Clustering and Cluster Inference of Complex Data Structures

Bader Lafi Q Alruwaili

Submitted in fulfilment of the requirements for the

Degree of Doctor of Philosophy

School of Mathematics and Statistics

College of Science and Engineering

University of Glasgow

March 2019

Abstract

Finite mixtures provide a flexible and powerful tool for fitting univariate and multivariate distributions that cannot be captured by standard statistical distributions. In particular, multivariate mixtures have been widely used to perform modelling and cluster analysis of high-dimensional data in a wide range of applications. Modes of mixture densities have been used with great success for organizing mixture components into homogenous groups. But the results are limited to normal mixtures. Beyond the clustering application existing research in this area has provided fundamental results regarding the upper bound of the number of modes, but they too are limited to normal mixtures.

This thesis provides new modality theorems and important analytical results on the upper bound of the number of modes for multivariate t-mixtures and compares them with existing results on normal mixtures . Graphical tools for merging t-mixtures and the effect of degrees-of-freedom are also thoroughly examined.

The most important contribution of this thesis are a set of fundamental results on the modality of skewed normal and skewed normal mixtures. First, we show that the topography of high-dimensional skew normal mixtures can be analyzed rigorously in lower dimensions by defining the corresponding ridgeline manifold that contains all critical points, as well as the ridges of the density. But unlike the normal or t-mixtures we need to solve an implicit equation to obtain this manifold. The plot of the elevations on the ridgeline can still be used to develop tools to explore the number of modes and

for merging mixture components. Though analytical results on the number of modes cannot be explored any more, the elevation plots lead to a new conjecture on the upper bound on the number of modes of skew normal mixture.

Unlike the normal and t-distribution, for skew normal distributions even the one-component counterpart have very interesting modal features. Firstly, as the modes cannot be written in closed form, we design and provide software tools to calculate the modes in any dimensions. We also provide a thorough study exploring the relationship between the means and modes of skew normals and provide fundamental results on the limiting behaviour of the mean and mode as the skewness parameter increases. We also provide another new result showing that though the mean can vary widely as the skewness parameter varies, the mode is a much more robust measure of the central tendency as the mode of skew distribution only varies within a smaller range.

Two R-package available on `github` containing the numerical tools for calculating the modes of skew normals and function specific to merging of skew normal components is provided as part of this thesis. Additionally, application of the merging tool developed of skew normal mixtures is demonstrated using flow-cytomtery data.

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I am sincerely thankful to my God, Allah, for his countless blessings in my life.

My main appreciation and gratitude must go to my supervisor, Dr. Surajit Ray, for his unlimited help and unfailing support. He was always there to give guidance on and encouragement with my PhD research work. His invaluable comments and suggestions were really useful throughout all the stages of my PhD research work.

I also wish to express my thanks to my parents for their unlimited love and support throughout all the various stages in my life. I am very grateful to the people I am now closest to, my dear wife, Tamani, for her unfailing support and constant encouragement throughout my studies abroad, and my lovely sons, Khalid and Rakan, for their love and smiles. I am also thankful to my brothers and sisters and all my friends for their support and encouragement.

Finally, I am grateful to my country, Saudi Arabia, for its support and for funding me to complete my studies in the USA and in the UK. I am also indebted to Glasgow University for making all their facilities available to me so that I could successfully complete my PhD studies.

Declaration

I declare that, all the work presents in this thesis has been done by myself under the supervision of Dr. Surajit Ray, except where otherwise stated. This thesis represent work completed, between 2015 and 2019 in Statistics in the School of Mathematics and Statistics at the University of Glasgow. None of the work described has been submitted to any other university or institute.

Chapter 1

Introduction

In this chapter we will start by stating the general definition and concepts of the finite mixture models. We will then discuss the importance of choosing the number of components in the mixture model and it's connection to modal analyses. In the last two sections of this chapter we will introduce the thesis statement of this research work and provide a brief outline of this dissertation.

1.1 General introduction to finite mixture models

Within the context of increasing scholarly attention over the past century, the whole concept of finite mixture has been extensively researched. Finite mixtures are one of the most popular model based approaches that are routinely used as a powerful tool for modelling heterogeneous data with a finite number of unobserved sub-populations (McLachlan and Peel, 2000). Finite mixtures have applications in several major research areas, such as clustering McLachlan and Basford (1988) and classification Titterington et al. (1985) as well as others. A general definition of the finite mixture model is given below:

Definition 1.1. *Let X be a set of D-dimensional continuous variables, and $x =$*

$(x_1, x_2, ...x_D)$ is an observation of X. The probability density function of the mixture model in general can be written as:

$$g(x) = \sum_{i=1}^{K} \pi_i f_i(x|\theta_i), \tag{1.1}$$

where K is the number of components, $f_i(x|\theta_i)$ are the component densities, and π_i are the mixing proportions of the components with the restriction

$$0 < \pi_i \leq 1 \quad \forall i \qquad \text{and} \qquad \sum_{i=1}^{K} \pi_i = 1.$$

Note, that the component densities of the mixture may have different parameters. For example, in this dissertation the mixture model will be used with different families of distributions, such as multivariate normal distribution, multivariate t-distribution, and multivariate skew normal distribution. As a result, the parameters of the mixture model will change each time a different distribution is being used. The general definition of the mixture in (1.1) also includes mixtures of members from different families, as such as a mixture of t and normal distribution.

The mixture of multivariate normal densities can be written as:

$$g(x) = \sum_{i=1}^{K} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), \quad \boldsymbol{x} \in \mathbb{R}^D, \tag{1.2}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$ represents multivariate normal densities, with mean of the components being $\boldsymbol{\mu_i}$, and the covariance $\boldsymbol{\Sigma_i}$, and π_i being the mixing proportions of the components with the restriction

$$0 \leq \pi_i \leq 1 \quad \forall i \qquad \text{and} \qquad \sum_{i=1}^{K} \pi_i = 1.$$

For the multivariate t-distributions, the mixture of multivariate t-distribution can be written as:

$$g(\boldsymbol{x}) = \sum_{i=1}^{K} \pi_i f(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, n_i), \tag{1.3}$$

where $f(\boldsymbol{x}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}, n_i)$ represents t-multivariate densities, with the mean of the components being $\boldsymbol{\mu_i}$, the covariance $\boldsymbol{\Sigma_i}$, the n_i being the degrees of freedom of t-multivariate components, and π_i being the mixing proportions of the components with the restriction

$$0 < \pi_i \leq 1 \quad \forall i \qquad and \qquad \sum_{i=1}^{K} \pi_i = 1.$$

We want to clarify that in this dissertation we will focus on finite mixture. Infinite mixture is beyond the scope of this dissertation.

1.2 Choosing the number of components in the mixture model

In this section we will address the challenge of choosing the number of components in the mixture models. Choosing the number of components in the mixture model has already been examined but nevertheless still remains as an open question, subject to considerable research. In particular, the mixture model is used and applied to data for two main purposes (McLachlan and Peel, 2000). Firstly, it can be used to provide a semiparametric framework to model unknown distribution. The other main purpose of using a mixture model is to perform model-based clustering. In both of these situations, the question that arises is: how many components should be used when a mixture model is used?

We intend to present a general overview in order to answer this question. Choosing the number of components in the mixture model depends on the family of the

components of the mixture model, as well as the subject matter question we are addressing. For example, in a case when the mixture model is used to provide an appealing semiparametric framework to model an unknown distribution, choosing the number of components is not the focal problem and is consequently also not the most important question (Titterington et al., 1985). However, this is totally different for the case when the mixture model is used to perform model-based clustering. Choosing the number of components in this case, namely, using the mixture model to perform model-based clustering is important, and the number of components must be chosen carefully and also critically addressed (Titterington et al., 1985). Often the components are equated to homogeneous clusters. In practice, using the mixture model with too many numbers of components may indeed make the mixture model overfit the data and thus give poor and unclear interpretations. Furthermore, using the mixture model with only a limited number of components gives the mixture model less flexibility to estimate, and approximate the true underlying data structure, and might fail to distinguish between heterogeneous groups.

There are a wide variety methods available to determine the number of components in the mixture model. The most common of these methods for determining the number of components are AIC (Akaike, 1974) and BIC (Schwarz, 1978). These two methods depend on the likelihood function and the information theoretic criteria to determine the components of the finite mixture model. Leroux (1992) proved some investigating results on the properties of these two methods, namely, AIC (Akaike's Information Criterion) and BIC (Bayesian Information Criterion). By using the maximum penalized likelihood method Keribin (2000) studied the estimation of the number of the components for the mixture models. More recently in the context of high dimensions Ray and Lindsay (2008) proposed a quadratic-risk-based approach to select the right number of components for the finite multivariate mixture. They used a quadratic risk model to

select the number of components with a multivariate normal mixture as an example. Rusakov and Geiger (2012) developed a new asymptotic model selection for a naive Bayesian network model. Choosing the number of components is still the main problem when using the mixture model. Our approach will be different from the previously mentioned approaches. We will start with an over estimate of the number of components for fitting the data. These components do not directly correspond to clusters. Instead we merge all pairwise components using the criterion of whether they display a single mode or not. This technique has been previously used by Hartigan and Hartigan (1985), but only for univariate normal components. We wish to design the mode merging technique for a wide variety of distributions which can optimally merge components in any dimensions. To implement the above approach we first need to understand the modal features of mixtures.

1.3 Number of modes

The mode or modality is a measure of central tendency that is used to describe the region of high concentration in a dataset. The mode is defined as the most frequently occuring value in the dataset. With respect to a probability density a mode is defined as a local maximum of the probability density. For widely used symmetric distributions, such as the normal distribution and the t-distribution, the mode is equal to the mean and median. On the other hand for skewed distributions, e.g skew normal distribution, the mode is usually different from the mean and median. Results relating to the modes of mixtures of distributions and skew normal distributions will be presented throughout the thesis.

Unlike the mean and median, there might be more than one mode in a distribution. A density with one mode is referred to as a unimodal density and a density with more

than one mode is called a multimodal density. For unimodal densities, the density shape can be described by using the concepts of skewness, and kurtosis. In the case of multimodal density, the density is primarily represented by the number and location of the modes. Finite mixture models are powerful tools to model multimodal densities. The important task of estimating the number of components of a mixture model can be intuitively equated to the number of modes in the density. There exists a number of results relating to the number of modes of univariate distributions (Ahsanullah et al., 2014).

