_p(1).
\end{aligned}
$$

Finally, we have the results

$$\begin{aligned}
\left|\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\sigma n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)\\
\left|\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)\\
\left|\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1)\\
\left|\frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)\\
\left|\frac{e}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{e}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1)\\
\left|\frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0}\right| &= \kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)\\
&\quad +\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p(1).
\end{aligned}$$

First we denote

$$\begin{aligned}
U_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right) &\xrightarrow{d} N(0,1),\\
V_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-1\right] &\xrightarrow{d} N(0,2),\\
\begin{pmatrix}U_n\\V_n\end{pmatrix}=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix}\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\\\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-1\end{pmatrix} &\xrightarrow{d} N\left(\mathbf{0},\begin{pmatrix}1&0\\0&2\end{pmatrix}\right),\\
U_n=O_p(1), \qquad V_n=O_p(1).&
\end{aligned}$$

By calculation

$$\begin{aligned}
\frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= -\frac{1}{\tilde{\sigma}^2} \\
\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= 0 \\
\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= -\frac{1}{n}\cdot\frac{2}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right) \\
&= -\frac{2U_n}{\tilde{\sigma}^2\sqrt{n}} \\
&= O_p\left(n^{-1/2}\right) \\
\frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= \frac{1}{\tilde{\sigma}^2}-\frac{1}{n}\cdot\frac{3}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 \\
&= -\frac{3V_n}{\tilde{\sigma}^2\sqrt{n}}-\frac{2}{\tilde{\sigma}^2} \\
&= -\frac{2}{\tilde{\sigma}^2}+O_p\left(n^{-1/2}\right) \\
\frac{e}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= 0.
\end{aligned}$$

Because as $n\to\infty$, $\frac{\lambda}{n}\to c_0>0$, we can derive

$$\begin{aligned}
\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} &= \frac{2}{\pi}-\frac{1}{n}\cdot\frac{2}{\pi}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-\frac{2\lambda}{n} \\
&= -\frac{2V_n}{\pi\sqrt{n}}-\frac{2\lambda}{n} \\
&= -2c_0+O_p\left(n^{-1/2}\right).
\end{aligned}$$

Then we can calculate

$$
\begin{aligned}
\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{1}{\tilde{\sigma}^2}+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)\\
\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{2U_n}{\tilde{\sigma}^2\sqrt{n}}+\kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&= O_p\left(n^{-1/2}\right)+\kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&= O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&= O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right)\\
\frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{3V_n}{\tilde{\sigma}^2\sqrt{n}}-\frac{2}{\tilde{\sigma}^2}+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
&= -\frac{2}{\tilde{\sigma}^2}+O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
&= -\frac{2}{\tilde{\sigma}^2}+O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
&= -\frac{2}{\tilde{\sigma}^2}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)\\
\frac{e}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right)\\
\frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{2V_n}{\pi\sqrt{n}}-\frac{2\lambda}{n}+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
&\quad +\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p\left(1\right)\\
&= -\frac{2\lambda}{n}+O_p\left(n^{-1/2}\right)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}^2\cdot O_p\left(1\right)\\
&= -\frac{2\lambda}{n}+O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\mu n}^2\cdot O_p\left(1\right)\\
&= -2c_0+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\cdot O_p\left(1\right).
\end{aligned}
$$

And as we know that

$$
nC_{11}^{-1} = \frac{1}{ad/n^2-c^2/n^2}\begin{pmatrix} d/n & -c/n \\ -c/n & a/n \end{pmatrix}.
$$

Therefore, after calculation and rearrangement,

$$\begin{aligned}
nC_{11}^{-} &= nC_{11}^{-1} + nC_{11}^{-1}C_{12}\left(C_{22} - C_{21}C_{11}^{-1}C_{12}\right)^{-1}C_{21}C_{11}^{-1} \\
&= \frac{\frac{1}{n^2}\begin{pmatrix} df - e^2 & be - cf \\ be - cf & af - b^2 \end{pmatrix}_{2\times 2}}{\frac{1}{n^3}\left(adf - c^2 f - b^2 d + 2bce - ae^2\right)} \\
C_{12}^{-} &= -C_{11}^{-1}C_{12}\left(C_{22} - C_{21}C_{11}^{-1}C_{12}\right)^{-1} \\
&= \frac{\begin{pmatrix} ce - bd \\ bc - ae \end{pmatrix}_{2\times 1}}{adf - c^2 f - b^2 d + 2bce - ae^2} \\
C_{21}^{-} &= -\left(C_{22} - C_{21}C_{11}^{-1}C_{12}\right)^{-1}C_{21}C_{11}^{-1} \\
&= \frac{\begin{pmatrix} ce - bd & bc - ae \end{pmatrix}_{1\times 2}}{adf - c^2 f - b^2 d + 2bce - ae^2} \\
C_{22}^{-} &= \left(C_{22} - C_{21}C_{11}^{-1}C_{12}\right)^{-1} \\
&= \frac{ad - c^2}{adf - c^2 f - b^2 d + 2bce - ae^2}.
\end{aligned}$$

Moreover,

$$\begin{pmatrix} \frac{\partial l_{inc}^{\star}}{\partial \mu} \\ \frac{\partial l_{inc}^{\star}}{\partial \sigma} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} = \begin{pmatrix} \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right) \\ -\frac{n}{\tilde{\sigma}} + \frac{1}{\tilde{\sigma}^3}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right)^2 \end{pmatrix}.$$

Finally we can reach

$$\frac{\frac{1}{n^2}\begin{pmatrix} e^2 - df & cf - be \\ cf - be & b^2 - af \end{pmatrix}_{2\times 2}}{\frac{1}{n^3}\left(adf - c^2 f - b^2 d + 2bce - ae^2\right)} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \frac{1}{\sqrt{n}}\begin{pmatrix} \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right) \\ -\frac{n}{\tilde{\sigma}} + \frac{1}{\tilde{\sigma}^3}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right)^2 \end{pmatrix} = \sqrt{n}\begin{pmatrix} \hat{\mu} - \tilde{\mu} \\ \hat{\sigma} - \tilde{\sigma} \end{pmatrix} \tag{6.2.6}$$

simultaneously with

$$\frac{\frac{1}{n^2}\begin{pmatrix} bd - ce & ae - bc \end{pmatrix}_{1\times 2}}{\frac{1}{n^3}\left(adf - c^2 f - b^2 d + 2bce - ae^2\right)} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \frac{1}{\sqrt{n}}\begin{pmatrix} \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right) \\ -\frac{n}{\tilde{\sigma}} + \frac{1}{\tilde{\sigma}^3}\sum_{i=1}^{n}\left(y_i - \tilde{\mu}\right)^2 \end{pmatrix} = \sqrt{n}\hat{\alpha}. \tag{6.2.7}$$

As we can calculate now

$$\begin{aligned}
\left(\frac{e}{n}\right)^2 |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}^2 O_p\left(n^{-1}\right) + \kappa_{\mu n}\kappa_{\sigma n} O_p\left(n^{-1/2}\right) + \kappa_{\sigma n}\kappa_{\alpha n} O_p\left(n^{-1/2}\right) \\
&\quad + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right)^2 \cdot O_p\left(1\right) \\
\frac{d}{n} \cdot \frac{f}{n} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{4c_0}{\tilde{\sigma}^2} + O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right)^2 \cdot O_p\left(1\right) + \kappa_{\sigma n} \cdot O_p\left(1\right) \\
\frac{c}{n} \cdot \frac{f}{n} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right) \cdot O_p\left(1\right) \\
\frac{b}{n} \cdot \frac{e}{n} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\mu n}\kappa_{\alpha n}^2 \cdot O_p\left(1\right) + \kappa_{\alpha n}^3 \cdot O_p\left(1\right) + \kappa_{\sigma n}^2 \cdot O_p\left(n^{-1/2}\right) \\
&\quad + \kappa_{\mu n}\kappa_{\sigma n} \cdot O_p\left(1\right) + \kappa_{\sigma n}\kappa_{\alpha n} \cdot O_p\left(1\right) + \kappa_{\mu n}\kappa_{\alpha n} \cdot O_p\left(n^{-1/2}\right) \\
&\quad + \kappa_{\alpha n}^2 \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}^2\kappa_{\alpha n} \cdot O_p\left(1\right) \\
\left(\frac{b}{n}\right)^2 |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^4 \cdot O_p\left(1\right) + \kappa_{\sigma n}\kappa_{\alpha n}^2 \cdot O_p\left(1\right) + \kappa_{\alpha n}^3 \cdot O_p\left(n^{-1/2}\right) \\
&\quad + \kappa_{\mu n}\kappa_{\alpha n}^3 \cdot O_p\left(1\right) + \kappa_{\sigma n}^2 \cdot O_p\left(1\right) + \kappa_{\sigma n}\kappa_{\alpha n} \cdot O_p\left(n^{-1/2}\right) \\
&\quad + \kappa_{\mu n}\kappa_{\sigma n}\kappa_{\alpha n} \cdot O_p\left(1\right) + \kappa_{\alpha n}^2 O_p\left(n^{-1}\right) \\
&\quad + \kappa_{\mu n}\kappa_{\alpha n}^2 \cdot O_p\left(n^{-1/2}\right) + \kappa_{\mu n}^2\kappa_{\alpha n}^2 \cdot O_p\left(1\right) \\
\frac{a}{n} \cdot \frac{f}{n} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{2c_0}{\tilde{\sigma}^2} + O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right)^2 \cdot O_p\left(1\right) + \kappa_{\sigma n} \cdot O_p\left(1\right).
\end{aligned}$$

Then

$$\begin{aligned}
\left[\left(\frac{e}{n}\right)^2 - \frac{d}{n} \cdot \frac{f}{n}\right] |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{4c_0}{\tilde{\sigma}^2} + O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right)^2 O_p\left(1\right) + \kappa_{\sigma n} \cdot O_p\left(1\right) \\
\left[\frac{c}{n} \cdot \frac{f}{n} - \frac{b}{n} \cdot \frac{e}{n}\right] |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right) \cdot O_p\left(1\right) \\
\left[\left(\frac{b}{n}\right)^2 - \frac{a}{n} \cdot \frac{f}{n}\right] |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{2c_0}{\tilde{\sigma}^2} + O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\alpha n}\right)^2 O_p\left(1\right) + \kappa_{\sigma n} O_p\left(1\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\left(\frac{a}{n}\cdot\frac{d}{n}\cdot\frac{f}{n}\right)|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{4c_0}{\tilde{\sigma}^4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p\left(1\right)+\kappa_{\sigma n}O_p\left(1\right)\\
\left[\left(\frac{c}{n}\right)^2\cdot\frac{f}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= O_p\left(n^{-1}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p\left(1\right)\\
\left[\left(\frac{b}{n}\right)^2\cdot\frac{d}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^4\cdot O_p\left(1\right)+\kappa_{\sigma n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)+\kappa_{\sigma n}^2\cdot O_p\left(1\right)+\kappa_{\sigma n}\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\sigma n}\kappa_{\alpha n}\cdot O_p\left(1\right)+\kappa_{\alpha n}^2 O_p\left(n^{-1}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}^2\kappa_{\alpha n}^2\cdot O_p\left(1\right)\\
\left(\frac{b}{n}\cdot\frac{c}{n}\cdot\frac{e}{n}\right)|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\mu n}\kappa_{\alpha n}^2\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\alpha n}^3\cdot O_p\left(1\right)\\
&\quad +\kappa_{\alpha n}^3\cdot O_p\left(n^{-1/2}\right)+\kappa_{\alpha n}^4\cdot O_p\left(1\right)+\kappa_{\sigma n}^2 O_p\left(n^{-1}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\kappa_{\sigma n}\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\sigma n}\kappa_{\alpha n}\cdot O_p\left(1\right)\\
&\quad +\kappa_{\sigma n}\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\sigma n}\kappa_{\alpha n}^2\cdot O_p\left(1\right)+\kappa_{\mu n}\kappa_{\alpha n}O_p\left(n^{-1}\right)\\
&\quad +\kappa_{\mu n}^2\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\alpha n}^2 O_p\left(n^{-1}\right)+\kappa_{\mu n}^3\kappa_{\alpha n}\cdot O_p\left(1\right)\\
\left[\frac{a}{n}\cdot\left(\frac{e}{n}\right)^2\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}^2 O_p\left(n^{-1}\right)+\kappa_{\mu n}\kappa_{\sigma n}O_p\left(n^{-1/2}\right)+\kappa_{\sigma n}\kappa_{\alpha n}O_p\left(n^{-1/2}\right)\\
&\quad +\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2\cdot O_p\left(1\right).
\end{aligned}$$

After calculation and rearrangement

$$\begin{aligned}
&\left[\frac{a}{n}\cdot\frac{d}{n}\cdot\frac{f}{n}-\left(\frac{c}{n}\right)^2\cdot\frac{f}{n}-\left(\frac{b}{n}\right)^2\cdot\frac{d}{n}+2\cdot\frac{b}{n}\cdot\frac{c}{n}\cdot\frac{e}{n}-\frac{a}{n}\cdot\left(\frac{e}{n}\right)^2\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}\\
= \; & -\frac{4c_0}{\tilde{\sigma}^4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p\left(1\right)+\kappa_{\sigma n}O_p\left(1\right).
\end{aligned}$$

As we have

$$\frac{1}{\sqrt{n}}\begin{pmatrix}\frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\\ -\frac{n}{\tilde{\sigma}}+\frac{1}{\tilde{\sigma}^3}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)^2\end{pmatrix} = \frac{1}{\tilde{\sigma}}\begin{pmatrix}U_n\\ V_n\end{pmatrix},$$

therefore, from (6.2.6) we can reach

$$\frac{\begin{pmatrix} -\frac{4c_0}{\tilde{\sigma}^2}+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p(1) & O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)\cdot O_p(1) \\ +\kappa_{\sigma n}\cdot O_p(1)+O_p\left(n^{-1/2}\right) & \\ O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)\cdot O_p(1) & -\frac{2c_0}{\tilde{\sigma}^2}+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p(1) \\ & +\kappa_{\sigma n}O_p(1)+O_p\left(n^{-1/2}\right) \end{pmatrix}}{-\frac{4c_0}{\tilde{\sigma}^4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p(1)+\kappa_{\sigma n}O_p(1)} \begin{pmatrix} \frac{U_n}{\tilde{\sigma}} \\ \\ \frac{V_n}{\tilde{\sigma}} \end{pmatrix}$$

$$= \sqrt{n}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \\ \hat{\sigma}-\tilde{\sigma} \end{pmatrix}.$$

As $\kappa_{\mu n}=o(1)$, $\kappa_{\sigma n}=o(1)$, $\kappa_{\alpha n}=o(1)$ and $U_n=O_p(1)$, $V_n=O_p(1)$, so

$$\begin{aligned} \sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right) &= \frac{-4c_0\cdot U_n+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p(1)}{-4c_0+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p(1)+\kappa_{\sigma n}O_p(1)} \\ &= U_n+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p(1) \\ &\xrightarrow{d} N(0,1) \\ \sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right) &= \frac{-2c_0\cdot V_n+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p(1)}{-4c_0+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2 O_p(1)+\kappa_{\sigma n}O_p(1)} \\ &= \frac{V_n}{2}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p(1) \\ &\xrightarrow{d} N\left(0,\frac{1}{2}\right). \end{aligned}$$

And we can solve (6.2.7) by calculating

$$\begin{aligned}
\left[\frac{b}{n}\cdot\frac{d}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)\\
\left[\frac{c}{n}\cdot\frac{e}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}O_p\left(n^{-1}\right)+\kappa_{\mu n}O_p\left(n^{-1/2}\right)+\kappa_{\alpha n}O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}^2\cdot O_p(1)+\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\alpha n}^2\cdot O_p(1)\\
\left[\frac{a}{n}\cdot\frac{e}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}\cdot O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\cdot O_p(1)+\kappa_{\alpha n}\cdot O_p(1)\\
\left[\frac{b}{n}\cdot\frac{c}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^2O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\alpha n}^2O_p(1)+\kappa_{\alpha n}^3O_p(1)\\
&\quad +\kappa_{\sigma n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}\kappa_{\sigma n}O_p(1)+\kappa_{\sigma n}\kappa_{\alpha n}O_p(1)\\
&\quad +\kappa_{\alpha n}O_p\left(n^{-1}\right)+\kappa_{\mu n}\kappa_{\alpha n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\kappa_{\alpha n}O_p(1)\\
\left[\frac{b}{n}\cdot\frac{d}{n}-\frac{c}{n}\cdot\frac{e}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\alpha n}^2\cdot O_p(1)+\kappa_{\sigma n}\cdot O_p(1)+\kappa_{\alpha n}O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\mu n}\kappa_{\alpha n}\cdot O_p(1)+\kappa_{\mu n}O_p\left(n^{-1/2}\right)+\kappa_{\mu n}^2\cdot O_p(1)\\
&= \left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2O_p(1)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)O_p\left(n^{-1/2}\right)\\
&\quad +\kappa_{\sigma n}O_p(1)\\
\left[\frac{a}{n}\cdot\frac{e}{n}-\frac{b}{n}\cdot\frac{c}{n}\right]|_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \kappa_{\sigma n}O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)O_p(1).
\end{aligned}$$

Thus,

$$\frac{\begin{pmatrix}\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2O_p(1)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)O_p\left(n^{-1/2}\right)+\kappa_{\sigma n}O_p(1)\\ \\ \kappa_{\sigma n}O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)O_p(1)\end{pmatrix}^T}{-\frac{4c_0}{\tilde{\sigma}^4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2O_p(1)+\kappa_{\sigma n}O_p(1)}\begin{pmatrix}\frac{U_n}{\tilde{\sigma}}\\ \\ \frac{V_n}{\tilde{\sigma}}\end{pmatrix} = \sqrt{n}\hat{\alpha}.$$

Therefore, we can get

$$\begin{aligned}
\sqrt{n}\begin{pmatrix}\hat{\alpha}\\ \tilde{\sigma}\end{pmatrix} &= \frac{\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)O_p(1)}{-4c_0+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\alpha n}\right)^2O_p(1)+\kappa_{\sigma n}O_p(1)}\\
&= \left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)O_p(1).
\end{aligned}$$

□

## Proof of Theorem 2

*Proof.* First with the results from the previous section we can calculate:

$$
\begin{aligned}
n\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right)^2 \cdot \frac{a}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{U_n^2}{\tilde{\sigma}^2} + \kappa_{\sigma n} \cdot O_p(1) + \kappa_{\alpha n}^2 \cdot O_p(1) \\
\sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right) \cdot \sqrt{n}\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right) \cdot \frac{b}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \left[\kappa_{\alpha n}^2 \cdot O_p(1) + \kappa_{\sigma n} \cdot O_p(1) + \kappa_{\alpha n} O_p\left(n^{-1/2}\right)\right. \\
&\quad \left. + \kappa_{\mu n}\kappa_{\alpha n} \cdot O_p(1)\right] \cdot (\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n}) \\
\sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right) \cdot \sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right) \cdot \frac{c}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= O_p\left(n^{-1/2}\right) + (\kappa_{\mu n} + \kappa_{\alpha n}) \cdot O_p(1) \\
n\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right)^2 \frac{d}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{V_n^2}{2\tilde{\sigma}^2} + O_p\left(n^{-1/2}\right) + (\kappa_{\mu n} + \kappa_{\alpha n})^2 \cdot O_p(1) \\
&\quad + \kappa_{\sigma n} \cdot O_p(1) \\
\sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right) \cdot \sqrt{n}\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right) \cdot \frac{e}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= (\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n}) \cdot \left[\kappa_{\sigma n} \cdot O_p\left(n^{-1/2}\right)\right. \\
&\quad \left. + (\kappa_{\mu n} + \kappa_{\alpha n}) \cdot O_p(1)\right] \\
n\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right)^2 \frac{f}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= (\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n})^2 \cdot O_p(1).
\end{aligned}
$$