Relatively few results have been published concerning the number of modes in the case of the multivariate mixture. Carreira-Perpiñán and Williams (2003) have shown that, for the normal mixture with any number of dimensions, the number of modes does not exceed the number of components if the covariance matrices of the mixture of normal components are equal. The most comprehensive study to determine the number of modes in the case of the multivariate mixture was carried out by Ray and Lindsay (2005). They explored the modality of the multivariate normal mixture by introducing graphical and analytical techniques that could facilitate the topographical study of the multivariate normal mixture in any dimension. In one of their examples Ray and Lindsay (2005) demonstrated the occurrence of three modes in two components and two dimensions. This example showed the possibility of there being more modes than the number of components.

Having more modes than the number of components enables the researcher to think about the maximum number of modes that can be derived from mixing two components. In a notable study on the upper bound of the number of modes in a mixture of multivariate densities, Ray and Ren (2012) showed that the upper bound of the number of modes from mixing two normal components in D dimensions is $D + 1$. For example, if we have two components in two dimensions, then the upper bound of the number of

modes in this case is three modes, and so on. Moreover, they showed that this bound is always achievable in any dimensions and this upper bound of $D+1$ is tight. Here the word tight refers to the fact that the upper bound is achievable.

A range of new research followed the paper by Ray and Lindsay, primarily based on the newly developed framework of ridgeline manifolds. Alexandrovich (2011) studied the number of modes for elliptical densities in general and the multivariate t-mixture in particular. The researcher noted that the ridgeline manifold of the mixture of multivariate normal densities is the same as the ridgeline manifold of the mixture of t-multivariate densities. Regarding the upper bound of the number of modes for the mixture of multivariate t-densities, Alexandrovich et al. (2013) demonstrated that the upper bound of the number of modes for any two components in D-dimensions is $2D+1$. For example, if we have two components with two dimensions, then the upper bound of the modes in this case is five modes, and so on.

In the next chapter we will introduce the previous results on the modality of the mixture of multivariate normal densities, and the modality of the mixture of the t-multivariate densities. One of the objectives of this dissertation is to investigate if the upper bound of the number of modes of the mixture of multivariate normal densities, and the upper bound of the number of mode of a mixture of multivariate t-densities are the same or not. Furthermore, these results will help in exploring the modality of mixtures of multivariate skew normal densities.

1.4 Thesis contribution

The main aims of this research work are to explore the modality of a mixture of multivariate t-densities and investigate if the upper bound of the number of modes of mixture of multivariate t-densities is the same as the upper bound of the number of modes of the mixture of multivariate normal densities. A further aim is to explore the

modality of skew normal densities and also to find the modes of mixtures of univariate and multivariate skew normal densities. Based on the modal properties of mixtures of skew-normal densities a new technique for merging the components of the mixture of multivariate skew normal will be introduced.

The main contributions of this thesis can be summarised as follows:

- To investigate if the upper bound of the number of modes for the mixture of multivariate t-densities is the same as for the upper bound of the number of modes for the mixture of multivariate normal densities.
- To explore the modality of the univariate skew normal density and find mathematical equations to calculate the modes of univariate skew normal densities.
- To explore the relationship between the mean and the mode of univariate skew normal densities and understand their limiting behaviour.
- To explore the behaviour of the univariate and multivariate mode of one component skew-normals.
- To explore the modality of the mixture of multivariate skew normal densities and to find numerical and graphical solutions to calculating the modes.
- To create the computing tools for merging components of skew-normal mixtures.

1.5 Outline of the thesis

This thesis is divided into seven chapters. A brief overview of each chapter and a description of the general structure of this thesis is given below:

In Chapter 2 we provide the reader with the background of related work on the modality of mixtures of multivariate normal densities and the modality of mixtures

of multivariate t-densities. The results from this chapter will be used to develop the research questions and proofs of results in chapters 3 and 5.

In Chapter 3 we investigate the upper bound of the number of modes of the mixture of multivariate t-densities. In particular, we will explore the modality of multivariate t-mixture in two ways, namely, graphically and analytically. Based on the modality studies we build the conjecture that upper bound of the number of modes of the mixture of multivariate t-densities is the same as the upper bound of the number of modes of the mixture of multivariate normal densities.

In Chapter 4 we start with an introduction and a review of the skew normal distribution. We first explore the modality of the one component univariate and multivariate skew normal density. It will be shown that to calculate the mode of a univariate skew normal we need to solve an implicit equation in one variable and a root finding technique can be used to find the numerical solutions. In this chapter we shall also explore the relationship between the mean and the mode of univariate skew normal and their limiting behaviour. We will also explore the behaviour of the mode as a function of the skewness parameter. In the later part of this chapter we will generalize the univariate results for the multivariate skew normal density.

In Chapter 5 we will build the ridgleline manifold for mixtures of skew-normals and develop this framework for graphical and analytical exploration of the modes of mixtures of skew-normal densities. We will use numerical tools to explore the elevation along the ridgeline manifold and to develop algorithms for merging components based on whether the two components display one or more modes.

In Chapter 6 we apply the tools developed in chapter 5 to analyse a real-life dataset on the analysis of flow-cytometry dataset.

In Chapter 7 we review the results of the research work in this thesis and discuss possible future work in this area.

Chapter 2

Statistical Background

In this chapter we will provide a detailed overview of the major works on modality of mixtures of multivariate normals and multivariate-t's. We will start by defining mixtures of densities in arbitrary dimensions and properties of modes of mixtures. We will then present the fundamental results on ridgeline manifolds and develop the framework for the exploration of modes that will be extensively used in the follow-up chapters in the dissertation which will present new results on skew-normal densities.

2.1 Multivariate normal distribution

Before defining mixtures of normals we will start with the definition of multivariate normal distribution. The multivariate normal distribution is parametrised by two important parameters, the mean vector $\boldsymbol{\mu}$, and the covariance matrix $\boldsymbol{\Sigma}$, and often the multivariate normal distribution is denoted by $\boldsymbol{X} \sim N_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Definition 2.1. *Suppose that $\boldsymbol{X}$ is a D-dimensional random vector from a multivariate normal distribution. The probability density function of the multivariate normal is given by:*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \exp^{\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}} \qquad \boldsymbol{x} \in \mathbb{R}^D \tag{2.1}$$

The parameters of the multivariate can be introduce as the following:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_D \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \dots & \sigma_{2D} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \sigma_{D3} & \dots & \sigma_{DD} \end{bmatrix}$$

where $\boldsymbol{\mu}$ is the vector of the means of multivariate normal distribution, and $\boldsymbol{\Sigma}$ is the $D \times D$ variance-covariance matrix of the multivariate normal distribution.

2.2 Mode of the distribution

The modes are defined as the local maximums of the probability density function. For unimodal densities, the density shape can be described by using the concepts of skewness, and kurtosis. When exploring and describing multimodal densities, researchers focus on the number and location of modes. The mode is also the most intuitive measure of central tendency of a theoretical unimodal densities. Graphically, the peak of the density curve indicates the mode of the distribution density, which is the most frequently occurring value in the dataset. Analytically, the density function $f(x)$ is maximized at the mode i.e. for a univariate density the mode is the solution to $f^{'}(x) = 0$, and $f^{''}(x) < 0$. In particular, the mode of univariate normal can be calculated easily by maximizing the $\log f(x)$ instead of the $f(x)$. In particular

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} exp^{\frac{-(x-\mu)^2}{(2\sigma^2)}} \\ \log f(x) &= \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{(x-\mu)^2}{2\sigma^2} \\ \frac{d \log f(x)}{dx} &= 0 - \frac{2(x-\mu)}{2\sigma^2} \end{aligned}$$

Now

$$\frac{d \log f(x)}{dx} = 0, \quad \text{gives } x = \mu.$$

$$\text{Also} \quad \frac{d^2 \log f(x)}{dx} = -\frac{1}{\sigma^2} < 0.$$

This implies the mode is at μ.

This helps us to arrive at the trivial result that the mode of a normal density is its mean and this is a unique mode so the normal density is unimodal. One can also derive

the mode by directly maximizing $f(x)$. For any distribution

$$\frac{d}{dx}\log f(x) = \frac{1}{f(x)}.f^{'}(x) \tag{2.2}$$

$$\text{or} \quad f^{'}(x) = \Big(\frac{d}{dx}\log f(x)\Big).f(x) \tag{2.3}$$

using the derivation in equation $\frac{d\log f(x)}{dx} = 0 - \frac{2(x-\mu)}{2\sigma^2}$ we can write $f^{'}(x) = -\frac{1}{\sigma^2}(x-\mu)f(x)$ which takes the value 0 at $x = \mu$. We can also check

$$\begin{aligned} f^{''}(x) &= -\frac{1}{\sigma^2}\big(1 \times f(x) + (x-\mu)f^{'}(x)\big) \\ &= -\frac{1}{\sigma^2}f(\mu) \quad \text{at} \quad x = \mu \\ \implies \quad f^{''}(\mu) &< 0 \quad \text{as } f(\mu) \text{ is always positive.} \end{aligned}$$

We will use the second technique to calculate the modes for skew normal density and the mixture of multivariate skew normal densities as the logarithm of the mixtures does not simplify the expression.

2.3 Modality of multivariate normal mixture

Unimodal densities are usually defined by the measure of central tendency, variance, skewness and kurtosis. When it comes to exploring and describing a multimodal density researchers focus on the number and location of modes. In this section we will introduce some important results for exploring the modality of multivariate mixture densities. Comprehensive results for understanding the modality of multivariate multimodal density were provided by Ray and Lindsay (2005). They introduced a series of tools to find the number and location of mode for a multivariate normal mixture. In this section we will present the tools to determine the number of modes for a multivariate normal mixture, which will later help us in generalising and facilitating the study of the

number of modes of a multivariate t-mixture and multivariate skew normal mixture. We will also use these results will be used to examine the upper bound of the number of modes of a mixture of multivariate t-densities.

2.3.1 Ridgeline function

The ridgeline function was first introduced by Ray and Lindsay (2005) to find modes in the context of a two component mixture of multivariate normals. Intuitively, the ridgeline function is the unique curve that travels from the mean of the first component to the mean of other component, and retains all the crucial information, such as the number and location of modes for the density, critical points, and saddle points of these components, on the curve of the ridgeline function. In this section we will review the concepts of the ridgeline function, and how the height along the ridgeline and the curvature of the height function can be used to explore the number and location of modes of the mixture of two normal components.

The probability density function of the mixture of multivariate normal densities with K components can be given by

$$g(\boldsymbol{x}) = \sum_{i=1}^{K} \pi_i \phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \quad 0 < \pi_i \leq 1 \quad \forall i \quad \sum_{i=1}^{K} \pi_i = 1.$$

where π_i's are the mixing proportions of the components of multivariate normal density $\phi(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. Ray and Lindsay (2005) introduced the ridgeline manifold of component mixture of multivariate normal densities by

$$x^*(\alpha) = [\alpha_1 \boldsymbol{\Sigma}_1^{-1} + \alpha_2 \boldsymbol{\Sigma}_2^{-1} + ... + \alpha_K \boldsymbol{\Sigma}_K^{-1}]^{-1} \times [\alpha_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \alpha_2 \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2 + ... + \alpha_K \boldsymbol{\Sigma}_K^{-1} \boldsymbol{\mu}_K],$$

where $\alpha_i \in [0, 1]$ is the range of the ridgeline function.

We will mostly be dealing with two component mixtures, so the probability density

function of a two component mixture of multivariate normal densities simplifies to

$$g(\boldsymbol{x}) = \pi_1\phi_1(\boldsymbol{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1) + \pi_2\phi_2(\boldsymbol{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2), \quad 0 < \pi_i \leq 1 \quad \forall i \quad \sum_{i=1}^{K}\pi_i = 1. \tag{2.4}$$

The ridgeline function of multivariate normal mixture (2.4) is given by

$$x^*(\alpha) = [\bar{\alpha}\boldsymbol{\Sigma}_1^{-1} + \alpha\boldsymbol{\Sigma}_2^{-1}]^{-1}[\bar{\alpha}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \alpha\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2], \tag{2.5}$$

where $\alpha \in [0,1]$ is the range of the ridgeline function, and $\bar{\alpha} = 1 - \alpha$. When $K > 2$ we will have a set of $\alpha_i's$ lying in a K-dimensional simplex, i.e. $\sum_{i=1}^{K}\alpha_i = 0$.

The image of the ridgeline function $x^*(\alpha)$ defines the curve from the mean of the first component to the mean of the second component. To illustrate the curve of the ridgeline function for a mixture of multivariate normal densities, we consider the following example.

Example 2.1. *(Two components, three modes, unequal variance) By considering the mixture of normal with $D = 2$, and $K = 2$ and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.8 & 0 \\ 0 & .07 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .07 & 0 \\ 0 & 0.8 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

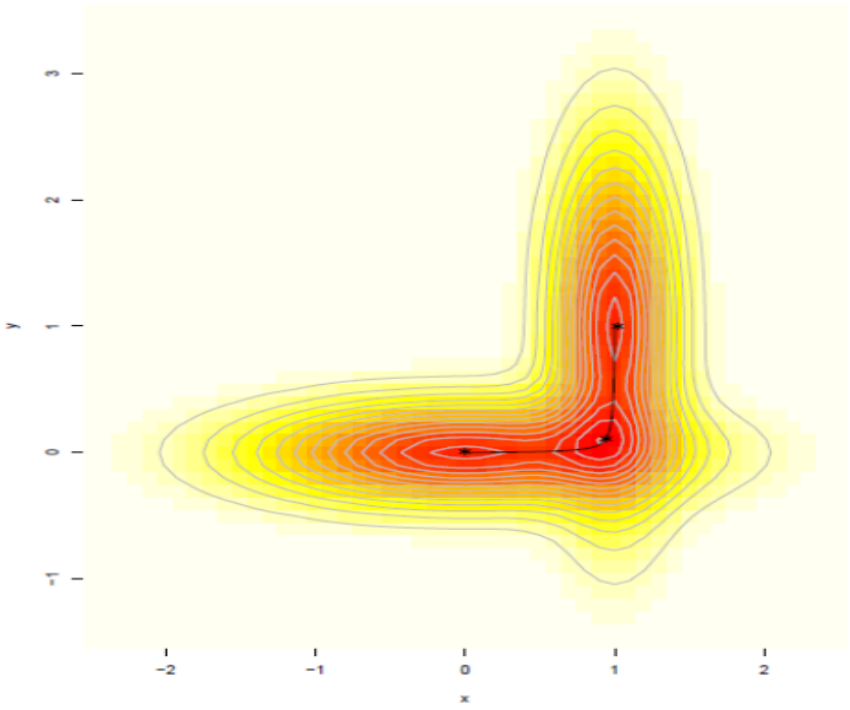


Figure 2.1: The contour plot and ridgeline curve of the mixture density given in the Example 2.1 showing the three modes on the ridgleline curve.

We now interpret Figure 2.1. The contours of the mixture density given in example 2.1 show that there are three modes, and the ridgeline function curve passes from the mean of the first component to the mean of the second component through all of the three modes and saddle-points of the density $g(\boldsymbol{x})$. The red colour denotes higher probability and the yellow lower probability, and the ridgeline is shown by the bold black line. The ridgeline function allows us to explore the modes of any D-dimensional normal by just exploring the ridgeline curve.

2.3.2 Height of ridgeline function

Now we know that the ridgleline passes through all the critical points, the next step to determine and explore the number and location of modes for a mixture of multivariate normal densities is by evaluating the height of the density along the curve and its critical

points. The ridgeline elevation is the height of the ridgeline function along the $x^*(\alpha)$ curve, the value of the ridgeline elevation function is defined by

$$h(\alpha) = \pi_1\phi_1(x^*(\alpha)) + \pi_2\phi_2(x^*(\alpha)), \qquad \alpha \in [0,1]. \tag{2.6}$$

The main advantage of using the ridgeline elevation plot is that it is a one dimensional curve but it allows us to determine the number and location of modes for a mixture of multivariate normal densities in any D dimensions, and are especially useful for $D > 2$ where the contour plot is not available.

To illustrate how to determine the number and location of the modes of the mixture of multivariate normal distribution by using the ridgeline elevation plot, we consider the two component mixture in example 2.1.

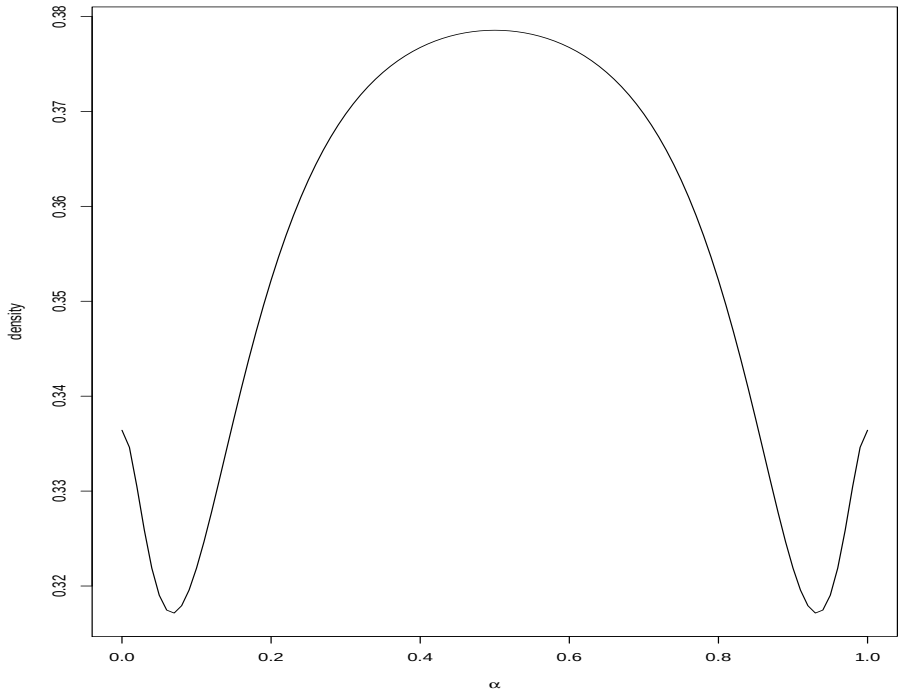


Figure 2.2: The ridgeline elevation plot of mixture density, given in example 2.1 showing three local modes, two at the ends ($\alpha = 0$ and 1) and one in the middle at $\alpha = 0.5$.

The ridgeline elevation plot given in Figure 2.2 shows that $h(\alpha)$ is one dimensional.

There are three local maxima, giving the three modes of the density in this plot. These modes are visible at α= 0, 0.5, and 1. The mode at $\alpha = 0$ corresponds to the mode at $x = (0, 0)$ in Figure 2.1, the mode at α=0.5 corresponds to the mode at the intersection of the two densities at $x = (0, 1)$ and finally the mode at α=1 corresponds to the mode at the other mean at $x = (1, 1)$.

Note that, for example 2.1 the elevation plot in Figure 2.2 just confirmed the three modes that were clearly visible from the contour plot in Figure 2.1. Now we move to a three dimensional example where we do not have the luxury to hunt for modes from the contour plot.

Example 2.2. *(Two components, four modes, unequal variance) Consider the mixture of normals with $D = 3$, and $K = 2$ and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0.6 \\ 2 \\ 0.6 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.05 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$\pi_1 = \pi_2 = 0.5$.

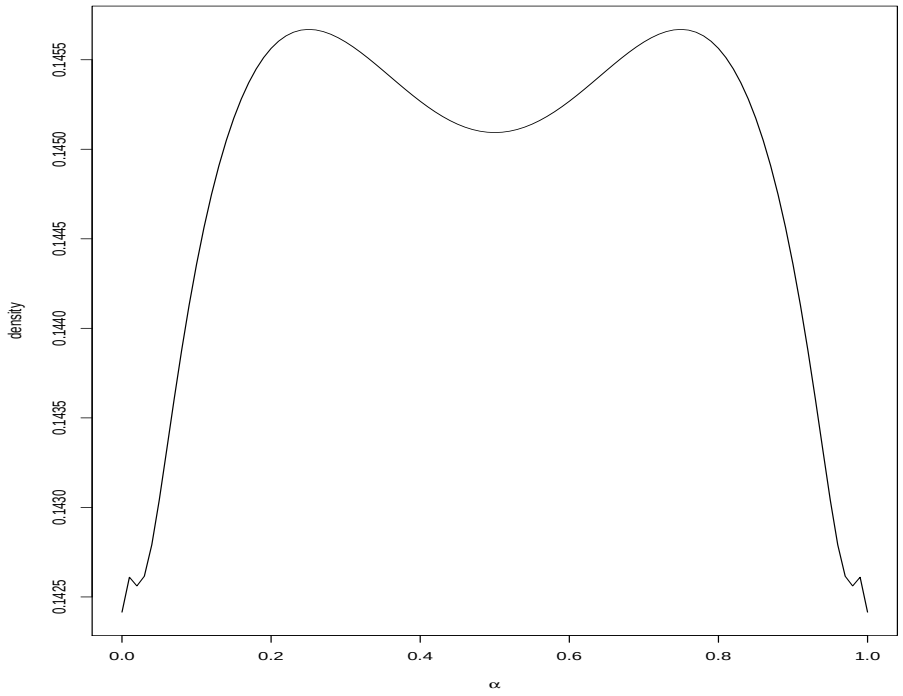


Figure 2.3: The ridgeline elevation plot of the mixture density given in Example 2.2 showing the local maximas corresponding to the four modes of the distribution.