Therefore, use multivariate Taylor series expansion of Lagrange form at point $\alpha = 0$, we can get

$$
\begin{aligned}
& l^{\star}_{inc}\left(\hat{\mu},\hat{\sigma},\hat{\alpha}\mid \mathbf{y}\right) \\
= \; & l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right)+\begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=0} \end{pmatrix}^{T}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix} \\
& +\frac{1}{2}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}^{T}\begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha^2} \end{pmatrix}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix} \\
= \; & l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right)+\frac{\sqrt{n}}{\tilde{\sigma}}\begin{pmatrix} U_n \\ V_n \\ 0 \end{pmatrix}^{T}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}+\frac{1}{2}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix}^{T}\begin{pmatrix} a & c & b \\ c & d & e \\ b & e & f \end{pmatrix}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}\begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha} \end{pmatrix} \\
= \; & l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right)+U_n\cdot\sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right)+V_n\cdot\sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right)+\frac{\tilde{\sigma}^2}{2}\left[n\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right)^2\cdot\frac{a}{n}\right. \\
& +2\sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right)\cdot\sqrt{n}\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right)\cdot\frac{b}{n}+2\sqrt{n}\left(\frac{\hat{\mu}-\tilde{\mu}}{\tilde{\sigma}}\right)\cdot\sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right)\cdot\frac{c}{n}+n\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right)^2\frac{d}{n} \\
& \left.+2\sqrt{n}\left(\frac{\hat{\sigma}-\tilde{\sigma}}{\tilde{\sigma}}\right)\cdot\sqrt{n}\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right)\cdot\frac{e}{n}+n\left(\frac{\hat{\alpha}}{\tilde{\sigma}}\right)^2\frac{f}{n}\right]\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
= \; & l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right)+U_n^2+\frac{V_n^2}{2}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right) \\
& +\frac{\tilde{\sigma}^2}{2}\left[-\frac{U_n^2}{\tilde{\sigma}^2}-\frac{V_n^2}{2\tilde{\sigma}^2}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right)\right] \\
= \; & l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right)+\frac{U_n^2}{2}+\frac{V_n^2}{4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right). \qquad (6.2.8)
\end{aligned}
$$

The definitions of $U_n$ and $V_n$ can refer to **Proof of Theorem 1**. Then as we have

$$
\begin{aligned}
l^{\star}_{inc}\left(\tilde{\mu},\tilde{\sigma},0\mid \mathbf{y}\right) &= -\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\tilde{\sigma}^2\right)-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 \\
&= -\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\tilde{\sigma}^2\right)-\frac{1}{2}\left(\sqrt{n}V_n+n\right).
\end{aligned}
$$

Therefore, (6.2.8) can be written as

$$
\begin{aligned}
l^{\star}_{inc}\left(\hat{\mu},\hat{\sigma},\hat{\alpha}\mid \mathbf{y}\right) \; = \; & -\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\tilde{\sigma}^2\right)-\frac{1}{2}\left(\sqrt{n}V_n+n\right) \\
& +\frac{U_n^2}{2}+\frac{V_n^2}{4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right).
\end{aligned}
$$

As $(\hat{\mu}^*, \hat{\sigma}^*, 0)$ is the maximum likelihood estimate when fixing $\alpha = 0$, so $\hat{\mu}^* = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{\mathbf{y}}$ and $\hat{\sigma}^{*2} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{\mathbf{y}})^2$ are the maximum likelihood estimates for the ordinary normal distribution. And

$$\begin{aligned}
\hat{\sigma}^{*2} &= \frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{\mathbf{y}})^2 \\
&= \frac{\tilde{\sigma}^2}{n}\sum_{i=1}^{n} \left(\frac{y_i - \tilde{\mu} + \tilde{\mu} - \bar{\mathbf{y}}}{\tilde{\sigma}}\right)^2 \\
&= \frac{\tilde{\sigma}^2}{n}\sum_{i=1}^{n} \left[\left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}}\right)^2 + \left(\frac{\tilde{\mu} - \bar{\mathbf{y}}}{\tilde{\sigma}}\right)^2 + 2\left(\frac{y_i - \tilde{\mu}}{\tilde{\sigma}}\right)\left(\frac{\tilde{\mu} - \bar{\mathbf{y}}}{\tilde{\sigma}}\right)\right] \\
&= \frac{\tilde{\sigma}^2}{n}\left[\left(\sqrt{n}V_n + n\right) + \frac{n(\tilde{\mu} - \bar{\mathbf{y}})^2}{\tilde{\sigma}^2} + 2\left(\frac{\tilde{\mu} - \bar{\mathbf{y}}}{\tilde{\sigma}}\right)\left(\frac{n\bar{\mathbf{y}} - n\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&= \frac{\tilde{\sigma}^2}{n}\left[\sqrt{n}V_n + n - \frac{n(\tilde{\mu} - \bar{\mathbf{y}})^2}{\tilde{\sigma}^2}\right] \\
&= \frac{\tilde{\sigma}^2}{n}\left(\sqrt{n}V_n + n - U_n^2\right) \\
\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}^*}{\hat{\sigma}^*}\right)^2 &= \frac{\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\mathbf{y}})^2} = n \\
l^{\star}_{inc}(\hat{\mu}^*, \hat{\sigma}^*, 0 \mid \mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\hat{\sigma}^{*2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}^*}{\hat{\sigma}^*}\right)^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left[\frac{\tilde{\sigma}^2}{n}\left(\sqrt{n}V_n + n - U_n^2\right)\right] - \frac{n}{2} \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\tilde{\sigma}^2\right) + \frac{n}{2}\log n - \frac{n}{2}\log\left(\sqrt{n}V_n + n - U_n^2\right) - \frac{n}{2}.
\end{aligned}$$

Finally, we can calculate the desired likelihood ratio

$$\begin{aligned}
l^{\star}_{inc}(\hat{\mu}, \hat{\sigma}, \hat{\alpha} \mid \mathbf{y}) - l^{\star}_{inc}(\hat{\mu}^*, \hat{\sigma}^*, 0 \mid \mathbf{y}) &= -\frac{\sqrt{n}V_n}{2} + \frac{n}{2}\log\left(\frac{V_n}{\sqrt{n}} + 1 - \frac{U_n^2}{n}\right) + \frac{U_n^2}{2} + \frac{V_n^2}{4} \\
&\quad + O_p\left(n^{-1/2}\right) + \left(\kappa_{\mu n} + \kappa_{\sigma n} + \kappa_{\alpha n}\right) \cdot O_p(1).
\end{aligned}$$

As Taylor series expansion can be used to calculate $\log\left(\frac{V_n}{\sqrt{n}} + 1 - \frac{U_n^2}{n}\right)$, so

$$\log\left(\frac{V_n}{\sqrt{n}} + 1 - \frac{U_n^2}{n}\right) = 0 + \frac{V_n}{\sqrt{n}} - \frac{U_n^2}{n} - \frac{V_n^2}{2n} + O\left(n^{-3/2}\right).$$

Then

$$-\frac{\sqrt{n}V_n}{2} + \frac{n}{2}\log\left(\frac{V_n}{\sqrt{n}} + 1 - \frac{U_n^2}{n}\right) = -\frac{\sqrt{n}V_n}{2} + \frac{\sqrt{n}V_n}{2} - \frac{U_n^2}{2} - \frac{V_n^2}{4} + O\left(n^{-1/2}\right).$$

Thus,

$$\begin{aligned} l^{\star}_{inc}\left(\hat{\mu},\hat{\sigma},\hat{\alpha}\mid\mathbf{y}\right)-l^{\star}_{inc}\left(\hat{\mu}^{*},\hat{\sigma}^{*},0\mid\mathbf{y}\right) &= -\frac{\sqrt{n}V_n}{2}+\frac{\sqrt{n}V_n}{2}-\frac{U_n^2}{2}-\frac{V_n^2}{4}+O\left(n^{-1/2}\right)+\frac{U_n^2}{2} \\ &\quad +\frac{V_n^2}{4}+O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right) \\ &= O_p\left(n^{-1/2}\right)+\left(\kappa_{\mu n}+\kappa_{\sigma n}+\kappa_{\alpha n}\right)\cdot O_p\left(1\right). \end{aligned}$$

□

## Proof of Theorem 3

*Proof.* From mean value theorem $(\bar{\mu}, \bar{\sigma}, \bar{\alpha})$ is the parameter values between the restricted MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ and underlying real parameter values $(\tilde{\mu}, \tilde{\sigma}, \tilde{\alpha})$. Then the multivariate Taylor series expansion of Lagrange form at the restricted MLE $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ can be written as

$$\begin{aligned}
\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}_{3\times 1} &= \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix}_{3\times 1} |_{\mu=\hat{\mu},\sigma=\hat{\sigma},\alpha=\hat{\alpha}} \\
&= \begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix}_{3\times 1} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} \\
&\quad + \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \sigma} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha^2} \end{pmatrix}_{3\times 3} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha}-\tilde{\alpha} \end{pmatrix}_{3\times 1}
\end{aligned}$$

where

$$\begin{pmatrix} \bar{\mu} \\ \bar{\sigma} \\ \bar{\alpha} \end{pmatrix} = \begin{pmatrix} \tilde{\mu} \\ \tilde{\sigma} \\ \tilde{\alpha} \end{pmatrix} + t \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha}-\tilde{\alpha} \end{pmatrix}, \qquad t \in (0,1)\,.$$

To make it simple, we use block matrix to simplify it.

$$-\begin{pmatrix} \frac{\partial l^{\star}_{inc}}{\partial \mu} \\ \frac{\partial l^{\star}_{inc}}{\partial \sigma} \\ \frac{\partial l^{\star}_{inc}}{\partial \alpha} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \hat{\mu}-\tilde{\mu} \\ \hat{\sigma}-\tilde{\sigma} \\ \hat{\alpha}-\tilde{\alpha} \end{pmatrix}$$

where

$$C_{11} = \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu^2} & \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma^2} \end{pmatrix}, \qquad C_{12} = \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l^{\star}_{inc}}{\partial \sigma \partial \alpha} \end{pmatrix}, \qquad C_{21} = \begin{pmatrix} \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \mu} & \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha \partial \sigma} \end{pmatrix}, \qquad C_{22} = \frac{\partial^2 l^{\star}_{inc}}{\partial \alpha^2}.$$

Then we denote it as

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}\,.$$

The inverse of it is

$$C^{-1} = \begin{pmatrix} C_{11}^{-} & C_{12}^{-} \\ C_{21}^{-} & C_{22}^{-} \end{pmatrix} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}$$

where

$$\begin{aligned} C_{11}^{-} &= C_{11}^{-1} + C_{11}^{-1} C_{12} \left( C_{22} - C_{21} C_{11}^{-1} C_{12} \right)^{-1} C_{21} C_{11}^{-1} \\ C_{12}^{-} &= -C_{11}^{-1} C_{12} \left( C_{22} - C_{21} C_{11}^{-1} C_{12} \right)^{-1} \\ C_{21}^{-} &= -\left( C_{22} - C_{21} C_{11}^{-1} C_{12} \right)^{-1} C_{21} C_{11}^{-1} \\ C_{22}^{-} &= \left( C_{22} - C_{21} C_{11}^{-1} C_{12} \right)^{-1}. \end{aligned}$$

Therefore, we can solve the equation now by

$$-C^{-1} \begin{pmatrix} \frac{\partial l_{inc}^{\star}}{\partial \mu} \\ \frac{\partial l_{inc}^{\star}}{\partial \sigma} \\ \frac{\partial l_{inc}^{\star}}{\partial \alpha} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} = \begin{pmatrix} \hat{\mu} - \tilde{\mu} \\ \hat{\sigma} - \tilde{\sigma} \\ \hat{\alpha} - \tilde{\alpha} \end{pmatrix}.$$

That is

$$\begin{cases} -C_{11}^{-} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l_{inc}^{\star}}{\partial \mu} \\ \frac{\partial l_{inc}^{\star}}{\partial \sigma} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} - C_{12}^{-} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \cdot \frac{\partial l_{inc}^{\star}}{\partial \alpha} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} = \begin{pmatrix} \hat{\mu} - \tilde{\mu} \\ \hat{\sigma} - \tilde{\sigma} \end{pmatrix} \\ -C_{21}^{-} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \begin{pmatrix} \frac{\partial l_{inc}^{\star}}{\partial \mu} \\ \frac{\partial l_{inc}^{\star}}{\partial \sigma} \end{pmatrix} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} - C_{22}^{-} |_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \cdot \frac{\partial l_{inc}^{\star}}{\partial \alpha} |_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} = \hat{\alpha} - \tilde{\alpha}. \end{cases} \tag{6.2.9}$$

As $y_i \sim SN\left(\tilde{\mu}-\tilde{\sigma}\tilde{\delta}\sqrt{\frac{2}{\pi}},\tilde{\sigma}^2,\tilde{\alpha}\right)$, so

$$\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\sim SN\left(0,1,\tilde{\alpha}\right), \qquad \left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)^2\sim x^2\left(1\right),$$

$$E\left[\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right]=0, \qquad \mathrm{Var}\left[\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right]=1-\frac{2\tilde{\delta}^2}{\pi},$$

$$E\left[\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right]=\tilde{\delta}\sqrt{\frac{2}{\pi}}, \qquad E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)^2\right]=1,$$

$$\mathrm{Var}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)^2\right]=2, \qquad E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)^3\right]=\frac{\tilde{\delta}}{\sqrt{2\pi}}\left(4+\frac{2}{1+\tilde{\alpha}^2}\right),$$

$$E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)^4\right]=3, \qquad E\left[\tilde{A}_i\right]=E\left[\tilde{\alpha}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}+\tilde{\delta}\sqrt{\frac{2}{\pi}}\right)\right]=\tilde{\alpha}\tilde{\delta}\sqrt{\frac{2}{\pi}},$$

$$E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]=1-\frac{2\tilde{\delta}^2}{\pi}, \qquad E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^4\right]=3-\frac{4\tilde{\delta}^2}{\pi}-\frac{8\tilde{\delta}^4}{\pi\tilde{\alpha}^2}-\frac{12\tilde{\delta}^4}{\pi^2},$$

$$E\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right]=\left(\frac{4}{\pi}-1\right)\tilde{\delta}^3\sqrt{\frac{2}{\pi}}, \qquad \mathrm{Var}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]=2-\frac{8\tilde{\delta}^4}{\pi\tilde{\alpha}^2}-\frac{16\tilde{\delta}^4}{\pi^2}.$$

For the convenience, we denote that

$$E_{10}=E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right]=\frac{\sqrt{2/\pi}}{\sqrt{1+\tilde{\alpha}^2}}, \qquad E_{11}=E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]=-\frac{2\tilde{\delta}^2}{\pi\tilde{\alpha}},$$

$$E_{12}=E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]=\sqrt{\frac{2}{\pi}}\frac{\tilde{\delta}^3}{\tilde{\alpha}}\left(\frac{1}{\tilde{\alpha}^2}+\frac{2}{\pi}\right),$$

$$E_{13}=E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right]=-\frac{2\tilde{\delta}^4}{\pi\tilde{\alpha}}\left(\frac{3}{\tilde{\alpha}^2}+\frac{2}{\pi}\right),$$

$$E_{20}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right], \qquad E_{21}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right], \qquad E_{22}=E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right].$$

With central limit theorem, we can derive

$$\begin{aligned}
\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)-0\right] &\xrightarrow{d} N\left(0, \quad 1-\frac{2\tilde{\delta}^2}{\pi}\right)\\
\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-\left(1-\frac{2\tilde{\delta}^2}{\pi}\right)\right] &\xrightarrow{d} N\left(0, \quad 2-\frac{8\tilde{\delta}^4}{\pi\tilde{\alpha}^2}-\frac{16\tilde{\delta}^4}{\pi^2}\right)\\
\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right]-E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right]\right\} &\xrightarrow{d} N\left(0, \quad E_{20}-\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)}\right)\\
\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]-E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]\right\} &\xrightarrow{d} N\left(0, \quad E_{22}-\frac{4\tilde{\delta}^4}{\pi^2\tilde{\alpha}^2}\right).
\end{aligned}$$

Denote them as $U_n$ , $V_n$, $Q_n$ and $R_n$:

$$\begin{aligned}
U_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right) &\xrightarrow{d} N\left(0, \quad 1-\frac{2\tilde{\delta}^2}{\pi}\right)\\
V_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-\left(1-\frac{2\tilde{\delta}^2}{\pi}\right)\right] &\xrightarrow{d} N\left(0, \quad 2-\frac{8\tilde{\delta}^4}{\pi\tilde{\alpha}^2}-\frac{16\tilde{\delta}^4}{\pi^2}\right)\\
Q_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}-\frac{\sqrt{2/\pi}}{\sqrt{1+\tilde{\alpha}^2}}\right] &\xrightarrow{d} N\left(0, \quad E_{20}-\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)}\right)\\
Q_n=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}-\frac{\sqrt{2/\pi}}{\sqrt{1+\tilde{\alpha}^2}}\right] &\xrightarrow{d} N\left(0, \quad E_{20}-\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)}\right)
\end{aligned}$$

then

$$\begin{aligned}
\frac{\partial l_{inc}^{\star}}{\partial \mu} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{1}{\tilde{\sigma}} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} + \tilde{\delta}\sqrt{\frac{2}{\pi}} \right) - \frac{\tilde{\alpha}}{\tilde{\sigma}} \sum_{i=1}^{n} \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \\
\frac{1}{\sqrt{n}} \frac{\partial l_{inc}^{\star}}{\partial \mu} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{U_n}{\tilde{\sigma}} + \frac{\sqrt{n}\tilde{\delta}}{\tilde{\sigma}}\sqrt{\frac{2}{\pi}} - \frac{\tilde{\alpha}}{\tilde{\sigma}} \left\{ Q_n + \sqrt{n} E\left[ \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \right] \right\} \\
&= \frac{U_n - \tilde{\alpha} Q_n}{\tilde{\sigma}} \\
&= \frac{1}{\tilde{\sigma}} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[ \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} - \frac{\phi\left(\tilde{A}_i\right)\tilde{\alpha}}{\Phi\left(\tilde{A}_i\right)} + \tilde{\delta}\sqrt{\frac{2}{\pi}} \right] \\
\mathrm{Var}\left[ \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} - \frac{\phi\left(\tilde{A}_i\right)\tilde{\alpha}}{\Phi\left(\tilde{A}_i\right)} + \tilde{\delta}\sqrt{\frac{2}{\pi}} \right] &= E\left[ \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} - \frac{\phi\left(\tilde{A}_i\right)\tilde{\alpha}}{\Phi\left(\tilde{A}_i\right)} + \tilde{\delta}\sqrt{\frac{2}{\pi}} \right)^2 \right] \\
&\quad - \left[ E\left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} - \frac{\phi\left(\tilde{A}_i\right)\tilde{\alpha}}{\Phi\left(\tilde{A}_i\right)} + \tilde{\delta}\sqrt{\frac{2}{\pi}} \right) \right]^2 \\
&= E\left[ \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right)^2 \right] - 2E\left[ \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \frac{\phi\left(\tilde{A}_i\right)\tilde{\alpha}}{\Phi\left(\tilde{A}_i\right)} \right] \\
&\quad + 2\tilde{\delta}\sqrt{\frac{2}{\pi}} E\left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) + \tilde{\alpha}^2 E\left[ \frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \right] - \tilde{\delta}^2 \frac{2}{\pi} - 0 \\
&= 1 + \tilde{\alpha}^2 E\left[ \frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \right].
\end{aligned}$$

Thus,

$$\frac{1}{\sqrt{n}} \frac{\partial l_{inc}^{\star}}{\partial \mu} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} = \frac{U_n - \tilde{\alpha} Q_n}{\tilde{\sigma}} \quad \xrightarrow{d} \quad N\left(0, \quad \frac{1}{\tilde{\sigma}^2} + \frac{\tilde{\alpha}^2 E_{20}}{\tilde{\sigma}^2}\right).$$

And

$$\begin{aligned}
\frac{\partial l_{inc}^{\star}}{\partial \sigma} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{n}{\tilde{\sigma}} + \frac{1}{\tilde{\sigma}\tilde{\alpha}} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \tilde{A}_i - \frac{\tilde{\alpha}}{\tilde{\sigma}} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \\
\frac{1}{\sqrt{n}} \frac{\partial l_{inc}^{\star}}{\partial \sigma} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{V_n}{\tilde{\sigma}} - \frac{2\tilde{\delta}^2}{\tilde{\sigma}\pi}\sqrt{n} + \frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}} U_n - \frac{\tilde{\alpha}}{\tilde{\sigma}} \left\{ R_n + \sqrt{n} E\left[ \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \right] \right\} \\
&= \frac{V_n}{\tilde{\sigma}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}} U_n - \frac{\tilde{\alpha}}{\tilde{\sigma}} R_n \\
&= \frac{1}{\tilde{\sigma}\sqrt{n}} \sum_{i=1}^{n} \left[ \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right)^2 - 1 + \tilde{\delta}\sqrt{\frac{2}{\pi}} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) - \tilde{\alpha} \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \left( \frac{y_i - \tilde{\mu}}{\tilde{\sigma}} \right) \right]
\end{aligned}$$

$$\text{Var}\left[\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2-1+\tilde{\delta}\sqrt{\frac{2}{\pi}}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)-\tilde{\alpha}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] = 2-\frac{2\tilde{\delta}^2}{\pi}-\frac{4\tilde{\delta}^4}{\pi}+\tilde{\alpha}^2E_{22}.$$