Notice that though we are considering a 3 dimensional mixture the ridgeline elevation plot in Figure 2.3 shows that the elevation plot $h(\alpha)$ is still one dimensional. There are four local maxima giving four modes. These modes are visible near $\alpha = 0$, 0.2, 0.8, and 1.

The ridgeline functions and the ridgeline elevation plot together provide important

visual and analytical information regarding the location and number of modes of the multivariate normal mixture. A series of analytical results are given in Ray and Lindsay (2005) which provides the framework for determining the number of modes of normal mixtures by using the first and second derivations of the height of the ridgeline function. We will reproduce those results in the next section and later on generalize them to explore the modality of mixtures of multivariate t-densities in Chapter 3.

2.3.3 The curvature function

Though the ridgeline elevation plots provide a graphical answer to the number of modes which can be equated to the number of local maxima, an analytical answer is a necessary first step to embark on a study of the general modal features of mixtures. Ray and Lindsay (2005) explored the curvature function of the ridgleline elevation function and showed that the number of modes of the mixture is a function of the number of zero crossings of the curvature. To be precise, the number of modes is $[N/2+1]$ where N is the number of number of zero crossings of the curvature function. The curvature function is the second derivative of the ridgeline elevation, and for a two component mixture it is given by

$$q(\alpha) = \frac{\phi_2''(\alpha)}{\phi_2(\alpha)}\frac{\phi_1'(\alpha)}{\phi_1(\alpha)} - \frac{\phi_1''(\alpha)}{\phi_1(\alpha)}\frac{\phi_2'(\alpha)}{\phi_2(\alpha)}, \qquad \alpha \in [0,1], \tag{2.7}$$

where $\phi_i(\alpha) = \phi(x^*(\alpha); \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. For the density mixture of the multivariate normal density with two components, the curvature function in (2.7) can be simplified as

$$q(\alpha) = [p(\alpha)]^2[1 - \alpha\bar{\alpha}p(\alpha)], \tag{2.8}$$

where $\bar{\alpha} = (1-\alpha)$, $p(\alpha) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, and $S_\alpha = [\bar{\alpha}\boldsymbol{\Sigma}_1^{-1} + \alpha\boldsymbol{\Sigma}_2^{-1}]$.

As $p(\alpha)$ is positive, so the zeroes of $[1 - \alpha\bar{\alpha}p(\alpha)]$ and the zeroes of $q(\alpha)$ are same.

Now let us define

$$k(\alpha) = 1 - \alpha\bar{\alpha}p(\alpha) \tag{2.9}$$

whose zero crossing or roots will provide the crucial information on the number of modes in the mixture.

To illustrate how to find the number of modes of a mixture of multivariate normal densities by using $k(\alpha)$, we consider the mixture given in example 2.1, and explore its curvature function.

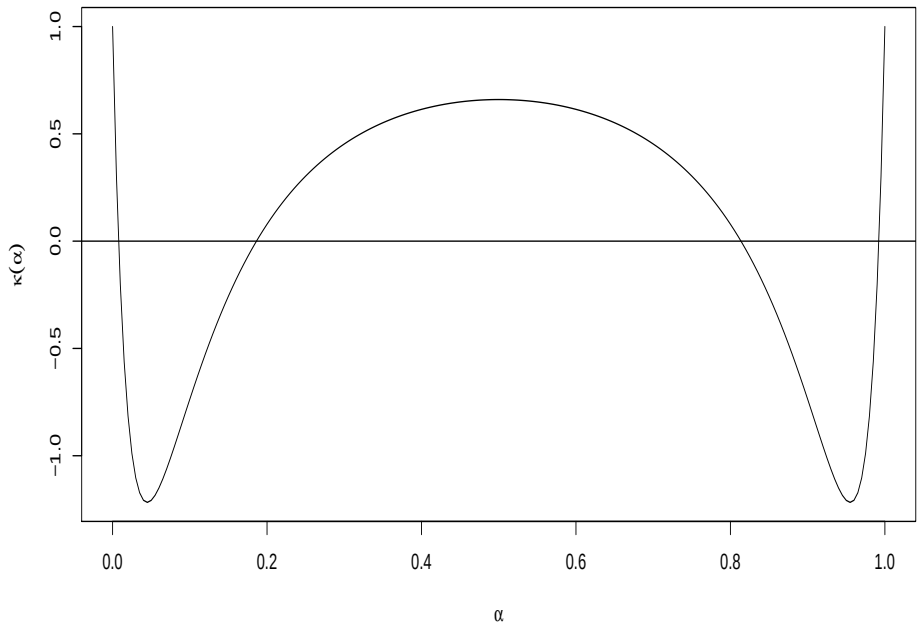


Figure 2.4: The curvature or $k(\alpha)$ plot of the mixture density given in Example 2.1 having three modes shows the four zero crossings.

The $k(\alpha)$ plot in Figure 2.4 shows that the $k(\alpha)$ curve crosses the zero line four times, and the total number of zeroes of $k(\alpha)$ is four. Now using the results that the number of mode is $N/2 + 1$, where N is the number of zeroes of the curvature function we can conclude that as the curvature function has 4 zeros the mixture has $4/2 + 1 = 3$

modes, which matches with the contour plot in Figure 2.1 and the elevation plot of Figure 2.2.

These powerful tools provide a rich framework for understanding the modality of the multivariate normal mixtures. Hennig (2010a) showed that the ridgeline function can be used for merging normal mixture components. The concept of the ridgeline function has been extended by Alexandrovich (2011) to explore the modality of a mixture of elliptical densities. These results also prompted researchers to start hypothesizing about fundamental results on the maximum number of the modes that one can get from mixing two normal components. The available results on this question were provided by Ray and Ren (2012). More results regarding this question will be presented in the next section. In this thesis we will generalize both the concepts of the ridgeline function and the elevation of the ridgeline function to explore the modality of the mixture of multivariate skew normal densities, areas of interest in this subject that have not yet been researches.

2.3.4 Upper bound of the number of modes of the multivariate normal mixture

Ray and Lindsay (2005) provided pathological examples showing the existence of 3 modes in 2 dimensions and 4 modes in 3 dimensions for normal mixtures of two components. But there was no general results relating the number of modes to the number of dimensions. Building on the rich framework of the ridgeline manifold and the behaviour of the curvature function, Ray and Ren (2012) later provided the important result on the number of modes of the multivariate normal mixture which states that the upperbound of D dimensional normal mixture of two components is $D + 1$ and the bound is tight. Here the word tight refers to the fact that the upper bound is achievable. Previously, Ray and Lindsay (2005) showed that the zeroes of the $k(\alpha)$

are the same zeroes as those of the $q(\alpha)$. Ray and Ren (2012) closely examined the $k(\alpha)$ by using a simplified class of the normal mixture where the first component was a standard normal but the second component had a diagonal covariance matrix. Ray and Ren (2012) showed that the upper bound of the number of modes of the mixture of multivariate normal densities could be computed using the number of the solutions to

$$k(\alpha) = 1 - \alpha(1-\alpha)p(\alpha) = 0, \tag{2.10}$$

where

$$p(\alpha) = \sum_{i=1}^{D} \frac{c_i}{[\alpha(\lambda_i - 1) + 1]^3}, \tag{2.11}$$

and $\alpha \in [0, 1]$, λ_i are the diagonal elements of $\boldsymbol{\Sigma}^*$, and $c_i = \lambda_i \boldsymbol{\mu}_i^2$. Where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is chosen to make one of the components identity variance and the other component to have a diagonal variance covariance matrix.

Finding the number of modes within the range $[0, 1]$ is a difficult task. Instead Ray and Ren (2012) used the transformation $\alpha = \frac{1}{s+1}$ to arrive at the function

$$k(s) = 1 - s(s+1) \sum_{i=1}^{D} \frac{c_i}{(s + \lambda_i)^3}, \tag{2.12}$$

where $s \in [0, \infty)$ corresponds to $\alpha \in [0, 1]$. Now, we need to find the number of positive roots of $k(s)$, which is a much easier task. $k(s)$ is a polynomial of degree $3D$ so it should have a maximum of $3D$ roots. Ray and Ren (2012) showed that D of the roots are negative, so we can get almost $2D$ positive roots of $k(s)$, which gives us $\frac{2D}{2} + 1 = D + 1$ modes. The details of the proof can be found in Ray and Ren (2012). In chapter 3 we will be using similar reasoning to explore the maximum number of modes of t-mixture.

2.4 Modality of the multivariate t-mixture

Now, we will move to non-normal mixtures. In this section we will give a brief background of the current research work on the modality of the multivariate t-mixtures. These results will be used in Chapter 3 to investigate if the maximum number of modes of the t-multivariate mixture is same as the maximum number of modes of the multivariate normal mixture.

2.4.1 Multivariate t-distribution

The multivariate t-distribution (or multivariate student's t-distribution) is defined as a generalization of the concept of univariate t-distributions. The multivariate t-distribution, and its properties has been introduced in many statistical books, such as McLachlan and Peel (2000). Note, there are more than one definition of multivariate t-distribution, but we will use the following definition throughout this dissertation.

Definition 2.2. *Suppose that $\boldsymbol{x}$ is a D-dimensional random vector of multivariate t-distribution with non-centrality parameter $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, and degrees of freedom n. The probability density function of multivariate t-distribution can be written as*

$$f(\boldsymbol{x}_i;\boldsymbol{\mu},\boldsymbol{\Sigma},n)=\frac{\Gamma(\frac{n+d}{2})\mid\boldsymbol{\Sigma}\mid^{-1/2}}{(\pi n)^{\frac{1}{2}D}\Gamma(\frac{n}{2})\{1+\frac{\delta(\boldsymbol{x},\boldsymbol{\mu};\boldsymbol{\Sigma})}{n}\}^{\frac{1}{2}(n+D)}},\qquad \boldsymbol{x}_i\in\mathbb{R}^D, \tag{2.13}$$

where $\delta(\boldsymbol{x}_i,\boldsymbol{\mu};\boldsymbol{\Sigma})=(\boldsymbol{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$ is the Mahalanobis squared distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$, and for easy of writing we will denote it by $\delta(\boldsymbol{x},\boldsymbol{\mu})$.

2.4.2 Mixture of multivariate t-distribution

Let K be the number of components of the mixture of multivariate t-densities. The probability density function of the mixture of multivariate t-densities can be given by

$$g(\boldsymbol{x}) = \sum_{i=1}^{K} \pi_i f(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, n_i), \quad \boldsymbol{x} \in \mathbb{R}^D, \tag{2.14}$$

where f is the t-density with location parameters $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Sigma}_i$, and n_i is the degree of freedom of multivaraite t-dsitrbtuion. π_i's are the mixing proportions with the constraint

$$0 < \pi_i \leq 1 \quad \forall i \qquad \text{and} \qquad \sum_{i=1}^{K} \pi_i = 1.$$

2.4.3 Upper bound of the number of modes of multivariate t-mixture

In this section we will present the current results on the modality of the mixture of multivariate t-densities. These results will then be used in Chapter 3 to explore the modality of the mixture of multivariate t-densities to find the upper bound of the number of modes. Though the ridgeline manifold results in Ray and Lindsay (2005) was initially derived to explore the modality of normal mixtures, Alexandrovich (2011) showed that the ridgeline function of the mixture of multivariate normal densities can indeed be used to explain the modes of the mixture of any elliptical densities. As t-distribution is a member of the elliptical family, Alexandrovich et al. (2013) explored the corresponding curvature function in equations (2.15) and (2.16) and showed that the upper bound of the number of modes of two t-component mixtures in D dimensions can be as high as $2D + 1$. However they failed to show that this bound is tight.

Let us now focus on the curvature function of the mixture of the multivariate t-distribution with two components, and sketch the proof for the upper bound of modes.

The curvature function of the t-mixture is given by

$$q_t(\alpha) = [p_t(\alpha)]^2[1 - \alpha\bar{\alpha}p_t(\alpha)f(n_1, n_2)], \tag{2.15}$$

where

$\bar{\alpha} = (1 - \alpha)$,

$p_t(\alpha) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$,

$f(n_1, n_2) = \left(\frac{n_2+D+2}{n_2}(1-\alpha)\left(1 + \frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1} + \frac{n_1+D+2}{n_1}\alpha\left(1 + \frac{\delta(x^*(\alpha),1)}{n_1}\right)^{-1}\right)$.

As $p_t(\alpha)$ is positive, the the zeroes of $q_t(\alpha)$ and the zeroes of $k_t(\alpha)$ are same, where the $k_t(\alpha)$ is given by

$$k_t(\alpha) = 1 - \alpha\bar{\alpha}p_t(\alpha)f(n_1, n_2). \tag{2.16}$$

Note that $k_t(\alpha)$ can also be written as

$$k_t(\alpha) = 1 - \alpha(1-\alpha)p_t(\alpha)\left(\frac{n_2 + D + 2}{n_2}(1-\alpha)\left(1 + \frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1} + \frac{n_1 + D + 2}{n_1}\alpha\left(1 + \frac{\delta(x^*(\alpha),1)}{n_1}\right)^{-1}\right), \tag{2.17}$$

where

$\delta(x^*(\alpha), 1) = (\boldsymbol{x} - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1)$,

$\delta(x^*(\alpha), 2) = (\boldsymbol{x} - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)$,

and n_i is the degrees of freedom of the mixture components where $i = 1, 2$.

It can be shown that $k_t(\alpha)$ has a total of almost $5D$ roots, $2D$ of which are coming from the new term $f(n_1, n_2)$. Using the same reasoning as the normal mixture it can be shown that D of them are negative. Thus at most $4D$ roots can be positive which implies that the maximum number of modes can be $4D/2 + 1 = 2D + 1$. The details of the proof can be found in Alexandrovich et al. (2013).

In the next chapter we will investigate the maximum number of modes for the mix-

ture of multivariate t-densities graphically, and analytically by focusing on the equation for the curvature function $k_t(\alpha)$. The goal for this is to check if the upper bound of the number of modes for the mixture of multivariate t-densities is the same as the upper bound of the number of modes for the mixture of multivariate normal densities.

Chapter 3

Modality of Multivariate t-Mixture

In this chapter we shall explore the modality of a mixture of multivariate t-densities. In the previous chapter we have stated that mixture models are a tool that can be used for modelling datasets arising from populations with subgroups. Finite mixtures show flexibility in modelling complex distributions, and heterogeneous populations, as well as performing cluster analyses McLachlan and Peel (2000). Normal mixture models are generally the most commonly used tools for modelling continuous multivariate data, but for continuous multivariate data with relatively large extreme values and fatter tails using the mixture t-densities provides a more appropriate alternative than the mixture of normal densities (Peel and McLachlan, 2000). In this chapter we will generalize the framework for ridgeline analysis introduced in the previous chapter to understand the modality of t-mixtures. We first provide a short description of how researchers have used the ridgeline analysis framework in recent years and developed algorithms and theoretical results.

The ridgeline function was used as a tool for merging components of the normal mixture to produce better cluster Hennig (2010a). The results of Ray and Lindsay (2005) in understanding the structures of the mixture of multivariate normal densities have also been extended to explore the structures of other densities. The concept of

the ridgeline function of multivariate normal densities mixtures has been extended to explore the modality of finite mixtures of elliptical densities by Alexandrovich et al. (2013). Later, the ridgeline function of the multivariate normal mixture has also been generalized to explore the modality of multivariate logistic mixtures (Liu and Unlü, 2015). Ray and Ren (2012) explored the modality of the mixture of the multivariate distribution in depth and showed that the upper bound of the number of modes that can be achieved from mixing two normal components with any D-dimensional is $D+1$ modes. Alexandrovich et al. (2013) explored the modality of the mixture of multivariate t-densities and came up with the upper bound of $2D+1$ modes from mixing two t-components in D-dimensions.

In this chapter we start with the conjecture that the upper bound of the number of modes for the mixture of two multivariate t-densities is the same as the upper bound of the number of modes for the mixture of two multivariate normal densities which is $D+1$. Note that Alexandrovich et al. (2013) was not able to prove the upper bound of $2D+1$ was tight. Here the word tight refers to the fact that the upper bound is achievable.

3.1 Upper bound of the number of modes of the mixture of multivariate t-distribution

In this section we will investigate the maximum number of modes that can be achieved from mixing two t-components and shall check if the upper bound of the number of modes for the mixture of two multivariate t-densities is the same as the upper bound of the number of modes for the mixture of two multivariate normal densities. The strongest results relating to the upper bound of the number of modes when mixing just two multivariate normal components were given by Ray and Ren (2012). They proved

that it was possible to get as many as $D+1$ modes from just two normal mixtures in any D-Dimension. Furthermore they supported their results by providing a recursive algorithm to construct the mean and variance parameters of the component densities, thus providing a tight upper bound. The available results for the maximum number of modes for the mixture of multivariate distribution were given by Alexandrovich et al. (2013). They show that after mixing two components of t-densities the maximum number of modes that can achieved in D-dimensions is $2D+1$ modes. However, we conjecture that the mixture of two multivariate t-densities has same maximum number of modes as the mixture of two multivariate normal densities.

To investigate the number of modes for the mixture of two multivariate t-densities, we first restate the probability density function of a two component t-mixture, which is a special case of (2.14) and is given by

$$g(\boldsymbol{x}) = \pi_1 f(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, n_1) + \pi_2 f(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, n_2), \qquad \boldsymbol{x} \in \mathbb{R}^D, \tag{3.1}$$

where n_1 and n_2 are the degrees of freedom of multivariate t-mixture.

Alexandrovich et al. (2013) showed that the ridgeline function of the multivariate normal mixture can be used for the mixture of multivariate t-densities. The ridgeline function of the two-component mixture of t-densities can be given by

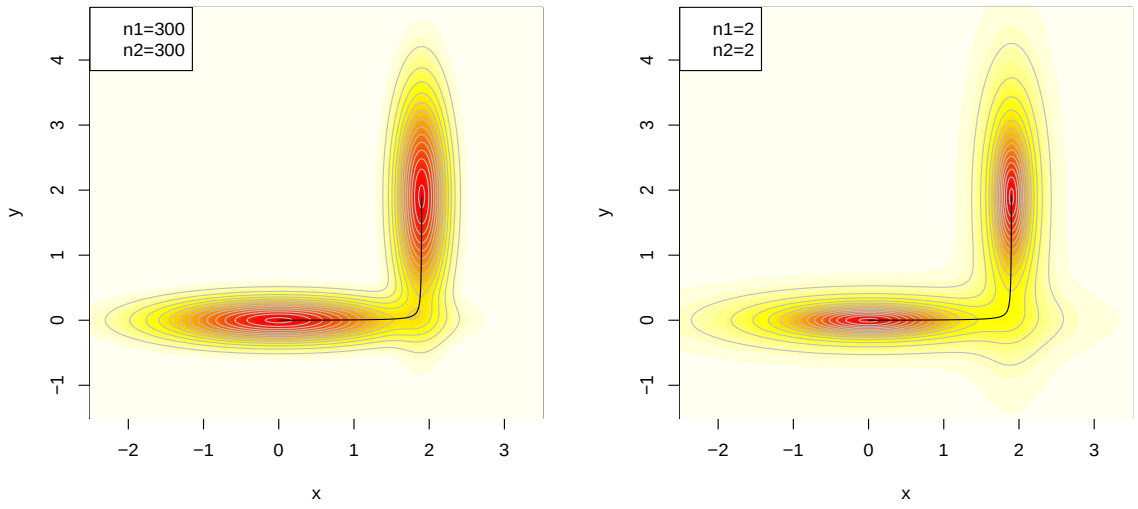
$$x^*(\alpha) = [\bar{\alpha}_1 \boldsymbol{\Sigma}_1^{-1} + \alpha_2 \boldsymbol{\Sigma}_2^{-1}]^{-1} [\bar{\alpha}_1 \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 + \alpha_2 \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2],$$

where α is defined as the range of the ridgeline function $\alpha \in [0,1]$, and $\bar{\alpha} = 1 - \alpha$.

The next example shows the ridgeline curve of the normal mixture can be used to explore the modes of the mixture of multivariate t-distribution.