Thus,

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial l_{inc}^{\star}}{\partial\sigma}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{V_n}{\tilde{\sigma}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}U_n-\frac{\tilde{\alpha}}{\tilde{\sigma}}R_n \\ &\xrightarrow{d} N\left(0,\quad \frac{2}{\tilde{\sigma}^2}-\frac{2\tilde{\delta}^2}{\pi\tilde{\sigma}^2}-\frac{4\tilde{\delta}^4}{\pi\tilde{\sigma}^2}+\frac{\tilde{\alpha}^2E_{22}}{\tilde{\sigma}^2}\right).\end{aligned}$$

And

$$\begin{aligned}\frac{\partial l_{inc}^{\star}}{\partial\alpha}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{\sqrt{2/\pi}}{\tilde{\alpha}\left(1+\tilde{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\tilde{A}_i+\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{B}_i-2\lambda\tilde{\alpha} \\ \frac{1}{\sqrt{n}}\frac{\partial l_{inc}^{\star}}{\partial\alpha}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{3/2}}U_n-\frac{2\tilde{\delta}\sqrt{n}}{\pi\left(1+\tilde{\alpha}^2\right)^{3/2}}+R_n+\sqrt{n}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\ &\quad+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)\left\{Q_n+\sqrt{n}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right]\right\}-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}} \\ &= -\frac{\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{3/2}}U_n+R_n+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)Q_n-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[-\frac{\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{3/2}}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)+\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right. \\ &\quad\left.+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}-\frac{2\tilde{\alpha}}{\pi\left(1+\tilde{\alpha}^2\right)^2}\right]-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}},\end{aligned}$$

$$\begin{aligned}&\text{Var}\quad\left[-\frac{\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{3/2}}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)+\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right. \\ &\quad\left.-\frac{2\tilde{\alpha}}{\pi\left(1+\tilde{\alpha}^2\right)^2}\right] \\ =\quad &-\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3}+E_{22}+2\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21}+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)^2E_{20}.\end{aligned}$$

As known, $\lambda/\sqrt{n} \to 0$ with $n \to \infty$, so

$$\begin{aligned}
& \frac{1}{\sqrt{n}} \frac{\partial l_{inc}^{\star}}{\partial \alpha} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} \\
= & -\frac{\sqrt{2/\pi}}{(1+\tilde{\alpha}^2)^{3/2}} U_n + R_n + \left( \tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2} \right) Q_n - 2\tilde{\alpha}\frac{\lambda}{\sqrt{n}} \\
\xrightarrow{d} & N\left( 0, \quad -\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3} + E_{22} + 2\left( \tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2} \right) E_{21} + \left( \tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2} \right)^2 E_{20} \right).
\end{aligned}$$

For simplicity, we denote

$$C_{11} = \begin{pmatrix} \frac{\partial^2 l_{inc}^{\star}}{\partial \mu^2} & \frac{\partial^2 l_{inc}^{\star}}{\partial \mu \partial \sigma} \\ \frac{\partial^2 l_{inc}^{\star}}{\partial \sigma \partial \mu} & \frac{\partial^2 l_{inc}^{\star}}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} a & c \\ c & d \end{pmatrix},$$

where $\bar{A}_i = \bar{\alpha}\left(\frac{y_i-\bar{\mu}}{\bar{\sigma}} + \bar{\delta}\sqrt{\frac{2}{\pi}}\right)$, $\bar{B}_i = \frac{y_i-\bar{\mu}}{\bar{\sigma}} + \bar{\delta}\sqrt{\frac{2}{\pi}} + \frac{\bar{\delta}\sqrt{2/\pi}}{1+\bar{\alpha}^2}$ and $\bar{\delta} = \bar{\alpha}\left(1+\bar{\alpha}^2\right)^{-1/2}$, then

$$\begin{aligned}
a \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l_{inc}^{\star}}{\partial \mu^2} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= -\frac{n}{\bar{\sigma}^2} - \frac{\bar{\alpha}^2}{\bar{\sigma}^2} \sum_{i=1}^{n} \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i - \frac{\bar{\alpha}^2}{\bar{\sigma}^2} \sum_{i=1}^{n} \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} \\
c \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l_{inc}^{\star}}{\partial \mu \partial \sigma} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= -\frac{1}{\bar{\sigma}^2\bar{\alpha}} \sum_{i=1}^{n} \bar{A}_i - \frac{1}{\bar{\sigma}^3} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right) + \frac{\bar{\alpha}}{\bar{\sigma}^2} \sum_{i=1}^{n} \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad -\frac{\bar{\alpha}^2}{\bar{\sigma}} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i - \frac{\bar{\alpha}^2}{\bar{\sigma}^3} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right) \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} \\
d \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l_{inc}^{\star}}{\partial \sigma^2} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{n}{\bar{\sigma}^2} - \frac{2}{\bar{\sigma}^3\bar{\alpha}} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right) \bar{A}_i - \frac{1}{\bar{\sigma}^4} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right)^2 + \frac{2\bar{\alpha}}{\bar{\sigma}^3} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad -\frac{\bar{\alpha}^2}{\bar{\sigma}^4} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right)^2 \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i - \frac{\bar{\alpha}^2}{\bar{\sigma}^4} \sum_{i=1}^{n} \left(y_i - \bar{\mu}\right)^2 \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}.
\end{aligned}$$

We also denote:

$$C_{12} = C_{21}^T = \begin{pmatrix} \frac{\partial^2 l_{inc}^{\star}}{\partial \mu \partial \alpha} \\ \frac{\partial^2 l_{inc}^{\star}}{\partial \sigma \partial \alpha} \end{pmatrix} = \begin{pmatrix} b \\ e \end{pmatrix},$$

where

$$
\begin{aligned}
b\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\mu\partial\alpha}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{n\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}} - \frac{1}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} + \frac{1}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2 \\
&\quad + \frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i + \frac{\bar{\alpha}}{\bar{\sigma}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i \\
e\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\sigma\partial\alpha}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right) - \frac{1}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad + \frac{1}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2 + \frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i \\
&\quad + \frac{\bar{\alpha}}{\bar{\sigma}^2}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i.
\end{aligned}
$$

And

$$
\begin{aligned}
C_{22}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\partial^2 l^{\star}_{inc}}{\partial\alpha^2}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= f\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} \\
&= \frac{3\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}}\sum_{i=1}^{n}\bar{A}_i - \frac{2n}{\pi\left(1+\bar{\alpha}^2\right)^3} + \frac{\sqrt{2/\pi}\left(2-\bar{\alpha}^2\right)}{\left(1+\bar{\alpha}^2\right)^{5/2}}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \\
&\quad - \sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\bar{B}_i{}^2 - \sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i{}^2 - 2\lambda.
\end{aligned}
$$

As $(\hat{\mu},\hat{\sigma},\hat{\alpha})$ is restricted to $|\hat{\mu}-\tilde{\mu}|\leq\kappa_{\mu n}\to 0$, $|\hat{\sigma}-\tilde{\sigma}|\leq\kappa_{\sigma n}\to 0$ and $|\hat{\alpha}-\tilde{\alpha}|\leq\kappa_{\alpha n}\to 0$, we can also give that

$$
|\bar{\mu}-\tilde{\mu}|\leq\kappa_{\mu n}, \qquad |\bar{\sigma}-\tilde{\sigma}|\leq\kappa_{\sigma n}, \qquad |\bar{\alpha}-\tilde{\alpha}|\leq\kappa_{\alpha n}.
$$

Then we also have

$$
|\bar{\mu}|\leq|\tilde{\mu}|+\kappa_{\mu n}, \qquad \tilde{\sigma}-\kappa_{\sigma n}\leq\bar{\sigma}\leq\tilde{\sigma}+\kappa_{\sigma n}, \qquad \frac{1}{\bar{\sigma}^2}\leq\frac{1}{\left(\tilde{\sigma}-\kappa_{\sigma n}\right)^2}.
$$

And we denote $\tilde{A}_i = \tilde{\alpha}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}} + \tilde{\delta}\sqrt{\frac{2}{\pi}}\right)$, $\tilde{B}_i = \frac{y_i-\tilde{\mu}}{\tilde{\sigma}} + \tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}$ and $\tilde{\delta} = \tilde{\alpha}\left(1+\tilde{\alpha}^2\right)^{-1/2}$

$$
\begin{aligned}
\left|\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}\right| &\leq \left|-\frac{1}{\bar{\sigma}^2} + \frac{1}{\tilde{\sigma}^2}\right| \\
&\quad + \left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i\right| \\
&\quad + \left|\frac{\bar{\alpha}^2}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right| \\
&\leq \kappa_{\sigma n}\cdot O\left(1\right) + \left(\kappa_{\alpha n} + \kappa_{\sigma n} + \kappa_{\mu n}\right)\cdot O_p\left(1\right) \\
&= o_p\left(1\right) \\
\left|\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}\right| &\leq \left|\frac{\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}} - \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}\right| \\
&\quad + \left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right| \\
&\quad + \left|\frac{1}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2 - \frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i{}^2\right| \\
&\quad + \left|\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\right. \\
&\quad \left. - \frac{\tilde{\alpha}^2\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i\right| \\
&\quad + \left|\frac{\bar{\alpha}}{\bar{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i - \frac{\tilde{\alpha}}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i\right| \\
&\leq \left(\kappa_{\alpha n} + \kappa_{\sigma n}\right)\cdot O\left(1\right) + \left(\kappa_{\alpha n} + \kappa_{\sigma n} + \kappa_{\mu n}\right)\cdot O_p\left(1\right) \\
&= o_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left| \frac{c}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{c}{n} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} \right| \\
\leq & \left| \frac{1}{\bar{\sigma}^2\bar{\alpha}} \frac{1}{n} \sum_{i=1}^{n} \bar{A}_i - \frac{1}{\tilde{\sigma}^2\tilde{\alpha}} \frac{1}{n} \sum_{i=1}^{n} \tilde{A}_i \right| \\
& + \left| \frac{1}{\bar{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu}) - \frac{1}{\tilde{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu}) \right| \\
& + \left| \frac{\bar{\alpha}}{\bar{\sigma}^2} \frac{1}{n} \sum_{i=1}^{n} \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{\tilde{\alpha}}{\tilde{\sigma}^2} \frac{1}{n} \sum_{i=1}^{n} \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \right| \\
& + \left| \frac{\bar{\alpha}^2}{\bar{\sigma}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu}) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i - \frac{\tilde{\alpha}^2}{\tilde{\sigma}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu}) \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \tilde{A}_i \right| \\
& + \left| \frac{\bar{\alpha}^2}{\bar{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu}) \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu}) \frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \right| \\
\leq & \left(\kappa_{\alpha n} + \kappa_{\sigma n} + \kappa_{\mu n}\right) \cdot O_p(1) + \left(\kappa_{\sigma n} + \kappa_{\mu n}\right) \cdot O_p(1) \\
= & o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
& \left| \frac{d}{n} \mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} - \frac{d}{n} \mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} \right| \\
\leq & \left| \frac{1}{\bar{\sigma}^2} - \frac{1}{\tilde{\sigma}^2} \right| + \left| \frac{2}{\bar{\sigma}^3\bar{\alpha}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu}) \bar{A}_i - \frac{2}{\tilde{\sigma}^3\tilde{\alpha}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu}) \tilde{A}_i \right| \\
& + \left| \frac{1}{\bar{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu})^2 - \frac{1}{\tilde{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu})^2 \right| \\
& + \left| \frac{2\bar{\alpha}}{\bar{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu}) \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} - \frac{2\tilde{\alpha}}{\tilde{\sigma}^3} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu}) \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \right| \\
& + \left| \frac{\bar{\alpha}^2}{\bar{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu})^2 \frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)} \bar{A}_i - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu})^2 \frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \tilde{A}_i \right| \\
& + \left| \frac{\bar{\alpha}^2}{\bar{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{\mu})^2 \frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^4} \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{\mu})^2 \frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \right| \\
\leq & \kappa_{\sigma n} \cdot O(1) + \left(\kappa_{\sigma n} + \kappa_{\mu n}\right) \cdot O_p(1) + \left(\kappa_{\alpha n} + \kappa_{\sigma n} + \kappa_{\mu n}\right) \cdot O_p(1) \\
= & o_p(1)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{e}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{e}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}\right| \\
\leq & \left|\frac{\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)-\frac{\sqrt{2/\pi}}{\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\right| \\
& +\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right| \\
& +\left|\frac{1}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i{}^2-\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i{}^2\right| \\
& +\left|\frac{\bar{\alpha}^2\sqrt{2/\pi}}{\bar{\sigma}^2\left(1+\bar{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i-\frac{\tilde{\alpha}^2\sqrt{2/\pi}}{\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i\right| \\
& +\left|\frac{\bar{\alpha}}{\bar{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\bar{\mu}\right)\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i-\frac{\tilde{\alpha}}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\tilde{\mu}\right)\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i\right| \\
\leq & \left(\kappa_{\alpha n}+\kappa_{\sigma n}+\kappa_{\mu n}\right)\cdot O_p\left(1\right) \\
= & o_p\left(1\right)
\end{aligned}
$$

$$
\begin{aligned}
& \left|\frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}-\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}\right| \\
\leq & \left|\frac{3\sqrt{2/\pi}}{\left(1+\bar{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\bar{A}_i-\frac{3\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\tilde{A}_i\right|+\left|\frac{2}{\pi\left(1+\bar{\alpha}^2\right)^3}-\frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3}\right| \\
& +\left|\frac{\sqrt{2/\pi}\left(2-\bar{\alpha}^2\right)}{\left(1+\bar{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}-\frac{\sqrt{2/\pi}\left(2-\tilde{\alpha}^2\right)}{\left(1+\tilde{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right| \\
& +\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)}{\Phi\left(\bar{A}_i\right)}\bar{A}_i\bar{B}_i{}^2-\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i\tilde{B}_i{}^2\right| \\
& +\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\bar{A}_i\right)^2}{\Phi\left(\bar{A}_i\right)^2}\bar{B}_i{}^2-\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i{}^2\right| \\
\leq & \left(\kappa_{\alpha n}+\kappa_{\sigma n}+\kappa_{\mu n}\right)\cdot O_p\left(1\right)+\kappa_{\alpha n}\cdot O\left(1\right) \\
= & o_p\left(1\right).
\end{aligned}
$$

Therefore,

$$
\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right),\qquad \frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right),
$$

$$
\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right),\qquad \frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right),
$$

$$
\frac{e}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{e}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right),\qquad \frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}}=\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}}+o_p\left(1\right).
$$

And with central limit theorem we have

$$
\begin{aligned}
\frac{a}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{1}{\tilde{\sigma}^2}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \\
&= -\frac{1}{\tilde{\sigma}^2}-\frac{\tilde{\alpha}^3}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]-\frac{\tilde{\alpha}^3\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right] \\
&\quad -\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right]+O_p\left(n^{-1/2}\right) \\
&= -\frac{1}{\tilde{\sigma}^2}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{20}+O_p\left(n^{-1/2}\right) \\
\frac{b}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}-\frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}+\frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}{\tilde{A}_i}^2 \\
&\quad +\frac{\tilde{\alpha}^2\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i+\frac{\tilde{\alpha}}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i \\
&= \frac{\sqrt{2/\pi}\tilde{\delta}^3}{\tilde{\sigma}\tilde{\alpha}^3}+\frac{1}{\tilde{\sigma}}\left(\frac{2\tilde{\alpha}^2\tilde{\delta}^2}{\pi}+\frac{2\tilde{\delta}^4}{\pi}-1\right)E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right] \\
&\quad +\frac{\tilde{\alpha}^2}{\tilde{\sigma}}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}\left(2\tilde{\alpha}^2+\tilde{\delta}^2\right) \\
&\quad \times E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}\tilde{\alpha}}\left(\tilde{\alpha}^2+\tilde{\delta}^2\right)E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right] \\
&\quad +O_p\left(n^{-1/2}\right) \\
&= \frac{\sqrt{2/\pi}\tilde{\delta}^3}{\tilde{\sigma}\tilde{\alpha}^3}-\frac{\sqrt{2/\pi}\tilde{\delta}}{\tilde{\sigma}\tilde{\alpha}}+\frac{\sqrt{2/\pi}\tilde{\delta}^3}{\tilde{\sigma}\tilde{\alpha}}+\frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{20} \\
&\quad +\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{21}+O_p\left(n^{-1/2}\right) \\
&= \frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{20}+\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{21}+O_p\left(n^{-1/2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{c}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{1}{\tilde{\sigma}^2\tilde{\alpha}}\frac{1}{n}\sum_{i=1}^{n}\tilde{A}_i - \frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right) + \frac{\tilde{\alpha}}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \\
&\quad -\tilde{\alpha}^2\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \\
&= -\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}^2} + \frac{\tilde{\alpha}}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right] - \tilde{\alpha}^3 E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] \\
&\quad -\tilde{\alpha}^3\tilde{\delta}\sqrt{\frac{2}{\pi}}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&\quad +O_p\left(n^{-1/2}\right) \\
&= -\tilde{\delta}^3\sqrt{\frac{2}{\pi}} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{21} + O_p\left(n^{-1/2}\right) \\
\frac{d}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{1}{\tilde{\sigma}^2} - \frac{2}{\tilde{\sigma}^2\tilde{\alpha}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\tilde{A}_i - \frac{1}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2 \\
&\quad +\frac{2\tilde{\alpha}}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i \\
&\quad -\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2} \\
&= \frac{2}{\tilde{\sigma}^2}\left(\frac{3\tilde{\delta}^2}{\pi}-1\right) + \frac{2\tilde{\alpha}}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&\quad -\frac{\tilde{\alpha}^3}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right] - \frac{\tilde{\alpha}^3\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] \\
&\quad -\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] + O_p\left(n^{-1/2}\right) \\
&= -\frac{2}{\tilde{\sigma}^2} + \frac{2\tilde{\delta}^2}{\pi\tilde{\sigma}^2} + \frac{4\tilde{\delta}^4}{\pi\tilde{\sigma}^2} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{22} + O_p\left(n^{-1/2}\right) \\
&= -\frac{2}{\tilde{\sigma}^2} + \frac{2\tilde{\delta}^2\left(1+2\tilde{\delta}^2\right)}{\pi\tilde{\sigma}^2} - \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{22} + O_p\left(n^{-1/2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{e}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)-\frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} \\
&\quad +\frac{1}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i{}^2+\frac{\tilde{\alpha}^2\sqrt{2/\pi}}{\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^{3/2}}\frac{1}{n} \\
&\quad \times\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i+\frac{\tilde{\alpha}}{\tilde{\sigma}}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i \\
&= \frac{1}{\tilde{\sigma}}\left(\frac{2\tilde{\alpha}^2\tilde{\delta}^2}{\pi}+\frac{2\tilde{\delta}^4}{\pi}-1\right)E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&\quad +\frac{\tilde{\alpha}^2}{\tilde{\sigma}}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right]+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}\left(2\tilde{\alpha}^2+\tilde{\delta}^2\right) \\
&\quad \times E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right]+\frac{\tilde{\alpha}}{\tilde{\sigma}}E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] \\
&\quad +\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}\tilde{\alpha}}\left(\tilde{\alpha}^2+\tilde{\delta}^2\right)E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right]+O_p\left(n^{-1/2}\right) \\
&= -\frac{2\tilde{\delta}^6}{\pi^2\tilde{\sigma}\tilde{\alpha}}+\frac{2\tilde{\delta}^2}{\pi\tilde{\sigma}\tilde{\alpha}}-\frac{2\tilde{\delta}^4}{\pi\tilde{\sigma}\tilde{\alpha}}+\frac{2\tilde{\delta}^6}{\pi\tilde{\sigma}\tilde{\alpha}^3}+\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{22} \\
&\quad +\frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21}+O_p\left(n^{-1/2}\right) \\
&= \frac{2\tilde{\alpha}\left(\pi+2\pi\tilde{\alpha}^2-\tilde{\alpha}^4\right)}{\pi^2\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^3}+\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{22}+\frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21} \\
&\quad +O_p\left(n^{-1/2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{f}{n}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{3\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{5/2}}\frac{1}{n}\sum_{i=1}^{n}\tilde{A}_i - \frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3} + \frac{\sqrt{2/\pi}\left(2-\tilde{\alpha}^2\right)}{\left(1+\tilde{\alpha}^2\right)^{5/2}}\frac{1}{n} \\
&\quad \times \sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\tilde{A}_i\bar{B}_i^2 - \frac{1}{n}\sum_{i=1}^{n}\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\tilde{B}_i^2 \\
&\quad -\frac{2\lambda}{n} \\
&= \frac{2\left(3\tilde{\alpha}^2-1\right)}{\pi\left(1+\tilde{\alpha}^2\right)^3} + \left[\frac{\sqrt{2/\pi}\left(2-\tilde{\alpha}^2\right)}{\left(1+\tilde{\alpha}^2\right)^{5/2}} - \frac{2\sqrt{2/\pi}\tilde{\delta}^3\left(\tilde{\alpha}^2+\tilde{\delta}^2\right)^2}{\pi\tilde{\alpha}^3}\right] \\
&\quad \times E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\right] - \tilde{\alpha}E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^3\right] - \frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\alpha}} \\
&\quad \times \left(3\tilde{\alpha}^2+2\tilde{\delta}^2\right)E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] - \frac{2\tilde{\delta}^2}{\pi\tilde{\alpha}^3}\left(\tilde{\alpha}^2+\tilde{\delta}^2\right) \\
&\quad \times \left(3\tilde{\alpha}^2+\tilde{\delta}^2\right)E\left[\frac{\phi\left(\tilde{A}_i\right)}{\Phi\left(\tilde{A}_i\right)}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] - E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)^2\right] \\
&\quad -\frac{2\sqrt{2/\pi}\tilde{\delta}}{\tilde{\alpha}^2}\left(\tilde{\alpha}^2+\tilde{\delta}^2\right)E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\left(\frac{y_i-\tilde{\mu}}{\tilde{\sigma}}\right)\right] \\
&\quad -\frac{2\tilde{\delta}^2}{\pi\tilde{\alpha}^4}\left(\tilde{\alpha}^2+\tilde{\delta}^2\right)^2E\left[\frac{\phi\left(\tilde{A}_i\right)^2}{\Phi\left(\tilde{A}_i\right)^2}\right] - \frac{2\lambda}{n} + O_p\left(n^{-1/2}\right) \\
&= \frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3} - E_{22} - 2\left(\tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21} \\
&\quad -\left(\tilde{\delta}\sqrt{\frac{2}{\pi}} + \frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)^2 E_{20} - \frac{2\lambda}{n} + O_p\left(n^{-1/2}\right).
\end{aligned}
$$

Then we can reach

$$\begin{aligned}
\frac{a}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\frac{1}{\tilde{\sigma}^2}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{20}+o_p\left(1\right)\\
\frac{b}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{20}+\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{21}+o_p\left(1\right)\\
\frac{c}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= -\tilde{\delta}^3\sqrt{\frac{2}{\pi}}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{21}+o_p\left(1\right)\\
\frac{d}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{2\tilde{\delta}^2\left(1+2\tilde{\delta}^2\right)-2\pi}{\pi\tilde{\sigma}^2}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{22}+o_p\left(1\right)\\
\frac{e}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{2\tilde{\alpha}\left(\pi+2\pi\tilde{\alpha}^2-\tilde{\alpha}^4\right)}{\pi^2\tilde{\sigma}\left(1+\tilde{\alpha}^2\right)^3}+\frac{\tilde{\alpha}}{\tilde{\sigma}}E_{22}+\frac{\tilde{\alpha}}{\tilde{\sigma}}\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21}+o_p\left(1\right)\\
\frac{f}{n}\mid_{\mu=\bar{\mu},\sigma=\bar{\sigma},\alpha=\bar{\alpha}} &= \frac{2}{\pi\left(1+\tilde{\alpha}^2\right)^3}-E_{22}-2\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)E_{21}\\
&\quad -\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)^2E_{20}-\frac{2\lambda}{n}+o_p\left(1\right)\\
\frac{1}{\sqrt{n}}\frac{\partial l_{inc}^{\star}}{\partial\mu}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{U_n-\tilde{\alpha}Q_n}{\tilde{\sigma}}\\
\frac{1}{\sqrt{n}}\frac{\partial l_{inc}^{\star}}{\partial\sigma}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= \frac{V_n}{\tilde{\sigma}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}U_n-\frac{\tilde{\alpha}}{\tilde{\sigma}}R_n\\
\frac{1}{\sqrt{n}}\frac{\partial l_{inc}^{\star}}{\partial\alpha}\mid_{\mu=\tilde{\mu},\sigma=\tilde{\sigma},\alpha=\tilde{\alpha}} &= -\frac{\sqrt{2/\pi}}{\left(1+\tilde{\alpha}^2\right)^{3/2}}U_n+R_n+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)Q_n-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}}.
\end{aligned}$$

From previous calculations we have

$$\begin{aligned}
nC_{11}^{-} &= \frac{\frac{1}{n^2}\begin{pmatrix} df-e^2 & be-cf\\ be-cf & af-b^2\end{pmatrix}_{2\times2}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)},\\
nC_{12}^{-} &= \frac{\frac{1}{n^2}\begin{pmatrix} ce-bd\\ bc-ae\end{pmatrix}_{2\times1}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)},\\
nC_{21}^{-} &= \frac{\frac{1}{n^2}\begin{pmatrix} ce-bd & bc-ae\end{pmatrix}_{1\times2}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)},\\
nC_{22}^{-} &= \frac{\frac{1}{n^2}\left(ad-c^2\right)_{1\times1}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)}.
\end{aligned}$$

Then (6.2.9) can be written as

$$
\begin{aligned}
&\sqrt{n}\begin{pmatrix}\hat{\mu}-\tilde{\mu}\\ \hat{\sigma}-\tilde{\sigma}\end{pmatrix}\\
=\;&\frac{\frac{1}{n^2}\begin{pmatrix}e^2-df & cf-be\\ cf-be & b^2-af\end{pmatrix}|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}\begin{pmatrix}\frac{U_n-\tilde{\alpha}Q_n}{\tilde{\sigma}}\\ \frac{V_n}{\tilde{\sigma}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}U_n-\frac{\tilde{\alpha}}{\tilde{\sigma}}R_n\end{pmatrix}_{2\times 1}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}}\\
&+\frac{\frac{1}{n^2}\begin{pmatrix}bd-ce\\ ae-bc\end{pmatrix}|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}\left(-\frac{\sqrt{2/\pi}}{(1+\tilde{\alpha}^2)^{3/2}}U_n+R_n+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)Q_n-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}}\right)}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}}
\end{aligned}
$$

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\alpha}-\tilde{\alpha}\right)\\
=\;&\frac{\frac{1}{n^2}\begin{pmatrix}bd-ce & ae-bc\end{pmatrix}|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}\begin{pmatrix}\frac{U_n-\tilde{\alpha}Q_n}{\tilde{\sigma}}\\ \frac{V_n}{\tilde{\sigma}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{\tilde{\sigma}}U_n-\frac{\tilde{\alpha}}{\tilde{\sigma}}R_n\end{pmatrix}_{2\times 1}}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}}\\
&+\frac{\frac{1}{n^2}\left(c^2-ad\right)|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}\left(-\frac{\sqrt{2/\pi}}{(1+\tilde{\alpha}^2)^{3/2}}U_n+R_n+\left(\tilde{\delta}\sqrt{\frac{2}{\pi}}+\frac{\tilde{\delta}\sqrt{2/\pi}}{1+\tilde{\alpha}^2}\right)Q_n-2\tilde{\alpha}\frac{\lambda}{\sqrt{n}}\right)}{\frac{1}{n^3}\left(adf-c^2f-b^2d+2bce-ae^2\right)|_{\bar{\mu},\bar{\sigma},\bar{\alpha}}}.
\end{aligned}
$$

Because $\frac{\lambda}{\sqrt{n}}\to 0$ as $n\to+\infty$, finally we can reach

$$\sqrt{n}\left(\hat{\mu}-\tilde{\mu}\right)\quad\xrightarrow{d}\quad N\left(0,\omega_1^2\right),$$

$$\sqrt{n}\left(\hat{\sigma}-\tilde{\sigma}\right)\quad\xrightarrow{d}\quad N\left(0,\omega_2^2\right),$$

$$\sqrt{n}\left(\hat{\alpha}-\tilde{\alpha}\right)\quad\xrightarrow{d}\quad N\left(0,\omega_3^2\right).$$

Here $\omega_1, \omega_2, \omega_3$ can be calculated using (2.8.1). First the Fisher information matrix is

$$\mathcal{I}(\mu,\sigma,\alpha) = \begin{pmatrix} \frac{1}{\tilde{\sigma}^2} + \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2} E_{20} & \ddots & \ddots \\ \frac{\sqrt{2/\pi}\tilde{\delta}^3}{\tilde{\sigma}^2} + \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2} E_{21} & \frac{2}{\tilde{\sigma}^2} - \frac{2\tilde{\alpha}^2\left(1+3\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^2(1+\tilde{\alpha}^2)^2} + \frac{\tilde{\alpha}^2}{\tilde{\sigma}^2} E_{22} & \ddots \\ -\frac{\sqrt{2/\pi}\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)}{\tilde{\sigma}(1+\tilde{\alpha}^2)^{3/2}} E_{20} & -\frac{2\tilde{\alpha}\left(1+2\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}(1+\tilde{\alpha}^2)^3} - \frac{\sqrt{2/\pi}\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)}{\tilde{\sigma}(1+\tilde{\alpha}^2)^{3/2}} E_{21} & -\frac{2}{\pi(1+\tilde{\alpha}^2)^3} + \frac{2\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)^2}{\pi(1+\tilde{\alpha}^2)^3} E_{20} \\ -\frac{\tilde{\alpha}}{\tilde{\sigma}} E_{21} & -\frac{\tilde{\alpha}}{\tilde{\sigma}} E_{22} & +\frac{2\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{(1+\tilde{\alpha}^2)^{3/2}} E_{21} + E_{22} \end{pmatrix}.$$

Then the determinant of it is

$$\begin{aligned} \det \mathcal{I}(\mu,\sigma,\alpha) &= 4 \cdot \frac{-\pi - 3\pi\tilde{\alpha}^2 - 3\pi\tilde{\alpha}^4 + (4-\pi)\,\tilde{\alpha}^6 + 10\tilde{\alpha}^8 + 4\tilde{\alpha}^{10}}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6} + \frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6} \\ &\quad \times \left[3\pi + (13\pi - 4)\cdot\tilde{\alpha}^2 + (22\pi - 20)\,\tilde{\alpha}^4 + (18\pi - 34)\,\tilde{\alpha}^6\right. \\ &\quad \left. + (7\pi - 20)\,\tilde{\alpha}^8 + (\pi - 4)\,\tilde{\alpha}^{10}\right] E_{20} + \frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}} \\ &\quad \left[2\pi + (7\pi - 4)\,\tilde{\alpha}^2 + (9\pi - 12)\,\tilde{\alpha}^4 + (5\pi - 8)\,\tilde{\alpha}^6 + \pi\tilde{\alpha}^8 + \tilde{\alpha}^{10}\right] E_{21} \\ &\quad + \frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi + (3\pi - 4)\,\tilde{\alpha}^2 + (3\pi - 8)\,\tilde{\alpha}^4 + \pi\tilde{\alpha}^6 + 2\tilde{\alpha}^8\right] E_{22} \\ &\quad + \frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3} \cdot \left[-\pi - 3\pi\tilde{\alpha}^2 - (2+3\pi)\,\tilde{\alpha}^4 - (2+\pi)\,\tilde{\alpha}^6\right] E_{21}^2 \\ &\quad + \frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left(1 - 2\tilde{\alpha}^2 - \tilde{\alpha}^4\right) \cdot E_{20}E_{21} + \frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3} \\ &\quad \times \left[\pi + 3\pi\tilde{\alpha}^2 + (3\pi - 4)\,\tilde{\alpha}^4 + (\pi - 2)\,\tilde{\alpha}^6\right] E_{20}E_{22}. \end{aligned}$$

And the diagonal elements of adjoint matrix are

$$
\begin{aligned}
\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{11} &= -\frac{4\left(\pi+3\pi\tilde{\alpha}^2+3\pi\tilde{\alpha}^4+\pi\tilde{\alpha}^6+\tilde{\alpha}^6\right)}{\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^6} \\
&\quad +\frac{4\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)^2\left[\pi+\left(2\pi-1\right)\tilde{\alpha}^2+\left(\pi-3\right)\tilde{\alpha}^4\right]}{\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^5}E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{9/2}} \\
&\quad \times\left[\pi+\left(3\pi-2\right)\tilde{\alpha}^2+3\left(\pi-2\right)\tilde{\alpha}^4+\left(\pi-3\right)\tilde{\alpha}^6\right]E_{21}+\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3} \\
&\quad \times\left[\pi+\left(3\pi-4\right)\tilde{\alpha}^2+\left(3\pi-8\right)\tilde{\alpha}^4+\left(\pi-3\right)\tilde{\alpha}^6\right]E_{22} \\
&\quad +\frac{2\tilde{\alpha}^4\left(2+\tilde{\alpha}^2\right)^2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}\left(E_{20}E_{22}-E_{21}^2\right) \\
\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{22} &= -\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}+\frac{2\tilde{\alpha}^2\left(3+4\tilde{\alpha}^2+\tilde{\alpha}^4\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}E_{20}+\frac{2\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{3/2}}E_{21} \\
&\quad +\frac{1}{\tilde{\sigma}^2}E_{22}+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{20}E_{22}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{21}^2 \\
\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{33} &= \frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+\left(3\pi-1\right)\tilde{\alpha}^2+\left(3\pi-4\right)\tilde{\alpha}^4+\left(\pi-4\right)\tilde{\alpha}^6\right] \\
&\quad +\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^2}\cdot\left[\pi+\left(2\pi-1\right)\tilde{\alpha}^2+\left(\pi-3\right)\tilde{\alpha}^4\right]E_{20} \\
&\quad -\frac{2\sqrt{2/\pi}\tilde{\alpha}^5}{\tilde{\sigma}^4}E_{21}+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^4}E_{22}+\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{20}E_{22}-\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{21}^2.
\end{aligned}
$$

Therefore, the variances we want are

$$
\begin{aligned}
\omega_1^2 &= \frac{\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{11}}{\det\mathcal{I}\left(\mu,\sigma,\alpha\right)} \\
&= \left\{-\frac{\left(\pi+3\pi\tilde{\alpha}^2+3\pi\tilde{\alpha}^4+\pi\tilde{\alpha}^6+\tilde{\alpha}^6\right)}{0.25\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^6}+\frac{4\tilde{\alpha}^2\left(2+\tilde{\alpha}^2\right)^2\left[\pi+\left(2\pi-1\right)\tilde{\alpha}^2+\left(\pi-3\right)\tilde{\alpha}^4\right]}{\pi^2\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^5}\right. \\
&\quad \times E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[\pi+\left(3\pi-2\right)\tilde{\alpha}^2+3\left(\pi-2\right)\tilde{\alpha}^4+\left(\pi-3\right)\tilde{\alpha}^6\right]E_{21} \\
&\quad +\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}\cdot\left[\pi+\left(3\pi-4\right)\tilde{\alpha}^2+\left(3\pi-8\right)\tilde{\alpha}^4+\left(\pi-3\right)\tilde{\alpha}^6\right]E_{22} \\
&\quad \left.+\frac{2\tilde{\alpha}^4\left(2+\tilde{\alpha}^2\right)^2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}\left(E_{20}E_{22}-E_{21}^2\right)\right\}\left\{\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+\left(4-\pi\right)\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{0.25\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\right. \\
&\quad +\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+\left(13\pi-4\right)\tilde{\alpha}^2+\left(22\pi-20\right)\tilde{\alpha}^4+\left(18\pi-34\right)\tilde{\alpha}^6+\left(7\pi-20\right)\tilde{\alpha}^8\right. \\
&\quad \left.+\left(\pi-4\right)\tilde{\alpha}^{10}\right]E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+\left(7\pi-4\right)\tilde{\alpha}^2+\left(9\pi-12\right)\tilde{\alpha}^4+\left(5\pi-8\right)\tilde{\alpha}^6\right. \\
&\quad \left.+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21}+\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+\left(3\pi-4\right)\tilde{\alpha}^2+\left(3\pi-8\right)\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22} \\
&\quad +\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\cdot\left[-\pi-3\pi\tilde{\alpha}^2-\left(2+3\pi\right)\tilde{\alpha}^4-\left(2+\pi\right)\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}} \\
&\quad \left.\times\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right)\cdot E_{20}E_{21}+\frac{2\tilde{\alpha}^2\left[\pi+3\pi\tilde{\alpha}^2+\left(3\pi-4\right)\tilde{\alpha}^4+\left(\pi-2\right)\tilde{\alpha}^6\right]}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}E_{20}E_{22}\right\}^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\omega_2^2 &= \frac{\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{22}}{\det\mathcal{I}\left(\mu,\sigma,\alpha\right)} \\
&= \left\{-\frac{2}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}+\frac{2\tilde{\alpha}^2\left(3+4\tilde{\alpha}^2+\tilde{\alpha}^4\right)}{\pi\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^3}E_{20}+\frac{2\sqrt{2/\pi}\tilde{\alpha}\left(2+\tilde{\alpha}^2\right)}{\tilde{\sigma}^2\left(1+\tilde{\alpha}^2\right)^{3/2}}E_{21}+\frac{1}{\tilde{\sigma}^2}E_{22}\right. \\
&\quad \left.+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{20}E_{22}-\frac{\tilde{\alpha}^2}{\tilde{\sigma}^2}E_{21}^2\right\}\left\{\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+\left(4-\pi\right)\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{0.25\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\right. \\
&\quad +\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+\left(13\pi-4\right)\tilde{\alpha}^2+\left(22\pi-20\right)\tilde{\alpha}^4+\left(18\pi-34\right)\tilde{\alpha}^6+\left(7\pi-20\right)\tilde{\alpha}^8\right. \\
&\quad \left.+\left(\pi-4\right)\tilde{\alpha}^{10}\right]E_{20}+\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+\left(7\pi-4\right)\tilde{\alpha}^2+\left(9\pi-12\right)\tilde{\alpha}^4+\left(5\pi-8\right)\tilde{\alpha}^6\right. \\
&\quad \left.+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21}+\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+\left(3\pi-4\right)\tilde{\alpha}^2+\left(3\pi-8\right)\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22} \\
&\quad +\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[-\pi-3\pi\tilde{\alpha}^2-\left(2+3\pi\right)\tilde{\alpha}^4-\left(2+\pi\right)\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}} \\
&\quad \left.\times\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right)E_{20}E_{21}+\frac{2\tilde{\alpha}^2\left[\pi+3\pi\tilde{\alpha}^2+\left(3\pi-4\right)\tilde{\alpha}^4+\left(\pi-2\right)\tilde{\alpha}^6\right]}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}E_{20}E_{22}\right\}^{-1},
\end{aligned}
$$

$$
\begin{aligned}
\omega_3^2 &= \frac{\mathrm{adj}\mathcal{I}\left(\mu,\sigma,\alpha\right)_{33}}{\det\mathcal{I}\left(\mu,\sigma,\alpha\right)} \\
&= \left\{\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+\left(3\pi-1\right)\tilde{\alpha}^2+\left(3\pi-4\right)\tilde{\alpha}^4+\left(\pi-4\right)\tilde{\alpha}^6\right]+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^2}\right. \\
&\quad \left.\times\left[\pi+\left(2\pi-1\right)\tilde{\alpha}^2+\left(\pi-3\right)\tilde{\alpha}^4\right]E_{20}-\frac{2\sqrt{2/\pi}\tilde{\alpha}^5}{\tilde{\sigma}^4}E_{21}+\frac{\tilde{\alpha}^2}{\tilde{\sigma}^4}E_{22}+\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{20}E_{22}-\frac{\tilde{\alpha}^4}{\tilde{\sigma}^4}E_{21}^2\right\} \\
&\quad \left\{4\cdot\frac{-\pi-3\pi\tilde{\alpha}^2-3\pi\tilde{\alpha}^4+\left(4-\pi\right)\tilde{\alpha}^6+10\tilde{\alpha}^8+4\tilde{\alpha}^{10}}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}+\frac{4\tilde{\alpha}^2}{\pi^2\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^6}\left[3\pi+\left(13\pi-4\right)\right.\right. \\
&\quad \left.\times\tilde{\alpha}^2+\left(22\pi-20\right)\tilde{\alpha}^4+\left(18\pi-34\right)\tilde{\alpha}^6+\left(7\pi-20\right)\tilde{\alpha}^8+\left(\pi-4\right)\tilde{\alpha}^{10}\right]E_{20} \\
&\quad +\frac{4\sqrt{2/\pi}\tilde{\alpha}}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left[2\pi+\left(7\pi-4\right)\tilde{\alpha}^2+\left(9\pi-12\right)\tilde{\alpha}^4+\left(5\pi-8\right)\tilde{\alpha}^6+\pi\tilde{\alpha}^8+\tilde{\alpha}^{10}\right]E_{21} \\
&\quad +\frac{2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+\left(3\pi-4\right)\tilde{\alpha}^2+\left(3\pi-8\right)\tilde{\alpha}^4+\pi\tilde{\alpha}^6+2\tilde{\alpha}^8\right]E_{22}+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3} \\
&\quad \times\left[-\pi-3\pi\tilde{\alpha}^2-\left(2+3\pi\right)\tilde{\alpha}^4-\left(2+\pi\right)\tilde{\alpha}^6\right]E_{21}^2+\frac{4\sqrt{2/\pi}\tilde{\alpha}^5\left(2+\tilde{\alpha}^2\right)}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^{9/2}}\left(1-2\tilde{\alpha}^2-\tilde{\alpha}^4\right) \\
&\quad \left.\times E_{20}E_{21}+\frac{2\tilde{\alpha}^2}{\pi\tilde{\sigma}^4\left(1+\tilde{\alpha}^2\right)^3}\left[\pi+3\pi\tilde{\alpha}^2+\left(3\pi-4\right)\tilde{\alpha}^4+\left(\pi-2\right)\tilde{\alpha}^6\right]E_{20}E_{22}\right\}^{-1}.
\end{aligned}
$$