Example 3.1. *Two components, three modes, and unequal variance, as a result of considering the mixture of t-densities with $D = 2$, $K = 2$ and the following parameters.*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 1.9 \\ 1.9 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



(a) (b)

Figure 3.1: Contour plot and ridgeline curve of the mixture density given in Example 3.1.

Figure 3.1 shows the contour plots of the mixture density, given in example 3.1 but with two different sets of degrees of freedom. Figure 3.1 (a) shows the contour plot of the t-mixture with $n_1 = n_2 = 300$ which incurs three modes. We can observe that the curve of the ridgeline function starts from the mode of the first component and ends at the mode of the second component and passes through the third mode and saddle-points of the t-mixture density. Figure 3.1 (b) shows the contour plot of the t-mixture with $n_1 = n_2 = 2$ which incurs two modes. The curve of the ridgeline function again passes from the mode of the first component to the mode of the second component through all the saddle-points of the t-mixture density. We notice that the number of modes changed from three to two with the change of degrees of freedom. But in both

cases the ridgeline curve remains the same and passes through all the critical points.

In particular the number and location of modes will depend on the height of the density along the same ridgleine curve. We will explore the number of modes of the mixture of two multivariate t-densities by the height function and its resulting curvature function. In the next section we will provide an extensive graphical analyses of the height along the ridgeline curve for different t-mixtures with the same mean and variance but varying degrees of freedom.

3.2 Graphical investigation into the number of modes

In this section we shall examine graphically the number of modes of the mixture of two multivariate t-densities by analyzing the behaviour of the curvature function of the multivariate t-mixture. The curvature function $k_t(\alpha)$ is defined as the second derivative of the ridgeline elevation function of the mixture of multivariate t-distribution (Alexandrovich et al., 2013). The curvature function $k_t(\alpha)$ of the two-component t-mixture can be written as

$$k_t(\alpha) = 1 - \alpha(1-\alpha)p(\alpha)\left(\frac{n_2+D+2}{n_2}(1-\alpha)\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1} + \frac{n_1+D+2}{n_1}\alpha\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)^{-1}\right) = 0, \quad (3.2)$$

where α is the range of the ridgeline function,

$p(\alpha) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} S_\alpha^{-1} \boldsymbol{\Sigma}_2^{-1} S_\alpha^{-1} \boldsymbol{\Sigma}_2^{-1} S_\alpha^{-1} \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$,

$S_\alpha = [\bar{\alpha}\boldsymbol{\Sigma}_1^{-1} + \alpha\boldsymbol{\Sigma}_2^{-1}]$,

$\delta(x^*(\alpha),1) = (\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1)$,

$\delta(x^*(\alpha),2) = (\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2)$,

and n_i is the degrees of freedom of the mixture components where $i = 1, 2$.

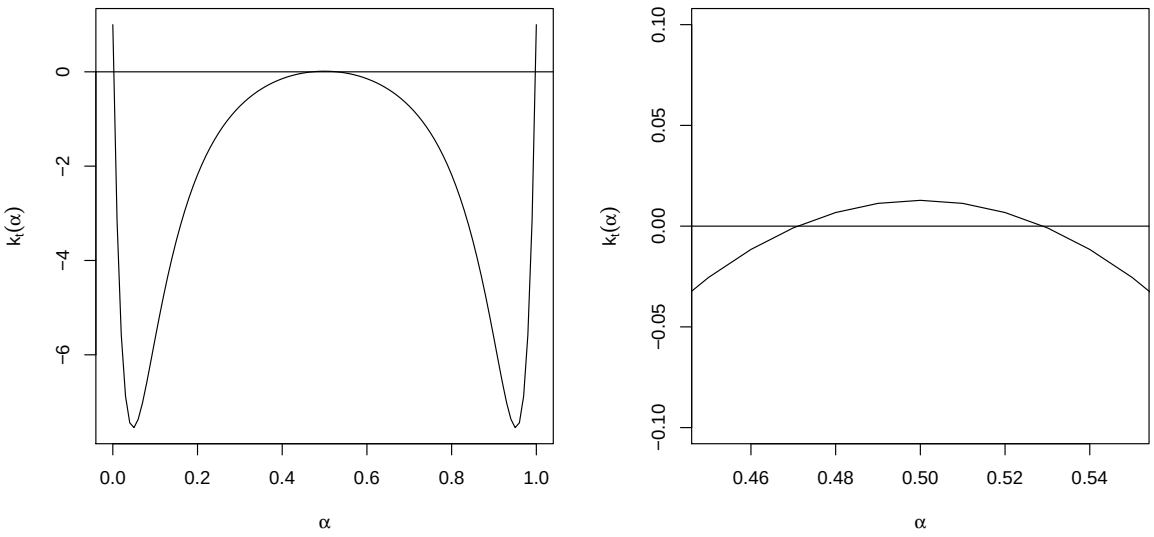
We will first explore the effect of degrees of freedom on the number of modes of multivariate t-mixtures of two components in two dimensions.

3.2.1 How degrees of freedom affects the number of modes in two D-dimensions

In this section we study the mixture of several two-component t-mixtures to understand the change in the number of modes when the degrees of freedom of the t-distribution changes. To isolate the effect of degrees of freedom we fix the mean and variance parameters of these two components for all the examples. We start with four examples each of which are two component mixtures of multivariate t-densities in two dimensions.

Example 3.2. *(Two components, unequal variance, three modes) Consider a two-component mixture of t-densities with $D = 2$, and degrees of freedom $n_1 = 30$ and $n_2 = 30$, and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2.45 \\ 2.45 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



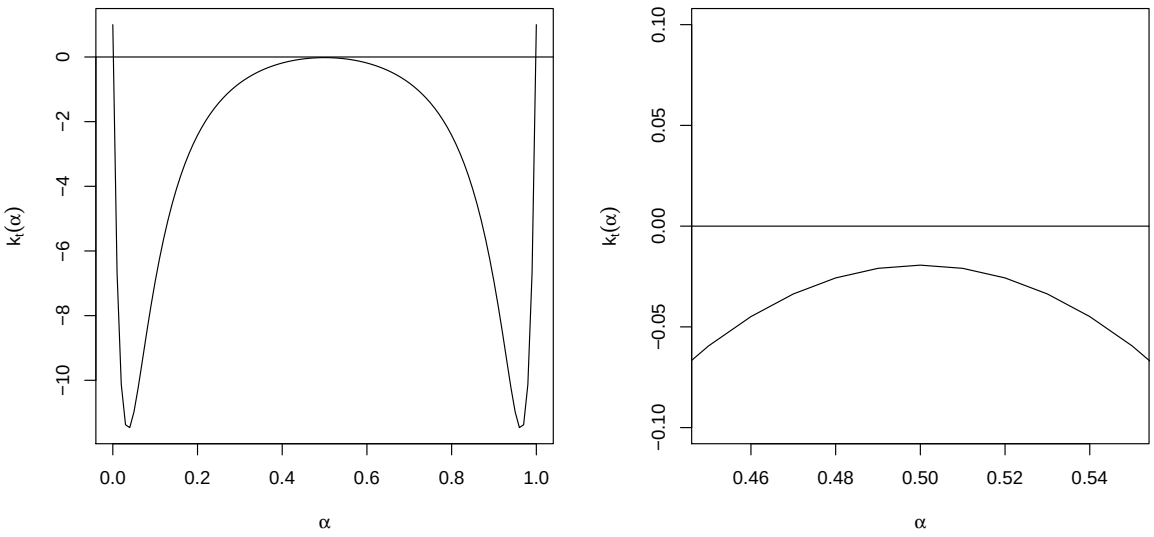
(a) (b)

Figure 3.2: The curvature function plot for the mixture density given in Example 3.2 where $n_1 = 30$, and $n_2 = 30$. (a) is the plot of the curvature function, and (b) is the zoomed version of (a) to check if the peak of the curve crosses the zero line.

Figure 3.2 shows the curvature function of the mixture density given in example 3.2. The left panel shows that the curve of the curvature function $k_t(\alpha)$ definitely crosses the zero line two times, but it is not entirely clear if the middle peak crosses the zero line. The right panel is a zoomed-in version of a subset of the entire range, which shows the curvature function $k_t(\alpha)$ crossing the zero-line two more times. Recall, that the total number of modes is $N/2 + 1$, where N is the number of zero crossing of $k_t(\alpha)$. The total number of modes for example 3.2 is therefore three.

Example 3.3. *(Two components, unequal variance, two modes) Now consider a two-component mixture of t-densities with $D = 2$, and degrees of freedom $n_1 = 95$ and $n_2 = 95$, and following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2.45 \\ 2.45 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



(a) (b)

Figure 3.3: The curvature function plot for the mixture density, given in Example 3.3, where $n_1 = 95$ and $n_2 = 95$. (a) is the plot of the curvature function, and (b) is the zoomed-in version of (a) to check if the peak of the curve crosses the zero line.

In this example, the mean, variance and proportion parameters have not been altered from those in example 3.2, but we have increased the degrees of freedom for both components of the t-mixture to $n_1 = n_2 = 95$ to assess how this affects the number of modes. Similar to the previous example the left panel again shows that the curvature function $k_t(\alpha)$ is crossing the zero-line once near 0 and again near 1. But unlike the previous example the zoomed-in plot (b) now reveals that the $k_t(\alpha)$ curve is not crossing the zero-line near the middle peak at 0.5. Thus the analysis of the curvature function for this example gives us only two modes instead of the three modes in example 3.2.

Notice that the increase in the degrees of freedom decreases the number of modes in example 3.2, whereas it increases the number of modes in example 3.1.