□

# Appendix C

## Skew Normal Regression

The response variable $y$ is related to the $n$-covariate row vector $x$ by the equation $y = x^T\boldsymbol{\beta} + \sigma(z - \delta\sqrt{2/\pi})$, where $\boldsymbol{\beta}$ is a column vector of unknown regression coefficients. Given observations $(x_i, y_i)_{1\leq i\leq n}$, let $\mathbf{X} = (x_1, ..., x_n)^T$, $\mathbf{y} = (y_1, ..., y_n)^T$. Suppose that $p < n$. If we replace $\mu$ in the skew normal model by $x^T\boldsymbol{\beta}$, the EM algorithm steps can be directly extended to the skew regression model. Let $Y = X\boldsymbol{\beta} + \sigma(Z - \delta\sqrt{2/\pi})$, where $\boldsymbol{\beta}$ and $\sigma$ are location and scale parameters respectively. Then $Y$ has the density function

$$g(y \mid x) = \frac{2}{\sigma}\phi\left(\frac{y - x^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y - x^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\right)$$

with $E[Y \mid X] = X^T\boldsymbol{\beta}$, $\text{var}(Y \mid X) = \sigma^2(1 - 2\delta^2/\pi)$ and Pearson's skewness index $\gamma_1$.

Given independent observations $(\mathbf{X}, \mathbf{y}) = (x_i, y_i)_{1\leq i\leq n}$, the **incomplete-data log-likelihood** is

$$\begin{aligned} l_{inc}(\beta, \eta, \theta \mid \mathbf{X}, \mathbf{y}) &= \frac{n}{2}\log\left(\frac{2}{\pi\sigma^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}\right)^2 \\ &\quad + \sum_{i=1}^{n}\log\left(\Phi\left(\alpha\left(\frac{y_i - x_i\boldsymbol{\beta}}{\sigma} + \delta\sqrt{\frac{2}{\pi}}\right)\right)\right). \end{aligned}$$

Augmenting $(\mathbf{X}, \mathbf{y})$ by $\mathbf{w} = (w_i)_{1\leq i\leq n}$, we form the complete data $(\mathbf{X}, \mathbf{y}, \mathbf{w}) = (x_i, y_i, w_i)_{1\leq i\leq n}$. The **complete-data log-likelihood** is

$$\begin{aligned} l_{com}(\boldsymbol{\beta}, \eta, \theta \mid \mathbf{X}, \mathbf{y}, \mathbf{w}) &= \sum_{i=1}^{n}\log\left(g(x_i, y_i, w_i)\right) \\ &= \frac{n}{2}\log\left(\frac{1}{\sigma^2\pi^2}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}\right)^2 \\ &\quad - \frac{1}{2}\sum_{i=1}^{n}\left(w_i - \alpha\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}\right)\right)^2 \end{aligned}$$

where $\alpha = \delta/\sqrt{1 - \delta^2}$, $\quad \delta = \alpha/\sqrt{1 + \alpha^2}$, $\quad \eta = \log(\sigma)$ and $\theta = 0.5\log\left(\frac{1+\delta}{1-\delta}\right)$.

**Initialisation**

Here the ordinary least squares estimator is used as the initial value for $\boldsymbol{\beta}$. The replacing $y_i$ by the residual $y_i - x_i^T\boldsymbol{\beta}^{(0)}$ and $\bar{y}$ by average of residuals which is $\overline{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(0)}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\boldsymbol{\beta}^{(0)}\right)$. The initialisations are as follows:

$$\begin{aligned}
\boldsymbol{\beta}^{(0)} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - x_i^T\boldsymbol{\beta}^{(0)} - \overline{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(0)}}\right)^2 \\
\gamma_1^{(0)} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T\boldsymbol{\beta}^{(0)} - \overline{\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^{(0)}}\right)^3}{S^3} \\
\delta^{(0)} &= \sqrt{\frac{\pi}{2}}\frac{\left(2\gamma_1^{(0)}/(4-\pi)\right)^{1/3}}{\sqrt{1+\left(2\gamma_1^{(0)}/(4-\pi)\right)^{2/3}}}, \quad \text{when } \left|\gamma_1^{(0)}\right| \leq 0.9952 \\
\theta^{(0)} &= 0.5\log\left(\frac{1-\delta^{(0)}}{1+\delta^{(0)}}\right) \\
\sigma^{(0)2} &= \frac{S^2}{1-2\delta^{(0)2}/\pi} \\
\eta^{(0)} &= \log\left(S/\sqrt{1-2\delta^{(0)2}/\pi}\right).
\end{aligned}$$

**E Step**

Given $\boldsymbol{\beta}^{(v)}$, $\eta^{(v)}$ and $\theta^{(v)}$, the estimates obtained in the $v$-th iteration, consider

$$\begin{aligned}
\Psi(\boldsymbol{\beta},\eta,\theta|\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}) &= E_{\mathbf{w}|\mathbf{X},\mathbf{y},\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}}[l_{com}(\boldsymbol{\beta},\eta,\theta \mid \mathbf{X},\mathbf{y},\mathbf{w})] \\
&= n\log(\frac{1}{\sigma\pi}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}\right)^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}E_{w_i|x_i,y_i,\boldsymbol{\beta}^{(v)},\sigma^{2(v)},\theta^{(v)}}\left[\left(w_i - \alpha\left(\frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}\right)\right)^2\right] \\
&= n\log(\frac{1}{\pi}) - n\eta - \frac{1}{2}\sum_{i=1}^{n}b_i^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\left(1 + (\alpha^{(v)}b_i^{(v)} - 2\alpha b_i)\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})} + (\alpha^{(v)}b_i^{(v)} - \alpha b_i)^2\right).
\end{aligned}$$

Let

$$\begin{aligned}
b_i^{(v)} &= \frac{y_i - x_i^T\boldsymbol{\beta}^{(v)}}{\sigma^{(v)}} + \delta^{(v)}\sqrt{2/\pi} \\
b_i &= \frac{y_i - x_i^T\boldsymbol{\beta}}{\sigma} + \delta\sqrt{2/\pi}.
\end{aligned}$$

We have

$$\begin{aligned}
\frac{\partial b_i}{\partial \boldsymbol{\beta}} &= -x_i/\sigma, \\
\frac{\partial b_i}{\partial \eta} &= -e^{-\eta}(y_i - x_i^T\boldsymbol{\beta}), \quad \frac{\partial^2 b_i}{\partial \eta^2} = e^{-\eta}(y_i - x_i^T\boldsymbol{\beta}) \\
\frac{\partial b_i}{\partial \theta} &= \frac{\partial \delta}{\partial \theta}\sqrt{2/\pi} = \frac{4e^{2\theta}}{(1+e^{2\theta})^2}\sqrt{2/\pi} \\
\frac{\partial^2 b_i}{\partial \theta^2} &= \frac{\partial^2 \delta}{\partial \theta^2}\sqrt{2/\pi} = \frac{8e^{2\theta}(1-e^{2\theta})}{(1+e^{2\theta})^3}\sqrt{2/\pi}.
\end{aligned}$$

**M-Step**

Calculate the derivatives of $\Psi$ block by block as follows:

$$\begin{aligned}
\frac{\partial \Psi}{\partial \boldsymbol{\beta}} &= -\sum_{i=1}^{n} b_i \frac{\partial b_i}{\partial \boldsymbol{\beta}} + \sum_{i=1}^{n} \alpha \frac{\partial b_i}{\partial \boldsymbol{\beta}} \left( \frac{\phi(\alpha^{(v)} b_i^{(v)})}{\Phi(\alpha^{(v)} b_i^{(v)})} + \alpha^{(v)} b_i^{(v)} - \alpha b_i \right) \\
&= \sum_{i=1}^{n} \frac{x_i b_i}{\sigma} - \sum_{i=1}^{n} \frac{x_i \alpha}{\sigma} \left( \frac{\phi(\alpha^{(v)} b_i^{(v)})}{\Phi(\alpha^{(v)} b_i^{(v)})} + \alpha^{(v)} b_i^{(v)} - \alpha b_i \right) \\
\frac{\partial^2 \Psi}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= -\frac{1}{\sigma^2}(1+\alpha^2) \sum_{i=1}^{n} x_i x_i^T = -\frac{1+\alpha^2}{\sigma^2} \mathbf{X}^T \mathbf{X}.
\end{aligned}$$

The other partial derivatives of $\Psi$ in the previous section can be adopted here.

Let

$$D(\alpha, \sigma) = \frac{\alpha\sigma}{1+\alpha^2} \left( \frac{\phi(\alpha^{(v)} b_1^{(v)})}{\Phi(\alpha^{(v)} b_1^{(v)})} + \alpha^{(v)} b_1^{(v)}, ..., \frac{\phi(\alpha^{(v)} b_n^{(v)})}{\Phi(\alpha^{(v)} b_n^{(v)})} + \alpha^{(v)} b_n^{(v)} \right)^T, \quad \mathbf{1} = (1, ..., 1)^T.$$

Then $\frac{\partial \Psi}{\partial \boldsymbol{\beta}} = 0$ implies

$$\boldsymbol{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \left( \mathbf{y} + \mathbf{1}\sigma\delta\sqrt{\frac{2}{\pi}} - D(\alpha, \sigma) \right).$$

So, we can update $\boldsymbol{\beta}^{(v)}$ by

$$\boldsymbol{\beta}^{(v+1)} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \left( \mathbf{y} + \mathbf{1}\sigma^{(v)}\delta^{(v)}\sqrt{\frac{2}{\pi}} - D(\alpha^{(v)}, \sigma^{(v)}) \right).$$

Secondly, similar to previous section, we update $\eta^{(v)}$ by a quadratic equation

$$e^{2\eta} - T_1(\boldsymbol{\beta}, \alpha) \cdot e^{\eta} - \frac{1+\alpha^2}{n} \sum_{i=1}^{n} (y_i - x_i^T \boldsymbol{\beta})^2 = 0$$

with

$$T_1(\boldsymbol{\beta},\alpha)=(1+\alpha^2)\frac{\delta}{n}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}(y_i-x_i^T\boldsymbol{\beta})-\frac{\alpha}{n}\sum_{i=1}^{n}(y_i-x_i^T\boldsymbol{\beta})\left(\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})}+\alpha^{(v)}b_i^{(v)}\right).$$

Solving the above quadratic equation, we have

$$\sigma=e^{\eta}=\frac{1}{2}T_1(\boldsymbol{\beta},\alpha)+\sqrt{\frac{1}{4}T_1(\boldsymbol{\beta},\alpha)^2+\frac{1+\alpha^2}{n}\sum_{i=1}^{n}(y_i-x_i^T\boldsymbol{\beta})^2}.$$

So, we update $\eta^{(v)}$ (namely $\sigma^{(v)}=e^{\eta^{(v)}}$) via

$$\begin{aligned}\sigma^{(v+1)} &= e^{\eta^{(v+1)}}\\ &= \frac{1}{2}T_1(\boldsymbol{\beta}^{(v+1)},\alpha^{(v)})+\sqrt{\frac{1}{4}T_1(\boldsymbol{\beta}^{(v+1)},\alpha^{(v)})^2+\frac{1+\alpha^{(v)2}}{n}\sum_{i=1}^{n}(y_i-x_i^T\boldsymbol{\beta}^{(v+1)})^2}.\end{aligned}$$

After fixing $(\boldsymbol{\beta},\eta)=(\boldsymbol{\beta}^{(v+1)},\eta^{(v+1)})$, we update $\theta^{(v)}$ the same as Section 4.2.1 by using **Newton-Raphson method**.

**Stopping rule**

Then the EM algorithm iteration alternates between **E-step** and **M-step** until:

$$\left|\frac{l_{inc}(\boldsymbol{\beta}^{(v+1)},\eta^{(v+1)},\theta^{(v+1)}\mid\mathbf{X},\mathbf{y})-l_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})}{l_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})}\right|<\varepsilon$$

where $\varepsilon$ is the tolerance ($10^{-8}$ is the default value). And $l_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})$ is the **incomplete-data log-likelihood** function calculated after $v$-th iteration:

$$\begin{aligned}l_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y}) &= \sum_{i=1}^{n}\log\left\{\frac{2}{\sigma^{(v)}}\phi\left(\frac{y_i-x_i^T\boldsymbol{\beta}^{(v)}}{\sigma^{(v)}}+\delta^{(v)}\sqrt{\frac{2}{\pi}}\right)\cdot\right.\\ &\quad\left.\Phi\left(\alpha^{(v)}\left(\frac{y_i-x_i^T\boldsymbol{\beta}^{(v)}}{\sigma^{(v)}}+\delta^{(v)}\sqrt{\frac{2}{\pi}}\right)\right)\right\},\end{aligned}$$

$$\sigma^{(v)}=e^{\eta^{(v)}},\qquad \delta^{(v)}=1-\frac{2}{e^{2\theta^{(v)}}+1},\quad \alpha^{(v)}=\frac{1}{2}\left(e^{\theta^{(v)}}-e^{-\theta^{(v)}}\right).$$

## Simulation

In the model $Y=\mathbf{X}\boldsymbol{\beta}+\sigma(Z-\delta\sqrt{2/\pi})$, same as ordinary skew normal simulation settings ($n=300,\ \mu=2,\ \sigma=0.5,\ \alpha=1.5$), $\mathbf{X}$ is a $300\times20$ matrix sampled from a multivariate normal

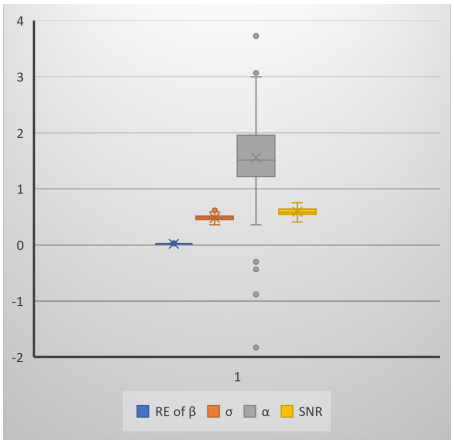



Figure 6.2.1: The estimation result for skew normal regression model under $n = 300$, $p = 20$, $\sigma = 0.5$, $\alpha = 1.5$. The averages of these estimated parameters after the 100 trials are $\overline{RE(\boldsymbol{\beta})} = 0.019$, $\bar{\sigma} = 0.482$, $\bar{\alpha} = 1.556$, $\overline{\text{SNR}} = 0.593$. RE means relative error.

distribution $N_{20}(\mu_x, \Sigma_x)$ where

$$\mu_x = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \Sigma_x = \begin{pmatrix} 1 & 0.1 & 0.01 & 0 & \cdots & 0 \\ 0.1 & 1 & 0.1 & 0.01 & \ddots & \vdots \\ 0.01 & 0.1 & 1 & \ddots & \ddots & 0 \\ 0 & 0.01 & \ddots & \ddots & 0.1 & 0.01 \\ \vdots & \ddots & \ddots & 0.1 & 1 & 0.1 \\ 0 & \cdots & 0 & 0.01 & 0.1 & 1 \end{pmatrix}.$$

Thus, $p = 20$ and $\boldsymbol{\beta}$ is a non-sparse vector with each element sampled from the 0.01-grid points of $[-2, 2]$. To illustrate the results, let's introduce two criteria. Relative error of $\hat{\boldsymbol{\beta}}$ is defined to measure the accuracy of the $\boldsymbol{\beta}$ estimation. With the help of Euclidean norm ($l_2-$norm), relative error of $\hat{\boldsymbol{\beta}}$ is defined as $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{real}\right\|_2 / \|\boldsymbol{\beta}_{real}\|_2$ where $\boldsymbol{\beta}_{real}$ is the real underlying coefficient vector. The other criterion signal-to-noise ratio (SNR) is defined as $\|\mathbf{X}\boldsymbol{\beta}_{real}\|_2 / (n\sigma_{real})$ where $\sigma_{real}$ is the real underlying skew normal scale parameter $\sigma$. After 100 trials, the estimation results with aforementioned methodology are shown in Figure 6.2.1.

## Penalised Skew Normal Regression

Dealing with high-dimensional case ($p$ close to $n$) or sometimes even encountering $p > n$ case, it poses challenges on our aforementioned method. Multicollinearity and overfitting can arise in high-dimensional dataset. To address these problems, we bring in a $l_1-$ norm penalty term to help apply variable selection for potential sparse $\boldsymbol{\beta}$.

**Incomplete and Complete data likelihoods**

$$
\begin{aligned}
pL_{inc} &= \prod_{i=1}^{n} \frac{2}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\right)\exp\left(-\lambda\left|\boldsymbol{\beta}\right|_1\right)\\
pL_{com} &= \prod_{i=1}^{n} \frac{2}{\sigma}\phi\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{\frac{2}{\pi}}\right)\phi\left(w_i-\alpha\left(\frac{y-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{2/\pi}\right)\right)\exp\left(-\lambda\left|\boldsymbol{\beta}\right|_1\right)\\
pl_{com} &= \log\left(pL_{com}\right)\\
&= n\log(\frac{1}{\sigma\pi})-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{2/\pi}\right)^2\\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\left(w_i-\alpha\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}}{\sigma}+\delta\sqrt{2/\pi}\right)\right)^2-n\lambda\left|\boldsymbol{\beta}\right|_1,
\end{aligned}
$$

where $|\boldsymbol{\beta}|_1$ is the Manhattan norm or $l_1-$ norm, known as $|\boldsymbol{\beta}|_1 = \sum_{i=1}^{p}|\boldsymbol{\beta}_i|$. And $\lambda$ is chosen through BIC values.

**Initialisation**

The only difference from the ordinary skew normal regression part is that $\boldsymbol{\beta}^{(0)}$ is initialised through LASSO estimator which is

$$
\boldsymbol{\beta}^{(0)} = \arg\min_{\beta}\left\{\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right\|_2^2+\lambda'\left|\boldsymbol{\beta}\right|_1\right\},
$$

where $\lambda'$ is chosen to give the minimum mean cross-validated error. The initialisation of other parameters keep the same as ordinary skew normal regression part.

**E Step**

Given $\boldsymbol{\beta}^{(v)}$, $\eta^{(v)}$ and $\theta^{(v)}$, the estimates obtained in the $v$-th iteration, consider

$$
\begin{aligned}
\Psi(\boldsymbol{\beta},\eta,\theta|\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}) &= E_{\mathbf{w}|\mathbf{y},\mathbf{X},\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}}[pl_{com}(\boldsymbol{\beta},\eta,\theta\mid\mathbf{y},\mathbf{X},\mathbf{w})]\\
&= n\log(\frac{1}{\pi})-n\eta-n\lambda\left|\boldsymbol{\beta}\right|_1-\frac{1}{2}\sum_{i=1}^{n}b_i^2\\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\left(1+(\alpha^{(v)}b_i^{(v)}-2\alpha b_i)\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})}+(\alpha^{(v)}b_i^{(v)}-\alpha b_i)^2\right).
\end{aligned}
$$

**M-Step**

Calculate the derivatives of $\Psi$ block by block as follows:

$$
\frac{\partial\Psi}{\partial\boldsymbol{\beta}} = \tfrac{1}{\sigma}\sum_{i=1}^{n}\mathbf{x}_i b_i-\frac{\alpha}{\sigma}\sum_{i=1}^{n}\mathbf{x}_i\left(\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})}+\alpha^{(v)}b_i^{(v)}-\alpha b_i\right)-n\lambda\mathrm{sgn}(\boldsymbol{\beta}).
$$

Then solve $\frac{\partial \Psi}{\partial \boldsymbol{\beta}} = \mathbf{0}$ to reach

$$\begin{aligned}\mathbf{0} &= \sum_{i=1}^{n}\left(y_i - \frac{\alpha^{(v)}}{\alpha}\sigma\delta^2 b_i^{(v)} + \sigma\delta\sqrt{\frac{2}{\pi}} - \frac{\sigma\delta^2}{\alpha}\cdot\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})}\right)\mathbf{x}_i \\ &\quad -\sum_{i=1}^{n}\mathbf{x}_i^T\boldsymbol{\beta}\mathbf{x}_i - \sigma^2\left(1-\delta^2\right)n\lambda\cdot\mathrm{sgn}(\boldsymbol{\beta})\end{aligned}$$

which can be represented as

$$\mathbf{X}^{*T}\mathbf{Y}^{*} - \mathbf{X}^{*T}\mathbf{X}^{*}\boldsymbol{\beta} - \lambda^{*}\mathrm{sgn}(\boldsymbol{\beta}) = 0,$$

where $\mathrm{sgn}(\boldsymbol{\beta}) = (\mathrm{sgn}(\beta_1),\ldots,\mathrm{sgn}(\beta_p))^T$, $\lambda^{*} = \sigma^2\left(1-\delta^2\right)n\lambda$, and $\mathbf{X}^{*}$ and $\mathbf{Y}^{*}$ are defined as

$$\begin{aligned}\mathbf{X}^{*} &= \begin{pmatrix}\mathbf{x}_1^T\\ \mathbf{x}_2^T\\ \vdots\\ \mathbf{x}_n^T\end{pmatrix}_{n\times p} = \mathbf{X}\\ \mathbf{Y}^{*} &= \begin{pmatrix} y_1 - \frac{\alpha^{(v)}}{\alpha}\sigma\delta^2 b_1^{(v)} + \sigma\delta\sqrt{\frac{2}{\pi}} - \frac{\sigma\delta^2}{\alpha}\cdot\frac{\phi(\alpha^{(v)}b_1^{(v)})}{\Phi(\alpha^{(v)}b_1^{(v)})}\\ y_2 - \frac{\alpha^{(v)}}{\alpha}\sigma\delta^2 b_2^{(v)} + \sigma\delta\sqrt{\frac{2}{\pi}} - \frac{\sigma\delta^2}{\alpha}\cdot\frac{\phi(\alpha^{(v)}b_2^{(v)})}{\Phi(\alpha^{(v)}b_2^{(v)})}\\ \vdots\\ y_n - \frac{\alpha^{(v)}}{\alpha}\sigma\delta^2 b_n^{(v)} + \sigma\delta\sqrt{\frac{2}{\pi}} - \frac{\sigma\delta^2}{\alpha}\cdot\frac{\phi(\alpha^{(v)}b_n^{(v)})}{\Phi(\alpha^{(v)}b_n^{(v)})}\end{pmatrix}_{n\times 1}\\ \mathbf{Y}^{*(v)} &= \begin{pmatrix} \frac{y_1}{1+\alpha^{(v)2}} + \frac{\alpha^{(v)2}}{1+\alpha^{(v)2}}\mathbf{x}_1^T\boldsymbol{\beta}^{(v)} + \frac{\alpha^{(v)}\sigma^{(v)}}{\left(1+\alpha^{(v)2}\right)^{3/2}}\sqrt{\frac{2}{\pi}} - \frac{\alpha^{(v)}\sigma^{(v)}}{1+\alpha^{(v)2}}\cdot\frac{\phi(\alpha^{(v)}b_1^{(v)})}{\Phi(\alpha^{(v)}b_1^{(v)})}\\ \frac{y_2}{1+\alpha^{(v)2}} + \frac{\alpha^{(v)2}}{1+\alpha^{(v)2}}\mathbf{x}_2^T\boldsymbol{\beta}^{(v)} + \frac{\alpha^{(v)}\sigma^{(v)}}{\left(1+\alpha^{(v)2}\right)^{3/2}}\sqrt{\frac{2}{\pi}} - \frac{\alpha^{(v)}\sigma^{(v)}}{1+\alpha^{(v)2}}\cdot\frac{\phi(\alpha^{(v)}b_2^{(v)})}{\Phi(\alpha^{(v)}b_2^{(v)})}\\ \vdots\\ \frac{y_n}{1+\alpha^{(v)2}} + \frac{\alpha^{(v)2}}{1+\alpha^{(v)2}}\mathbf{x}_n^T\boldsymbol{\beta}^{(v)} + \frac{\alpha^{(v)}\sigma^{(v)}}{\left(1+\alpha^{(v)2}\right)^{3/2}}\sqrt{\frac{2}{\pi}} - \frac{\alpha^{(v)}\sigma^{(v)}}{1+\alpha^{(v)2}}\cdot\frac{\phi(\alpha^{(v)}b_n^{(v)})}{\Phi(\alpha^{(v)}b_n^{(v)})}\end{pmatrix}_{n\times 1}\\ \lambda^{*(v)} &= \frac{\sigma^{(v)2}}{1+\alpha^{(v)2}}n\lambda = \sigma^{(v)2}\left(1-\delta^{(v)2}\right)n\lambda.\end{aligned}$$

To solve this equation, it suffices to solve the following LASSO problem

$$\boldsymbol{\beta}^{(v+1)} = \arg\min_{\boldsymbol{\beta}}\left\{\frac{1}{2}\left\|\mathbf{Y}^{*(v)} - \mathbf{X}^{*}\boldsymbol{\beta}\right\|_2^2 + \lambda^{*(v)}\left|\boldsymbol{\beta}\right|_1\right\}.$$

Similarly, to update $\eta^{(v)}$ (namely $\sigma^{(v)} = e^{\eta^{(v)}}$), we solve $\frac{\partial \Psi}{\partial \eta} = 0$ which is equivalent to the quadratic equation

$$e^{2\eta} - T_1(\boldsymbol{\beta},\alpha)\cdot e^{\eta} - \frac{1+\alpha^2}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2 = 0$$

with

$$T_1(\boldsymbol{\beta},\alpha) \;=\; (1+\alpha^2)\frac{\delta}{n}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})-\frac{\alpha}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})\left(\frac{\phi(\alpha^{(v)}b_i^{(v)})}{\Phi(\alpha^{(v)}b_i^{(v)})}+\alpha^{(v)}b_i^{(v)}\right).$$

Solving the above quadratic equation, we have

$$\sigma \;=\; e^{\eta}=\frac{1}{2}T_1(\boldsymbol{\beta},\alpha)+\sqrt{\frac{1}{4}T_1(\boldsymbol{\beta},\alpha)^2+\frac{1+\alpha^2}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\boldsymbol{\beta})^2}.$$

Therefore, we update $\eta^{(v)}$ (namely $\sigma^{(v)}=e^{\eta^{(v)}}$) via

$$\begin{aligned}\sigma^{(v+1)} &= e^{\eta^{(v+1)}}\\ &= \frac{1}{2}T_1(\boldsymbol{\beta}^{(v+1)},\alpha^{(v)})+\sqrt{\frac{1}{4}T_1(\boldsymbol{\beta}^{(v+1)},\alpha^{(v)})^2+\frac{1+\alpha^{(v)2}}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\boldsymbol{\beta}^{(v+1)})^2}.\end{aligned}$$

After fixing $(\beta,\eta)=(\beta^{(v+1)},\eta^{(v+1)})$, we can also update $\theta^{(v)}$ the same as before by using the Newton-Raphson approach.

**Stopping rule**

Then the EM algorithm iteration alternates between E-step and M-step until

$$\left|\frac{pl_{inc}(\boldsymbol{\beta}^{(v+1)},\eta^{(v+1)},\theta^{(v+1)}\mid\mathbf{X},\mathbf{y})-pl_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})}{pl_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})}\right| \;<\; \varepsilon,$$

where $\varepsilon$ is the tolerance ($10^{-8}$ in our programs). And $pl_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y})$ is the penalised incomplete-data log-likelihood function calculated after $v$-th iteration:

$$\begin{aligned}pl_{inc}(\boldsymbol{\beta}^{(v)},\eta^{(v)},\theta^{(v)}\mid\mathbf{X},\mathbf{y}) &= \sum_{i=1}^{n}\log\left\{\frac{2}{\sigma^{(v)}}\phi\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}^{(v)}}{\sigma^{(v)}}+\delta^{(v)}\sqrt{\frac{2}{\pi}}\right)\right.\\ &\quad\left.\times\Phi\left(\alpha^{(v)}\left(\frac{y_i-\mathbf{x}_i^T\boldsymbol{\beta}^{(v)}}{\sigma^{(v)}}+\delta^{(v)}\sqrt{\frac{2}{\pi}}\right)\right)\exp\left(-\lambda\left|\boldsymbol{\beta}\right|_1\right)\right\},\end{aligned}$$

$$\sigma^{(v)}=e^{\eta^{(v)}},\qquad \delta^{(v)}=1-\frac{2}{e^{2\theta^{(v)}}+1},\quad \alpha^{(v)}=\frac{1}{2}\left(e^{\theta^{(v)}}-e^{-\theta^{(v)}}\right).$$

## Simulation

For this modelling, the only difference from skew normal regression model lies in the underlying $\beta$. To explore the variable selection ability of the penalised skew normal regression model, $p$ is increased from 20 to 200, and $\boldsymbol{\beta}_{real}$ is set as a fixed sparse vector shown in Figure 6.2.2 for all 100 trials.

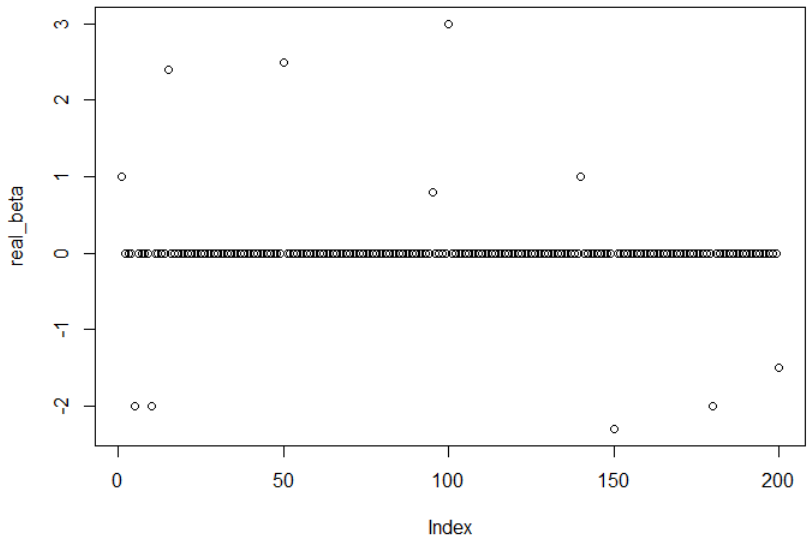



Figure 6.2.2: The fixed sparse underlying real $\beta$ for penalised skew normal regression model simulation trials.

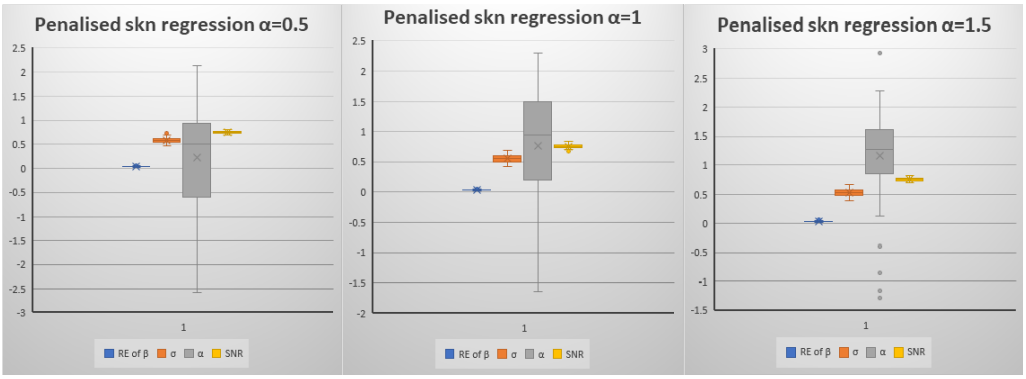



Figure 6.2.3: The penalised skew normal regression estimation results under $\alpha = 0.5, 1.0$ and $1.5$ with underlying real parameters ($n = 300,\ p = 200,\ \sigma = 0.5$). RE means relative error.

To test the estimation abilities under different skew normal shapes, $\alpha$ is set as 0.5, 1.0 and 1.5 for each 100-trial. After all these 300 simulation trials, the estimation results are exhibited in Figure 6.2.3. A smaller shape parameter $\alpha$ can bring in more erratic features as the sample skewness can approach the opposite (negative) direction.

## Inherent Skewness of Anti-Cancer Drug Data Analysed with $\alpha$-Penalised Skew Normal Model

This part provides the detailed table for Section 4.5. In Figure 6.2.4, the drug data has been centralised before calculation with EM algorithm. The drug names are highlighted in green, and the 3 estimates on the right of each drug are $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\alpha}$.