Example 3.4. *(Two components, unequal variance, three modes) Consider a two-component mixture of t-densities with $D = 2$, with different degrees of freedoms $n_1 = 30$ and $n_2 = 3$ and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2.45 \\ 2.45 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

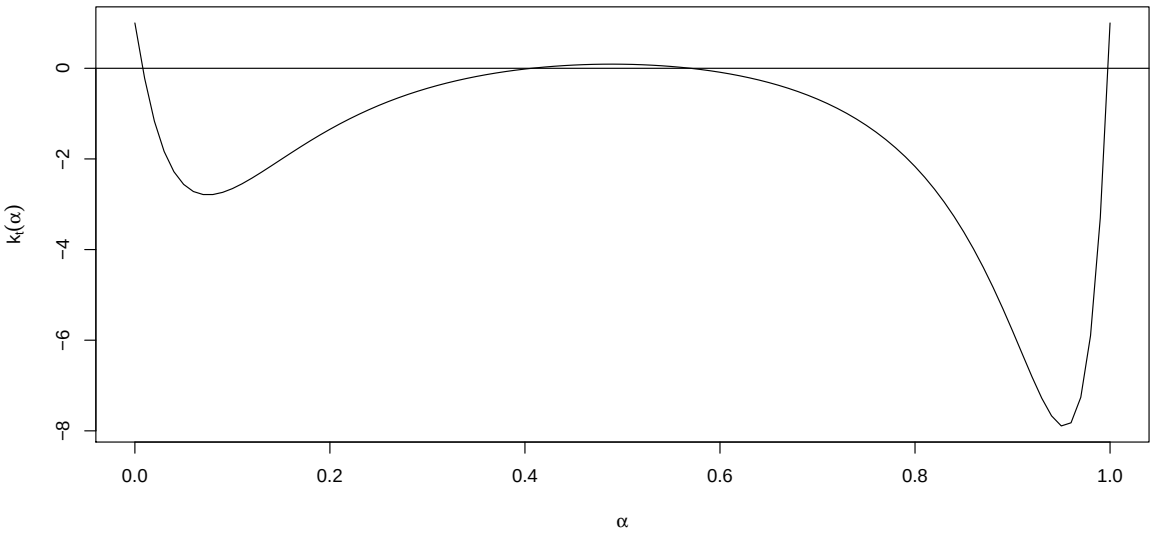


Figure 3.4: The curvature function plot for the mixture density, given in Example 3.4, where $n_1 = 30$ and $n_2 = 3$.

Figure 3.4 shows the plot of the curvature function with unbalanced degrees of freedom for the two components. The plot shows that when the degree of freedom is increasing for the first component but decreasing for the second component the total number of modes is three. Notice that the unbalanced degrees of freedom produces an asymmetric $k_t(\alpha)$ curve.

Example 3.5. *(Two components, unequal variance, three modes) Consider a two-component mixture of t-densities with $D = 2$, with different degrees of freedoms $n_1 = 30$ and $n_2 = 95$ and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 2.45 \\ 2.45 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

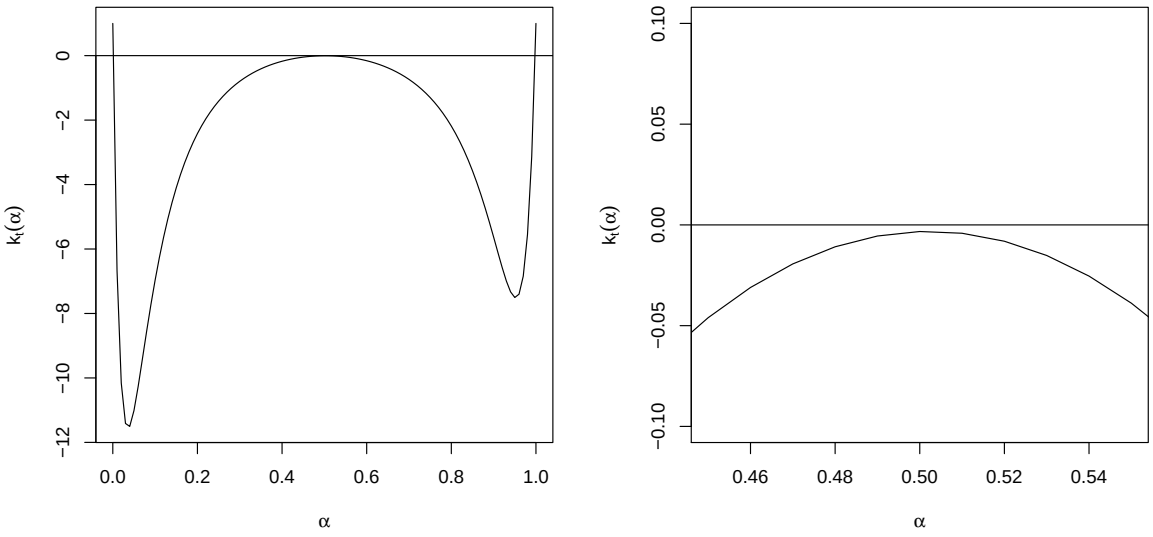


Figure 3.5: The curvature function plot for the mixture density, given in Example 3.5, where $n_1 = 30$ and $n_2 = 95$.

In this example, the mean, variance and proportion parameters have not been altered from those in example 3.2, but we have used unbalanced degrees of freedom for both components of the t-mixture where $n_1 = 30$ and $n_2 = 95$ to assess how this affects the number of modes. The left panel again shows that the curvature function $k_t(\alpha)$ is crossing the zero-line once near 0 and again near 1. But unlike the example 3.2 the zoomed-in plot now reveals that the $k_t(\alpha)$ curve is not crossing the zero-line near the middle peak at 0.5. Thus the analysis of the curvature function for this example gives us only two modes.

These examples help us to understand the effect of the degree of freedom on the number of modes of the mixture of two t-components and two dimensions. We observed from these examples that the degrees of freedom do indeed affect the number of modes of the mixture of multivariate t-densities. When we increased the degrees of freedom we lost one mode. In fact, based on this trend and the fact that normal is a limiting case of t when the degrees of freedom goes to infinity we can easily see that we can construct examples where we can create a new mode by changing the degrees of freedom.

For more investigations into the effect of the degrees of freedom on the number of modes for the mixture of two multivariate t-densities, we will now explore the number of modes for the mixture of multivariate t-densities in the case of using two components and three dimensions. One of the reason we are doing this investigation is to see if we can find examples where the degrees of freedom of the t-mixture can create extra modes compared to a normal mixture with same mean and variance. If we can create extra modes building a two component normal mixture which has already attained its upper bound of $D + 1$ modes in D dimension, we would be able to contradict our conjecture that the upper bound of the t-mixture is also $D + 1$.

3.2.2 How degrees of freedom affects the number of modes in three dimensions

In this section we shall investigate the maximum number of modes that can be obtained from the two-component mixture in three dimensions. Similar to the two dimensional case, we again fix the mean, variance and mixing proportions and only vary the degrees of freedom to understand its effect on the number of components.

Example 3.6. *(Two components, unequal variance) First consider, a set of eight two-component mixtures of t-densities with $D = 3$, differing by the degrees of freedom and but with same mean and variances and mixing proportions given by*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0.6 \\ 2 \\ 0.6 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

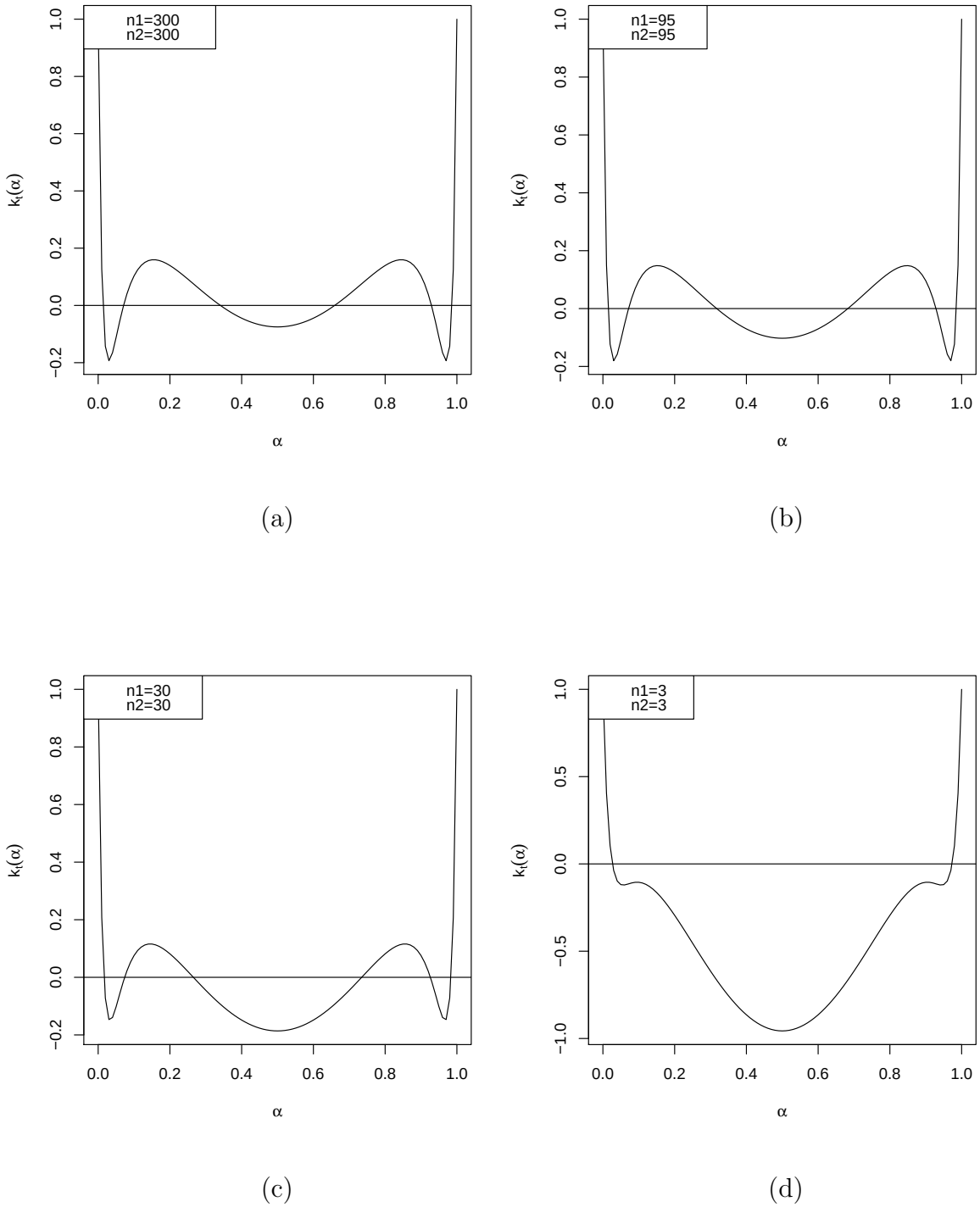
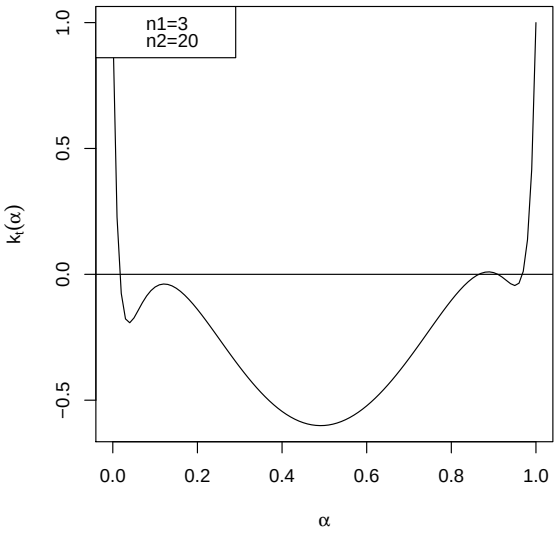
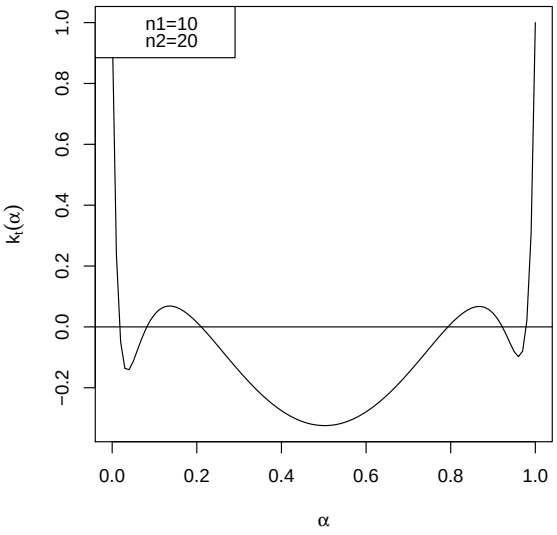