| Drug names | Esti_μ | Esti_σ | Esti_α | Drug names | Esti_μ | Esti_σ | Esti_α | Drug names | Esti_μ | Esti_σ | Esti_α |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Doxorubicin | 0.0064 | 2.0324 | 2.4351 | JW.7.24.1 | 0.0022 | 2.3295 | 1.9653 | Navitoclax | -0.0479 | 3.1136 | -9.0964 |
| Etoposide | 0.0042 | 2.5579 | 2.2213 | NPK76.II.72.1 | 0.0452 | 2.8762 | -3.0881 | Vorinostat | 0.0008 | 1.3617 | 1.2679 |
| Gemcitabine | 0.0505 | 3.7687 | 5.5675 | STF.62247 | -0.0075 | 1.3770 | -2.7379 | Nilotinib | -0.2124 | 2.2723 | -8.5091 |
| Mitomycin.C | 0.0023 | 2.2171 | 2.4459 | NG.25 | -0.0389 | 2.8072 | -2.2232 | CI.1040 | 0.0000 | 2.1069 | -1.8908 |
| Vinorelbine. | 0.0367 | 2.1528 | 4.5727 | TL.1.85 | -0.0321 | 2.9871 | -4.9176 | Temsirolimus | 0.0006 | 1.9519 | -1.4265 |
| NSC.87877 | -0.0076 | 1.3342 | -2.4573 | VX.11e | 0.0248 | 2.5033 | -4.4672 | Veliparib | 0.0007 | 1.0463 | -2.3070 |
| QS11 | -0.0114 | 2.1437 | 2.5178 | FR.180204 | 0.0012 | 1.3951 | -2.9114 | Bosutinib | 0.0000 | 1.4294 | -0.0202 |
| CP466722 | 0.0033 | 2.1465 | 2.2572 | Tubastatin.A | 0.0203 | 2.3260 | -4.7648 | Lenalidomid | -0.0026 | 1.1159 | -2.9341 |
| Midostaurin | -0.0003 | 1.4316 | 1.1844 | Zibotentan | -0.0042 | 0.7546 | -1.4568 | Axitinib | -0.0179 | 1.8517 | -1.8443 |
| Ponatinib | -0.0453 | 2.8326 | -4.1994 | Sepantroniur | 0.3617 | 3.1026 | 27.8853 | AZD7762 | -0.0065 | 1.7654 | -1.1587 |
| JNK.9L | 0.0201 | 1.3462 | 1.9478 | NSC.207895 | 0.0000 | 1.4796 | -0.0164 | GW441756 | -0.0133 | 1.6920 | -3.5447 |
| PF.562271 | 0.0005 | 1.1754 | 1.2141 | VNLG.124 | -0.0542 | 1.5305 | -4.3790 | Lestaurtinib | 0.0000 | 1.3171 | -0.1737 |
| HG6.64.1 | -0.0011 | 1.8100 | -0.8312 | AR.42 | 0.0060 | 2.6686 | 2.7604 | Tanespimycir | -0.0011 | 2.0499 | 0.7032 |
| JQ12 | -0.0021 | 2.0770 | 1.4184 | CUDC.101 | 0.0096 | 2.5567 | 2.0274 | VX.702 | -0.0148 | 1.2325 | -3.2108 |
| DMOG | -0.0003 | 1.5575 | 0.6710 | Belinostat | 0.0469 | 3.0607 | 3.7045 | Motesanib | -0.0480 | 1.5183 | -4.0807 |
| FTI.277 | -0.0060 | 1.1541 | -2.1867 | I.BET.762 | 0.0544 | 2.9495 | -4.3188 | KU.55933 | 0.0002 | 1.2523 | -1.4119 |
| OSU.03012 | 0.0093 | 1.4835 | 1.7872 | CAY10603 | 0.0242 | 2.8101 | 2.6148 | Elesclomol | 0.0045 | 2.2353 | 1.9514 |
| Shikonin | 0.0020 | 1.5121 | 1.3218 | Linifanib | -0.1653 | 1.6313 | -4.6832 | Vismodegib | 0.0034 | 1.3033 | -2.5698 |
| Embelin | 0.0092 | 1.3188 | 2.5082 | BIX02189 | 0.0015 | 1.9038 | -2.3944 | BX795 | -0.0026 | 1.3345 | -0.9664 |
| FH535 | 0.0071 | 1.3618 | 1.2908 | Alectinib | -0.0817 | 1.5467 | -4.3022 | NU7441 | 0.0024 | 1.1164 | -1.2961 |
| PAC.1 | 0.0000 | 1.2419 | -0.0248 | Pelitinib | 0.0074 | 2.5626 | -1.6461 | SL0101 | 0.0012 | 0.9631 | -1.5926 |
| IPA.3 | -0.0186 | 2.5445 | 3.7608 | Omipalisib | -0.0026 | 2.6081 | 1.4572 | Doramapimc | -0.0030 | 1.3976 | -3.0706 |
| GSK650394 | -0.0065 | 2.1126 | 1.5899 | KIN001.236 | 0.0086 | 1.5720 | -2.8909 | JNK.Inhibitor | 0.0014 | 1.0012 | -1.2895 |
| BAY.61.3606 | -0.0006 | 1.9726 | 1.6793 | KIN001.244 | -0.0020 | 1.6331 | -0.9581 | Nutlin.3a.... | 0.0118 | 2.3042 | -5.7845 |
| X5.Fluoroura | 0.0000 | 1.8219 | -0.0169 | WHI.P97 | 0.0092 | 1.6254 | -3.7254 | PD173074 | 0.0008 | 1.6500 | -4.3315 |
| Thapsigargin | 0.0618 | 2.4210 | 3.6143 | KIN001.260 | -0.0011 | 1.9679 | -4.1225 | ZM447439 | 0.0000 | 1.1350 | 0.0129 |
| Obatoclax.M | 0.0070 | 2.1519 | 2.2619 | KIN001.266 | 0.0412 | 1.6771 | 3.8010 | RO.3306 | 0.0171 | 1.6869 | -2.8289 |
| BMS.754807 | -0.0042 | 2.2875 | -1.8306 | Masitinib | -0.0132 | 1.8686 | -1.3681 | MK.2206 | 0.0002 | 1.4059 | -0.5130 |
| Linsitinib | -0.0015 | 2.4949 | -6.7760 | Amuvatinib | -0.0561 | 2.5376 | -3.7717 | Palbociclib | -0.0003 | 1.6709 | 0.5184 |
| Bexarotene | -0.0007 | 1.3183 | -1.1275 | MPS.1.IN.1 | -0.0001 | 1.4811 | 0.4098 | Dactolisib | 0.0030 | 1.2629 | 1.4194 |
| Bleomycin | -0.0030 | 3.0052 | 2.3066 | NVP.BHG712 | -0.0205 | 2.7668 | -2.8435 | AZD8055 | 0.0003 | 1.1497 | -0.9761 |
| LFM.A13 | 0.0014 | 0.9335 | -1.5935 | OSI.930 | -0.0168 | 1.7822 | -4.0582 | PD0325901 | 0.0050 | 2.5865 | -1.4617 |
| GW.2580 | -0.0234 | 1.0004 | -2.0395 | OSI.027 | -0.0807 | 3.9356 | 4.2235 | SB590885 | -0.0396 | 1.8888 | -3.8586 |
| Luminespib | 0.0434 | 1.7035 | 4.4241 | CX.5461 | 0.1029 | 3.5497 | -5.1427 | CCT007093 | -0.0017 | 1.0592 | -2.0000 |
| Phenformin | 0.0006 | 2.4527 | -1.2496 | PHA.793887 | 0.0149 | 2.9927 | -1.7393 | EHT.1864 | 0.0000 | 0.8941 | -0.0131 |
| Bryostatin.1 | 0.0082 | 1.4804 | -5.9105 | PI.103 | -0.0457 | 3.8463 | 2.8569 | Cetuximab | -0.0608 | 1.6003 | -8.0032 |
| Pazopanib | 0.0117 | 2.2975 | -4.9792 | PIK.93 | -0.0002 | 2.4083 | -1.3242 | PF.4708671 | -0.0010 | 1.3961 | -1.9597 |
| Dacinostat | 0.0154 | 1.5635 | 2.5527 | SB52334 | 0.0086 | 2.1175 | -3.3111 | Serdemetan | 0.0007 | 1.1235 | 1.0958 |
| Epothilone.B | 0.0180 | 2.3933 | 2.8082 | TPCA.1 | 0.0011 | 2.0533 | -0.7968 | TW.37 | 0.0013 | 1.2912 | 1.0536 |
| GSK1904529 | -0.0073 | 0.9885 | -2.1695 | Fedratinib | 0.0000 | 1.6520 | -0.0168 | CCT.018159 | -0.0030 | 1.4429 | 2.8437 |
| BMS.345541 | 0.0000 | 1.2390 | 0.0366 | Foretinib | 0.0096 | 1.9682 | 1.2960 | Rucaparib | 0.0069 | 1.5151 | -2.0023 |
| Tipifarnib | 0.0023 | 1.9448 | 1.2006 | Y.39983 | 0.0090 | 2.3748 | -4.8096 | SB505124 | -0.0025 | 1.3736 | -2.7186 |
| Ruxolitinib | -0.0133 | 1.0364 | -2.3466 | YM201636 | 0.0000 | 1.3487 | 0.0159 | Tamoxifen | -0.0011 | 1.0982 | -2.5577 |
| AS601245 | 0.0130 | 1.4024 | 2.0098 | Tivozanib | -0.0391 | 1.1970 | -2.7081 | PFI.1 | 0.0036 | 1.4870 | 2.3364 |
| Ispinesib.Mes | -0.0650 | 2.8983 | 3.9126 | GSK690693 | -0.1143 | 3.0620 | -7.8466 | IOX2 | -0.0086 | 0.9987 | -2.1396 |
| TL.2.105 | -0.0435 | 2.1878 | -4.3138 | SNX.2112 | -0.0219 | 3.4572 | 5.6145 | YK.4.279 | 0.0254 | 1.6529 | 3.1586 |
| AT.7519 | -0.0276 | 2.9796 | 2.6457 | QL.XI.92 | 0.0231 | 2.1177 | -3.0842 | X.5Z..7.Oxoz | -0.0199 | 2.1168 | -2.3315 |
| TAK.715 | 0.0001 | 1.2822 | -0.5135 | XMD13.2 | 0.0002 | 1.4747 | -0.4740 | Piperlongum | 0.0253 | 1.3009 | 3.5875 |
| BX.912 | -0.0013 | 3.0098 | -3.2222 | QL.X.138 | -0.0090 | 2.5636 | 2.0417 | Daporinad | -0.1214 | 4.7752 | 7.3214 |
| ZSTK474 | 0.0000 | 1.8036 | -0.0172 | XMD15.27 | -0.0071 | 1.3365 | -2.9232 | Talazoparib | -0.0022 | 2.6894 | -2.2314 |
| AS605240 | -0.0032 | 2.4920 | -1.7030 | T0901317 | -0.0166 | 1.4785 | -2.6331 | rTRAIL | -0.0643 | 2.2566 | -9.1337 |
| Genentech.C | 0.0000 | 1.8636 | -0.0171 | Selisistat | -0.0059 | 0.9021 | -1.8229 | UNC1215 | 0.0010 | 0.6829 | -1.5421 |
| GSK1070916 | -0.1148 | 3.9317 | -7.7784 | THZ.2.49 | 0.0000 | 2.4094 | 0.0184 | SGC0946 | -0.0008 | 0.7412 | -1.8024 |
| Enzastaurin | 0.0000 | 1.7452 | -0.0172 | KIN001.270 | 0.0052 | 1.3689 | -2.1280 | ICL1100013 | 0.0000 | 1.5869 | 0.0160 |
| GSK429286A | -0.0281 | 1.9189 | -4.3005 | THZ.2.102.1 | 0.0103 | 4.1537 | 3.8515 | XAV939 | -0.0045 | 1.3740 | -2.8923 |
| FMK | -0.0022 | 1.2650 | -2.9247 | AICA.Ribonu | 0.0000 | 1.2760 | 0.0776 | Trametinib | 0.0000 | 2.5828 | -0.1794 |
| QL.XII.47 | 0.0097 | 2.9505 | 4.6097 | Vinblastine | 0.0047 | 1.4732 | 1.5160 | Dabrafenib | -0.1767 | 3.4595 | -7.2048 |
| Idelalisib | -0.0262 | 2.1979 | -5.1443 | Cisplatin | 0.0004 | 1.3546 | 0.9913 | Temozolomi | -0.0082 | 1.1829 | -2.8922 |
| Cabozantinib | -0.0181 | 2.0919 | -1.8242 | Cytarabine | 0.0009 | 1.8965 | 1.1284 | Bleomycin..5 | 0.0000 | 1.9870 | 0.0217 |
| WZ3105 | -0.0108 | 2.6205 | 2.2202 | Docetaxel | -0.0017 | 1.6201 | 1.4565 | SN.38 | -0.0061 | 1.9816 | 1.8426 |
| XMD14.99 | -0.0042 | 1.5258 | -3.2117 | Methotrexate | -0.0023 | 2.8258 | -7.3362 | PFI.3 | -0.0026 | 0.8910 | -2.3885 |
| Quizartinib | -0.1136 | 1.6691 | -3.7435 | Tretinoin | -0.0497 | 2.0216 | -4.5335 | | | | |
| CP724714 | -0.0045 | 1.2523 | -2.3163 | Gefitinib | -0.0414 | 1.9436 | -8.0513 | | | | |

Figure 6.2.4: The detailed estimation results of parameters $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$ for 187 anti-cancer drugs after applied with $\alpha$-penalised skew normal model and corresponding EM algorithm computations.

# Appendix D

## Mixture of Skew Normal Regressions

Following Section 5.1, when it comes to skew normal mixture regressions, the location parameter $\mu_k$ becomes $\mathbf{x}_i^T\boldsymbol{\beta}_k$, then we take VSVA model (see Chapter 5) as an example. Let's consider a finite mixture model in which a set of independent data $(\mathbf{x}_i, y_i)_{1\leq i\leq n}$ come from a $K-$component mixture of skew normal densities

$$g\left(\mathbf{x}_i, y_i; \boldsymbol{\Theta}\right) = \sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma_k}\phi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k}{\sigma_k} + \delta_k\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha_k\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k}{\sigma_k} + \delta_k\sqrt{\frac{2}{\pi}}\right)\right),$$

and $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T, \mathbf{y} = (y_1, ..., y_n)^T$. Here $\boldsymbol{\pi}_{1:K} = (\pi_1, \ldots, \pi_K)$ are the mixing probabilities with constraints $\pi_i \geq 0$ for $i = 1, 2, \ldots, K$ and $\sum_{i=1}^{K} \pi_i = 1$. And $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_K)$ with $\Theta_k = (\pi_k, \boldsymbol{\beta}_k, \sigma_k, \alpha_k)$ being specific parameters for $k$th component.

The **incomplete-data log-likelihood** is

$$l_{inc}(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \cdot \frac{2}{\sigma_k}\phi\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k}{\sigma_k} + \delta_k\sqrt{\frac{2}{\pi}}\right)\Phi\left(\alpha_k\left(\frac{y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k}{\sigma_k} + \delta_k\sqrt{\frac{2}{\pi}}\right)\right)\right).$$

Same as before, we augment $(\mathbf{X}, \mathbf{y})$ by $\mathbf{w} = (w_{ik})_{1\leq i\leq n, 1\leq k\leq K}$, and also introduce a set of latent component indicators $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iK})$ with $i = 1, 2, \ldots, n$. We define it with binary values

$$Z_{ik} = \begin{cases} 1 & \text{if } y_i \text{ belongs to component } k, \\ 0 & \text{otherwise.} \end{cases}$$

Thus we form the complete data $(\mathbf{X}, \mathbf{y}, \mathbf{w}, \mathbf{Z}) = (\mathbf{x}_i, y_i, \mathbf{w}_i, \mathbf{Z_i})_{1\leq i\leq n}$. The complete-data log-likelihood is

$$l_{com}(\boldsymbol{\Theta} \mid \mathbf{X}, \mathbf{y}, \mathbf{w}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{k=1}^{K} Z_{ik}\left\{\log\pi_k + \log\frac{1}{\sigma_k\pi} - \frac{1}{2}b_{ik}^2 - \frac{1}{2}\left(w_{ik} - \alpha_k b_{ik}\right)^2\right\}$$

where $b_{ik} = \left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)/\sigma_k + \delta_k\sqrt{2/\pi}$.

## Iterate with EM Algorithm

### E Step

Given $\boldsymbol{\Theta}^{(v)}$, the estimates obtained in the $v$-th iteration, consider

$$
\begin{aligned}
\Psi(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(v)}) &= E_{\mathbf{Z},\mathbf{w}|\mathbf{X},\mathbf{y},\boldsymbol{\Theta}^{(v)}}[l_{com}(\boldsymbol{\Theta} \mid \mathbf{X},\mathbf{y},\mathbf{w},\mathbf{Z})] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \frac{1}{\sigma_k \pi} \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} E\left[Z_{ik}\left(w_{ik}-\alpha_k b_{ik}\right)^2 \mid \mathbf{X},\mathbf{y},\boldsymbol{\Theta}^{(v)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \frac{1}{\sigma_k \pi} \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} E\left[E\left[Z_{ik}\left(w_{ik}-\alpha_k b_{ik}\right)^2 \mid Z_{ik},\mathbf{X},\mathbf{y},\boldsymbol{\Theta}^{(v)}\right]\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \frac{1}{\sigma_k \pi} \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} E\left[Z_{ik}\left(w_{ik}-\alpha_k b_{ik}\right)^2 \mid \mathbf{X},\mathbf{y},\boldsymbol{\Theta}^{(v)}, Z_{ik}=1\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \log \pi_k + \log \frac{1}{\pi}\cdot \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} - \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \eta_k \\
&\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} b_{ik}^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} \hat{Z}_{ik}^{(v)} \left[1 + \left(\alpha_k^{(v)} b_{ik}^{(v)} - 2\alpha_k b_{ik}\right)\right. \\
&\quad \left. \times \frac{\phi(\alpha_k^{(v)} b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)} b_{ik}^{(v)})} + \left(\alpha_k^{(v)} b_{ik}^{(v)} - \alpha_k b_{ik}\right)^2\right]
\end{aligned}
$$

where $b_{ik}^{(v)} = \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k^{(v)}\right)/\sigma_k^{(v)} + \delta_k^{(v)}\sqrt{2/\pi}$ and

$$
\begin{aligned}
\hat{Z}_{ik}^{(v)} &= E\left[Z_{ik} \mid \mathbf{x}_i, y_i, \boldsymbol{\Theta}^{(v)}\right] \\
&= \frac{\hat{\pi}_k^{(v)} \cdot \frac{2}{\hat{\sigma}_k^{(v)}} \phi\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_k^{(v)}}{\hat{\sigma}_k^{(v)}} + \hat{\delta}_k^{(v)}\sqrt{\frac{2}{\pi}}\right) \Phi\left(\hat{\alpha}_k^{(v)}\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_k^{(v)}}{\hat{\sigma}_k^{(v)}} + \hat{\delta}_k^{(v)}\sqrt{\frac{2}{\pi}}\right)\right)}{\sum_{t=1}^{K} \hat{\pi}_t^{(v)} \cdot \frac{2}{\hat{\sigma}_t^{(v)}} \phi\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_t^{(v)}}{\hat{\sigma}_t^{(v)}} + \hat{\delta}_t^{(v)}\sqrt{\frac{2}{\pi}}\right) \Phi\left(\hat{\alpha}_t^{(v)}\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_t^{(v)}}{\hat{\sigma}_t^{(v)}} + \hat{\delta}_t^{(v)}\sqrt{\frac{2}{\pi}}\right)\right)}.
\end{aligned}
$$

**M-Step**

Same as before, we calculate the derivatives of $\Psi$ with respect to $\Theta_k$ block by block. Update the block of $\pi_k$ by

$$\hat{\pi}_k^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}.$$

Then find the partial derivative of $\boldsymbol{\beta}_k$ with the help of $\partial b_{ik}/\partial\boldsymbol{\beta}_k = -\mathbf{x}_i/\sigma_k$ and get

$$\frac{\partial\Psi}{\partial\boldsymbol{\beta}_k} = \tfrac{1}{\sigma_k}\quad\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}b_{ik}\mathbf{x}_i - \frac{\alpha_k}{\sigma_k}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\mathbf{x}_i\left(\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} + \alpha_k^{(v)}b_{ik}^{(v)} - \alpha_k b_{ik}\right).$$

Solve $\frac{\partial\Psi}{\partial\boldsymbol{\beta}_k} = 0$ and reach

$$\begin{aligned}&\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i + \delta_k\sigma_k\sqrt{\frac{2}{\pi}} - \frac{\alpha_k\sigma_k}{1+\alpha_k^2}\cdot\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} - \frac{\alpha_k\sigma_k}{1+\alpha_k^2}\cdot\alpha_k^{(v)}b_{ik}^{(v)}\right)\mathbf{x}_i\\ =\ &\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\mathbf{x}_i^T\boldsymbol{\beta}_k\mathbf{x}_i.\end{aligned}$$

Write it as matrices

$$\mathbf{X}^T\mathrm{diag}\left(\hat{\boldsymbol{Z}}_k^{(v)}\right)\left(\mathbf{y} + \mathbf{1}\sigma_k\delta_k\sqrt{\frac{2}{\pi}} - D(\alpha_k,\sigma_k)\right) = \mathbf{X}^T\mathrm{diag}\left(\hat{\boldsymbol{Z}}_k^{(v)}\right)\mathbf{X}\boldsymbol{\beta}_k,$$

where

$$D(\alpha_k,\sigma_k) = \frac{\alpha_k\sigma_k}{1+\alpha_k^2}\left(\frac{\phi(\alpha_k^{(v)}b_{1k}^{(v)})}{\Phi(\alpha_k^{(v)}b_{1k}^{(v)})} + \alpha_k^{(v)}b_{1k}^{(v)},\ldots,\frac{\phi(\alpha_k^{(v)}b_{nk}^{(v)})}{\Phi(\alpha_k^{(v)}b_{nk}^{(v)})} + \alpha_k^{(v)}b_{nk}^{(v)}\right)^T.$$

After rearrangement we can update $\boldsymbol{\beta}_k$ by

$$\boldsymbol{\beta}_k^{(v+1)} = \left(\mathbf{X}^T\mathrm{diag}\left(\hat{\boldsymbol{Z}}_k^{(v)}\right)\mathbf{X}\right)^{-1}\mathbf{X}^T\mathrm{diag}\left(\hat{\boldsymbol{Z}}_k^{(v)}\right)\left(\mathbf{y} + \mathbf{1}\sigma_k^{(v)}\delta_k^{(v)}\sqrt{\frac{2}{\pi}} - D(\alpha_k^{(v)},\sigma_k^{(v)})\right)$$

in which

$$D(\alpha_k^{(v)},\sigma_k^{(v)}) = \frac{\alpha_k^{(v)}\sigma_k^{(v)}}{1+\alpha_k^{(v)2}}\left(\frac{\phi(\alpha_k^{(v)}b_{1k}^{(v)})}{\Phi(\alpha_k^{(v)}b_{1k}^{(v)})} + \alpha_k^{(v)}b_{1k}^{(v)},\ldots,\frac{\phi(\alpha_k^{(v)}b_{nk}^{(v)})}{\Phi(\alpha_k^{(v)}b_{nk}^{(v)})} + \alpha_k^{(v)}b_{nk}^{(v)}\right)^T.$$

As $b_{ik}$ can also be written as $b_{ik} = \exp(-\eta_k)\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right) + \delta_k\sqrt{2/\pi}$, so we can derive that $\partial b_{ik}/\partial\eta_k = -\exp(-\eta_k)\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)$. Then the partial derivative of $\eta_k$ is

$$\begin{aligned}\frac{\partial\Psi}{\partial\eta_k} &= -\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)} + \left(1+\alpha_k^2\right)e^{-2\eta_k}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)^2 \\ &\quad + \left(1+\alpha_k^2\right)e^{-\eta_k}\delta_k\sqrt{\frac{2}{\pi}}\cdot\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right) \\ &\quad -\alpha_k e^{-\eta_k}\cdot\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)\left(\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} + \alpha_k^{(v)}b_{ik}^{(v)}\right).\end{aligned}$$

Then solve $\partial\Psi/\partial\eta_k = 0$ and reach

$$\begin{aligned}0 &= \sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}e^{2\eta_k} - \left[\left(1+\alpha_k^2\right)\delta_k\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)\right. \\ &\quad \left. -\alpha_k\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)\cdot\left(\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} + \alpha_k^{(v)}b_{ik}^{(v)}\right)\right]e^{\eta_k} \\ &\quad -\left(1+\alpha_k^2\right)\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)^2.\end{aligned}$$

We use $T_6\left(\boldsymbol{\beta}_k, \alpha_k\right)$ to denote

$$\begin{aligned}T_6\left(\boldsymbol{\beta}_k, \alpha_k\right) &= \left(1+\alpha_k^2\right)\delta_k\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right) \\ &\quad -\alpha_k\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right)\cdot\left(\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} + \alpha_k^{(v)}b_{ik}^{(v)}\right).\end{aligned}$$

So we update $\sigma_k = \exp(\eta_k)$ by

$$\begin{aligned}\sigma_k^{(v+1)} &= e^{\eta_k^{(v+1)}} \\ &= \frac{1}{2\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}}T_6\left(\boldsymbol{\beta}_k^{(v+1)}, \alpha_k^{(v)}\right) + \frac{1}{\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}} \\ &\quad \times\sqrt{\frac{1}{4}T_6\left(\boldsymbol{\beta}_k^{(v+1)}, \alpha_k^{(v)}\right)^2 + \left(1+\alpha_k^{(v)2}\right)\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\cdot\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(v+1)}\right)^2}\end{aligned}$$

in which

$$\begin{aligned}T_6\left(\boldsymbol{\beta}_k^{(v+1)}, \alpha_k^{(v)}\right) &= \left(1+\alpha_k^{(v)2}\right)\delta_k^{(v)}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(v+1)}\right) \\ &\quad -\alpha_k^{(v)}\sum_{i=1}^{n}\hat{Z}_{ik}^{(v)}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(v+1)}\right)\cdot\left(\frac{\phi(\alpha_k^{(v)}b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)}b_{ik}^{(v)})} + \alpha_k^{(v)}b_{ik}^{(v)}\right).\end{aligned}$$

From $b_{ik} = \exp(-\eta_k)\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k\right) + (1 - 2/(\exp(2\theta_k)+1))\sqrt{2/\pi}$, we have

$$\frac{\partial b_{ik}}{\partial \theta_k} = \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2}\sqrt{\frac{2}{\pi}}, \quad \frac{\partial \alpha_k}{\partial \theta_k} = \frac{1}{2}\left(e^{\theta_k} + e^{-\theta_k}\right).$$

Then the partial derivative of $\theta_k$ is

$$\begin{aligned}\frac{\partial \Psi}{\partial \theta_k} &= -\sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2} \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} b_{ik} + \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \left[\frac{1}{2}\left(e^{\theta_k} + e^{-\theta_k}\right) b_{ik}\right. \\ &\quad \left. + \alpha_k \sqrt{\frac{2}{\pi}} \cdot \frac{4e^{2\theta_k}}{\left(e^{2\theta_k}+1\right)^2}\right] \cdot \left(\frac{\phi(\alpha_k^{(v)} b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)} b_{ik}^{(v)})} + \alpha_k^{(v)} b_{ik}^{(v)} - \alpha_k b_{ik}\right).\end{aligned}$$

To use the Newton-Raphson approach, we calculate the second-order partial derivative of $\theta_k$

$$\begin{aligned}\frac{\partial^2 \Psi}{\partial \theta_k^2} &= -\frac{8e^{2\theta_k}}{\pi\left(1+e^{2\theta_k}\right)^2} \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} - \sqrt{\frac{2}{\pi}} \cdot 2\delta_k^3 \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} b_{ik} - \frac{1}{4}\left(e^{\theta_k} + e^{-\theta_k}\right)^2 \\ &\quad \times \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} b_{ik}^2 + \sum_{i=1}^{n} \hat{Z}_{ik}^{(v)} \left[\alpha_k b_{ik} - \sqrt{\frac{2}{\pi}} \cdot \frac{2e^{\theta_k}\left(e^{4\theta_k} - 6e^{2\theta_k} + 1\right)}{\left(e^{2\theta_k}+1\right)^3}\right] \\ &\quad \times \left(\frac{\phi(\alpha_k^{(v)} b_{ik}^{(v)})}{\Phi(\alpha_k^{(v)} b_{ik}^{(v)})} + \alpha_k^{(v)} b_{ik}^{(v)} - \alpha_k b_{ik}\right),\end{aligned}$$

where $\theta_k = 0.5 \log\left(\frac{1+\delta_k}{1-\delta_k}\right)$, $b_{ik}^{(v)} = \exp\left(-\eta_k^{(v)}\right)\left(y_i - x_i^T \beta_k^{(v)}\right) + \delta_k^{(v)}\sqrt{2/\pi}$, $\eta_k^{(v)} = \log\left(\sigma_k^{(v)}\right)$ and $\alpha_k^{(v)} = \delta_k^{(v)}/\sqrt{1-\delta_k^{(v)2}}$, $\delta_k = \alpha_k/\sqrt{1+\alpha_k^2}$.

**Stopping rule**

Then the EM algorithm iteration alternates between **E-step** and **M-step** until

$$\left|\frac{l_{inc}(\boldsymbol{\Theta}^{(v+1)} \mid \mathbf{X}, \mathbf{y}) - l_{inc}(\boldsymbol{\Theta}^{(v)} \mid \mathbf{X}, \mathbf{y})}{l_{inc}(\boldsymbol{\Theta}^{(v)} \mid \mathbf{X}, \mathbf{y})}\right| < \varepsilon,$$

where $\varepsilon$ is the tolerance. And $l_{inc}(\boldsymbol{\Theta}^{(v)} \mid \mathbf{X}, \mathbf{y})$ is the incomplete-data log-likelihood function calculated after $v$-th iteration:

$$\begin{aligned}l_{inc}(\boldsymbol{\Theta}^{(v)} \mid \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k^{(v)} \cdot \frac{2}{\sigma_k^{(v)}} \phi\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k^{(v)}}{\sigma_k^{(v)}} + \delta_k^{(v)}\sqrt{\frac{2}{\pi}}\right)\right. \\ &\quad \left. \cdot \Phi\left(\alpha_k^{(v)}\left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}_k^{(v)}}{\sigma_k^{(v)}} + \delta_k^{(v)}\sqrt{\frac{2}{\pi}}\right)\right)\right\},\end{aligned}$$

$$\sigma_k^{(v)} = e^{\eta_k^{(v)}}, \qquad \delta_k^{(v)} = 1 - \frac{2}{e^{2\theta_k^{(v)}}+1}, \qquad \alpha_k^{(v)} = \frac{1}{2}\left(e^{\theta_k^{(v)}} - e^{-\theta_k^{(v)}}\right).$$

## Initialisation

Inspired by the thresholded LASSO (refer to Section 2.2.6), we came up with this two-step method to initialise for the mixture skew normal regression problem.

Different from previous initialisation parts, mixture skew normal regression estimation with EM algorithm needs a two-step initialisation to improve the clustering result. The first step is carried out to obtain classification result as some simulation trials showed that the classification accuracy is not sensitive to the initialisation settings, but the parameter estimation accuracy is. Given the component number $K$, for each component $k \quad (k = 1, 2, \ldots, K)$ the same initial parameter values are applied in the first trial run. The detailed first initialisation is as follows:

$$
\begin{aligned}
\boldsymbol{\beta}_k^{(0)} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
S_k^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(0)} - \overline{\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_k^{(0)}}\right)^2 \\
\gamma_{1k}^{(0)} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_k^{(0)} - \overline{\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_k^{(0)}}\right)^3}{S_k^3} \\
\delta_k^{(0)} &= \sqrt{\frac{\pi}{2}}\frac{\left(2\gamma_{1k}^{(0)}/\left(4-\pi\right)\right)^{1/3}}{\sqrt{1+\left(2\gamma_{1k}^{(0)}/\left(4-\pi\right)\right)^{2/3}}}, \quad \text{when } \left|\gamma_{1k}^{(0)}\right| \leq 0.9952 \\
\theta_k^{(0)} &= 0.5\log\left(\frac{1-\delta_k^{(0)}}{1+\delta_k^{(0)}}\right) \\
\sigma_k^{(0)2} &= \frac{S_k^2}{1-2\delta_k^{(0)2}/\pi} \\
\eta_k^{(0)} &= \log\left(S_k/\sqrt{1-2\delta_k^{(0)2}/\pi}\right).
\end{aligned}
$$

After the first trial run, the classification result $(C_1, C_2, \ldots C_K)$ is reached, where $C_k$ is a vector to indicate the rows belonging to Group $k$. So $\mathbf{y}_{C_k}$ is the sub-vector with all elements from Group $k$, and $\mathbf{X}_{C_k}$ is the sub-matrix with all rows from Group $k$. Then for all $k \in \{1, 2, \ldots K\}$,

the second step initialisation (the one after thresholding) is shown here:

$$
\begin{aligned}
\beta_k^{(0)} &= \left(\mathbf{X}_{C_k}^T \mathbf{X}_{C_k}\right)^{-1} \mathbf{X}_{C_k}^T \mathbf{y}_{C_k} \\
S_k^2 &= \frac{1}{n_k - 1} \sum_{i=1}^{n_k} \left( y_{C_k,i} - \mathbf{x}_{C_k,i}^T \boldsymbol{\beta}_k^{(0)} - \overline{\mathbf{y}_{C_k} - \mathbf{X}_{C_k} \boldsymbol{\beta}_k^{(0)}} \right)^2 \\
\gamma_{1k}^{(0)} &= \frac{\frac{1}{n} \sum_{i=1}^{n} \left( y_{C_k,i} - \mathbf{x}_{C_k,i}^T \boldsymbol{\beta}_k^{(0)} - \overline{\mathbf{y}_{C_k} - \mathbf{X}_{C_k} \boldsymbol{\beta}_k^{(0)}} \right)^3}{S_k^3} \\
\delta_k^{(0)} &= \sqrt{\frac{\pi}{2}} \frac{\left( 2\gamma_{1k}^{(0)} / (4 - \pi) \right)^{1/3}}{\sqrt{1 + \left( 2\gamma_{1k}^{(0)} / (4 - \pi) \right)^{2/3}}}, \quad \text{when } \left| \gamma_{1k}^{(0)} \right| \leq 0.9952 \\
\theta_k^{(0)} &= 0.5 \log \left( \frac{1 - \delta_k^{(0)}}{1 + \delta_k^{(0)}} \right) \\
\sigma_k^{(0)2} &= \frac{S_k^2}{1 - 2\delta_k^{(0)2} / \pi} \\
\eta_k^{(0)} &= \log \left( S_k / \sqrt{1 - 2\delta_k^{(0)2} / \pi} \right).
\end{aligned}
$$

# Bibliography


[1] Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. 2nd International Symposium on Information Theory. 73. 1033-1055. 10.1007/978-1-4612-1694-0_15.

[2] Almeida, C. A. and Barry, S. A. (2011). Cancer: Basic Science and Clinical Aspects. John Wiley & Sons, Hoboken.

[3] Arellano-Valle, R. B. and Azzalini, A. (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis.* 99. 1362-1382.

[4] Azzalini, A. (1985). A class of distributions which includes the normal ones. Scand. J. Statist. 12, 171-178.

[5] Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. Statistica, 46(2), 199-208.

[6] Azzalini, A. (2005). The skew-normal distribution and related multivariate families. Scand. J. Stat. 32, pp. 159–188.

[7] Azzalini, A. (2014). The Skew-Normal and Related Families. Cambridge University Press, Cambridge. 10.1017/CBO9781139248891.

[8] Azzalini, A. and Arellano-Valle, R. B. (2012). Maximum penalized likelihood estimation for skew-normal and skew-$t$ distributions. Journal of Statistical Planning and Inference. 143. 10.1016/j.jspi.2012.06.022.

[9] Azzalini, A. & Capitanio, A. (1999). Statistical applications of the multivariate skew-normal distribution. Journal of the Royal Statistical Society, Series B, 61, pp. 579-602.

[10] Baghfalaki, T. and Ganjali, M. (2011). An EM Estimation Approach for Analyzing Bivariate Skew Normal Data with Non monotone Missing Values. Communications in Statistics - Theory and Methods. 40:9, 1671-1686.

[11] Bai, Z. and Silverstein, J. W. (2010). Spectral Analysis of Large Dimensional Random Matrices. 2nd ed. Springer, New York.

[12] Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49**, 803-821.

[13] Bastian M., Heymann S. and Jacomy M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

[14] Bühlmann, P. and Meier, L. (2008). Discussion: One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics* **36**(4):1534-1541.

[15] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods. Theory and Applications*. Springer, New York.

[16] Carter, B. Z., Mak, P. Y., Mu, H., Zhou, H., Mak, D. H., Schober, W., Leverson, J. D., Zhang, B., Bhatia, R., Huang, X., Cortes, J., Kantarjian, H., Konopleva, M. and Andreeff, M. (2016). Combined Targeting of BCL-2 and BCR-ABL Tyrosine Kinase Eradicates Chronic Myeloid Leukemia Stem Cells. Science Translational Medicine. 8. 355ra117-355ra117. 10.1126/scitranslmed.aag1180.

[17] D'Agostino, R. B. (1970). Transformation to Normality of the Null Distribution of G1. Biometrika, 57, 3, 679-681.

[18] Daubechies, I., DeVore, R., Fornasier, M. and Güntürk, C. S. (2010). Iteratively Reweighted Least Squares Minimization for Sparse Recovery. *Communications on Pure and Applied Mathematics*. 63. 1 - 38. 10.1002/cpa.20303.

[19] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). 'Maximum likelihood from incomplete data via the EM algorithm'. *Journal of the Royal Statistical Society, Series B (Methodological)*, pp.1-38.

[20] Dodge, Y. (2003). The Oxford Dictionary of Statistical Terms. Oxford University Press, Oxford.

[21] Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57.

[22] Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*. 56 (293): 52–64.

[23] Elk, R. and Landrine, H. (2012). Cancer Disparities: Causes and Evidence-Based Solutions. American Cancer Society. Springer, New York.

[24] Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*. 96. 1348-1360. 10.1198/016214501753382273.

[25] Fan, J. and Li, R. (2002). Variable Selection for Cox's Proportional Hazards Model and Frailty Model. The Annals of Statistics, 30, 74–99.

[26] Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10 (4): 507–521.

[27] Fraley, C. and Raftery, A.E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97/458, pp. 611-631.

[28] Friedman, J., Hastie, T., Simon, N. and Tibshirani, R. (2015). *glmnet: lasso and elastic-net regularized generalized linear models*. R package version 2.0.

[29] Friedman J., Hastie T., Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

[30] Garnett, M. J., Edelman, E. J., Gill, S., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., Liu, Q., Iorio, F., Surdez, D., Chen, L., Milano, R., Bignell, G., Tam, A., Davies, H., Stevenson, J. and Benes, C. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature. 483. 570-5. 10.1038/nature11005.

[31] Geary, R. C. (1947). Testing for Normality. Biometrika, 34, 209-242.

[32] Giraud, C. (2021). Introduction to High-Dimensional Statistics. 10.1201/9781003158745.

[33] Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., Hao, Y. (2015). Improved Variable Selection Algorithm Using a LASSO-Type Penalty, with an Application to Assessing Hepatitis B Infection Relevant Factors in Community Residents. PLoS One, 27;10(7):e0134151.

[34] Gupta, M. R.and Chen, Y. (2011). 'Theory and use of the EM algorithm'. *Foundations and Trends in Signal Processing*: Vol. 4: No. 3, pp. 223-296.

[35] Gupta, M. and Ibrahim, J. (2007). Variable Selection in Regression Mixture Modeling for the Discovery of Gene Regulatory Networks. *Journal of the American Statistical Association*, 102.867-880.10.1198/016214507000000068.

[36] Hastie, T., Mazumder, R., Lee, J.D., Zadeh, R. (2015). Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares. Journal of Machine Learning Research, 16: 3367-3402.

[37] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, Florida.

[38] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), pp. 55-67.

[39] Huang, M. and Yao, W. (2012). Mixture of Regression Models with Varying Mixing Proportions: A Semiparametric Approach. *Journal of the American Statistical Association*, 107(498): 711-724.

[40] Huang, M., Li, R. and Wang, S. (2013). Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association*, 108(503): 929-941.

[41] Huang, M., Yao, W., Wang, S. and Chen, Y. (2018). Statistical Inference and Applications of Mixture of Varying Coefficient Models. *Scandinavian Journal of Statistics*, 45(3): 618-643.

[42] Iorio, F., Knijnenburg, T. A., Vis, D. J., Bignell, G. R., Menden, M. P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., Cokelaer, T., Greninger, P., van Dyk, E., Chang, H., de Silva, H., Heyn, H., Deng, X. M., Egan, R. K., Liu, Q. S., Mironenko, T., Mitropoulos, X., Richardson, L., Wang, J. H., Zhang, T. H., Moran, S., Sayols, S., Soleimani, M., Tamborero, D., Lopez-Bigas, N., Ross-Macdonald, P., Esteller, M., Gray, N. S., Haber, D. A., Stratton, M. R., Benes, C. H., Wessels, L. F. A., Saez-Rodriguez, J., McDermott, U. and Garnett, M. J. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*, **166**, 3, 740-754.

[43] Khalili, A. and Chen, J. (2007). Variable Selection in Finite Mixture of Regression Models. Journal of the American Statistical Association. 102. 1025-1038. 10.1198/016214507000000590.

[44] Khalili, A., Chen, J. and Lin, S. (2011). Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. Biostatistics 12, 156–172.

[45] Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone 1. *New Phytologist*. 11(2): 37-50.

[46] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning. Vol. 112, Springer.

[47] Janku, F., Hong, D. S., Fu, S., Piha-Paul, S. A., Naing, A., Falchook, G. S., Tsimberidou, A. M., Stepanek, V. M. and Moulder, S. L. (2014). Assessing PIK3CA and PTEN in early-phase trials with PI3K/AKT/mTOR inhibitors. *Cell Reports*. 6(2): 377-387.

[48] Kidd, B. A., Readhead, B. P., Eden, C., Parekh, S. and Dudley, J. T. (2015). Integrative network modeling approaches to personalized cancer medicine. Personalized Medicine, 12. 245-257. 10.2217/pme.14.87.

[49] Lai, M. J., Xu, Y. and Yin, W. (2013). Improved iteratively reweighted least squares for unconstrained smoothed lq minimization. SIAM Journal on Numerical Analysis, 5(2), 927-957, 2013. DOI: 10.1137/110840364.

[50] Lin, T. I., Lee, J. C. and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. Stat. Sinica 17, pp. 909–927.

[51] Little, R. J. and Rubin, D. B. (2014). Statistical Analysis with Missing Data. John Wiley & Sons, Hoboken.

[52] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281-297.

[53] Mazumder, R., Hastie, T. and Tibshirani R. (2010). Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 11, 2287-2322.

[54] McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York.

[55] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

[56] Mclachlan, G. J., Lee, S. X. and Rathnayake, S. I. (2019). Finite Mixture Models. Annual Review of Statistics and Its Application, 6. 10.1146/annurev-statistics-031017-100325.

[57] Meinshausen, N. (2007). Relaxed Lasso. Computational Statistics & Data Analysis. 52. 374-393.

[58] Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A. and Sutherland, R. L. (2011). Cyclin D as a therapeutic target in cancer. *Nature Reviews. Cancer*. 11(8): 558–572.

[59] Oftadeh, E. (2017). Complex modelling of multi-outcome data with applications to cancer biology. *Doctor of Philosophy (PhD) thesis*. University of Kent.

[60] Pan, W. and Shen, X. (2007). Penalized Model-Based Clustering with Application to Variable Selection. Journal of Machine Learning Research. 8. 1145-1164.

[61] Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A* **185**, 71-110.

[62] Pearson, K. (1906). Walter Frank Raphael Galton, 1860-1906: a memoir. *Biometrika* **5**, 1-52.

[63] Pewsey, A. (2000). Problems of inference for Azzalini's skewnormal distribution. Journal of Applied Statistics, 27:7, 859-870, DOI: 10.1080/02664760050120542.

[64] Prentice, R. L. (1975). Discrimination among some parametric models. Biometrika, 62, 607-614.

[65] Rand, W. (1971). Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association. 66. 846-850. 10.2307/2284239.

[66] Raftery, A. and Dean, N. (2006). Variable Selection for Model-Based Clustering. Journal of the American Statistical Association, 101. 168-178. 10.1198/016214506000000113.

[67] Rotnitzky, A., Cox, D.R., Bottai, M. and Robins, J. (2000). Likelihood-based inference with singular information matrix. Bernoulli, Vol. 6, 243–284.

[68] Rousseeuw, P. J.; Croux, C. (1993). Alternatives to the median absolute deviation. Journal of the American Statistical Association, 88 (424): 1273–1283.

[69] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Hoboken.

[70] Schwab, M. (2008). *Encyclopedia of Cancer*. Springer Science & Business Media.

[71] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2): 461–464. doi:10.1214/aos/1176344136. MR468014.

[72] Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016). 'mclust 5: clustering, classification and density estimation using Gaussian finite mixture models'. *The R Journal* **8/1**, pp. 205-233.

[73] Shaphiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika. 52. 591-611.

[74] Teimouri, M. (2020). EM algorithm for mixture of skew-normal distributions fitted to grouped data. Journal of Applied Statistics, DOI: 10.1080/02664763.2020.1759032.

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

[76] Turner, M. E. (1960). On heuristic estimation methods. Biometrics, Vol 16, 299-301.

[77] Wang, P., Puterman, M., Cockburn, I. and Le, N. (1996). Mixed Poisson Regression Models with Covariate Dependent Rates. Biometrics. 52. 381-400. 10.2307/2532881.

[78] Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S. A., Bindal, N., Beare, D., Smith, J. A., Thompson, I. R., Ramaswamy, S., Futreal, P. A., Haber, D. A., Stratton, M. R., Benes, C. H., McDermott, U., and Garnett, M. J. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Research, 41, D955 - D961.

[79] Zhang, J. (2017). Screening and clustering of sparse regressions with finite non-Gaussian mixtures. *Biometrics* **73**(2), 540-550.

[80] Zhou, H., Pan, W. and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. Electronic journal of statistics, 3. 1473-1496. 10.1214/09-EJS487.

[81] Zhou, S. (2010). Thresholded Lasso for high dimensional variable selection and statistical estimation. *arXiv: Statistics Theory*. 1002.1583.

[82] Zhu, Z., Wang, T. and Samworth, R. (2019). High Dimensional Principal Component Analysis with Heterogeneous Missingness. ArXiv:1906.12125.

[83] Zhu, Z., Wang, T. and Samworth, R. (2019). primePCA: projected refinement for imputation of missing entries in principal component analysis. R package version 1.0, https://cran.r-project.org/web/packages/primePCA/.

[84] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

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