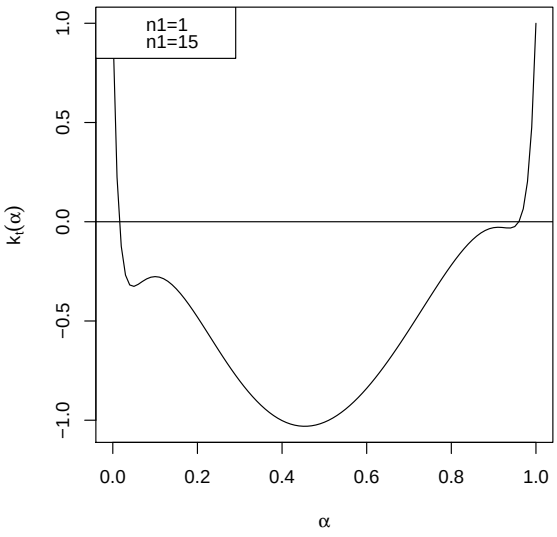
Figure 3.6: The curvature function plots for the mixture density, given in Example 3.7, with balanced degrees of freedom. These plots show that there are either 2 or 6 crossings with the zero line resulting in either 2 or 4 modes.



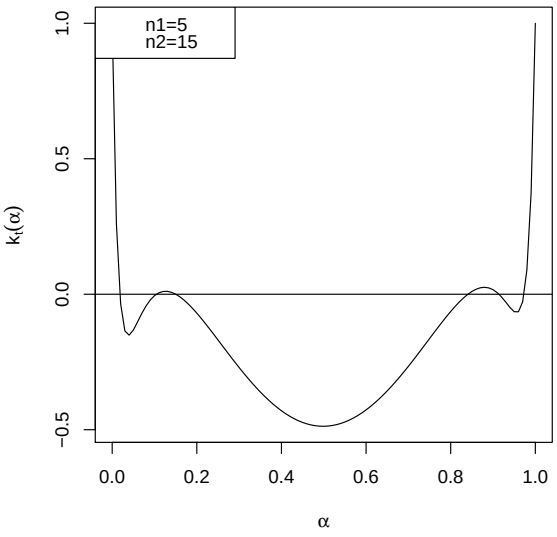
(e)



(f)



(g)



(h)

Figure 3.7: The curvature function plots for the mixture density, given in Example 3.7, with unbalanced degrees of freedom. These plots show that there are either 2, 4 or 6 crossings with the zero line resulting in either 2,3 or 4 modes.

The eight plots in figures 3.6 and 3.7 shows the plots of the curvature function for the parameters given in example 3.7, with balanced and unbalanced degrees of freedoms. The first thing we notice is that when the two components have different degrees of freedom the $k_t(\alpha)$ function is asymmetric. More importantly we notice that, with the same mean and variance and mixing proportions parameters the number of modes of the t-mixture changes considerably with the change of the degrees of freedom of the two components.

For example, in sub-plot (a) when the degrees of freedom of the two components are 300 we can conclude that there are 6 zero crossings which implies $6/2 + 1 = 4$ modes. When the degrees of freedom are both 95 we again see 6 zero crossings in sub-plot (b). But when the degrees of freedom of the two components are 3, which is represented in sub-plot (d) we can only observe 2 zero crossings which implies that the mixture has only two modes. In sub-plot (e), where the degrees of freedom are 3 and 20, we have even demonstrated an example, with 4 zero crossings or 3 modes.

In summary, The total number of modes that can be obtained from using different degrees of freedom can range from two to four. Finally, the maximum number of modes that we achieved in the case of using two t-components and three dimensions was four.

3.2.3 Systematic study of change in modes with change in degrees of freedom

We now provide a systematic study to explore the relationship between the number of mode and the degrees of freedom in a two component t-mixture.

Example 3.7. *(Two components, unequal variance) Consider a two-component mixtures of t-densities with $D = 3$, with degrees of freedoms ranging from 1 to 20 and the following parameters:*

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\mu}_2 = \begin{pmatrix} 0.6 \\ 2 \\ 0.6 \end{pmatrix}, \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & .05 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} .05 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

Table 3.1 summarises the number of modes for changing degrees of freedom for both components, while keeping the means, variances and mixing proportions same. For these set of parameters we notice that when both components have high degrees of freedom we get 4 modes, which is same as the number of modes for a normal mixture with similar parameters. We also observe that for very low degrees of freedom we only get 2 modes. There are some instance for moderate degrees of freedom where we get 3 modes, but notice that odd number of modes is only possible for the above parameter setting when we have unequal degrees of freedom. In other words, if the degrees of freedom are same we jump from 2 modes (at 6 degrees of freedom) to 4 modes (at 7 degrees of freedom).

Table 3.1: Number of modes for the mixture density, given in Example 3.7, for degrees of freedom $n_1 = 1, 2, ..., 20$ and $n_2 = 1, 2, ..., 20$.

n2 \ n1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
4	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
5	2	2	2	2	2	2	2	3	3	3	3	4	4	4	4	4	4	4	4	4
6	2	2	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4
7	2	2	2	2	2	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
8	2	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
9	2	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
10	2	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
11	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
12	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
13	2	2	2	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
14	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
15	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
16	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
17	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
18	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
19	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
20	2	2	3	3	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4

3.2.4 Conclusions from graphical studies of the effect of degrees of freedom

By exploring the effect of the degrees of freedom on the number of modes for the mixture of multivariate t-distribution in the case of using two components in two dimensions we have observed that the degrees of freedom plays a vital role in determining the number of modes that the t-mixtures incur. We were able to increase the number of modes by changing the degrees of freedom, but only in the situation where the t-mixtures have not attained the $D+1$ limit which is true for normal mixture. Once we start with a mixture with $D+1$ modes in D dimension we were not able to increase the number of modes any further which supports our conjecture that the upper bound of normal and t-mixtures are same, and that it is a tight upper-bound. Here the word tight refers to the fact that the upper bound is achievable. Thus we were unable to create more than $D+1$ modes in two and three dimensions and did not find any evidence to contradict our conjecture.

To undertake further investigation into understanding the number of modes for the mixture of multivariate t-distribution, we would need to explore the curvature function $k_t(\alpha)$ given in 3.3 analytically.

3.3 An analytical investigation into the number of modes

In the previous section we explored the curvature function $k_t(\alpha)$ graphically and could not find any evidence to disprove our conjecture that the number of modes for the mixture of two multivariate t-densities was the same as the number of modes for the mixture of two multivariate normal densities. In this section we shall explore the curvature function $k_t(\alpha)$ in depth in order to provide more support for our conjecture

that the upper bound of the number of modes for the mixture of two multivariate t-densities is same as the upper bound of the number of modes for the mixture of two multivariate normal densities.

Recall, the curvature function of the two-component t-mixture was:

$$k_t(\alpha) = 1 - \alpha(1-\alpha)p(\alpha)\left(\frac{n_2+D+2}{n_2}(1-\alpha)\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1} + \frac{n_1+D+2}{n_1}\alpha\left(1+\frac{\delta(x^*(\alpha),1}{n_1}\right)^{-1}\right)$$

We rewrite the equation in the following form to isolate the effect of the degrees of freedom:

$$k_t(\alpha) = 1 - \alpha(1-\alpha)p(\alpha)f(n_1,n_2) \tag{3.3}$$

where

$p(\alpha) = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_2^{-1}S_\alpha^{-1}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$,

$f(n_1,n_2) = \left(\frac{n_2+D+2}{n_2}(1-\alpha)\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1} + \frac{n_1+D+2}{n_1}\alpha\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)^{-1}\right)$,

$S_\alpha = [\bar{\alpha}\boldsymbol{\Sigma}_1^{-1} + \alpha\boldsymbol{\Sigma}_2^{-1}]$,

$\delta(x^*(\alpha),1) = (\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)$,

$\delta(x^*(\alpha),2) = (\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)$,

n_1 and n_2 are the degrees of freedom of a two component t-mixtures.

First note that the only term that contatins the degrees of freedom parameters is the $f(n_1,n_2)$. So to study the effect of degrees freedom on the behaviour of the $k_t(\alpha)$ function we can focus on the function $f(n_1,n_2)$. Intuitively a normal is a limiting distribution to a t-distribution when the degrees of freedom goes to infinity. The next lemma shows that the curvature function of normal and t-densities with the same mean and variance are same when the degrees of freedom of the two components, n_1 and n_2 $\to \infty$.

Lemma 3.1. *If the degrees of freedom of the two mixture components $n_1 \to \infty$ and $n_2 \to \infty$ then we have $f(n_1,n_2) = 1$.*

Proof.

$$f(n_1,n_2)=\left(\frac{n_2+D+2}{n_2}(1-\alpha)\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)^{-1}+\frac{n_1+D+2}{n_1}\alpha\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)^{-1}\right). \tag{3.4}$$

$$\begin{aligned}
f(n_1,n_2) &= \left(\frac{\frac{(n_2+D+2)(1-\alpha)}{n_2}}{\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)}+\frac{\frac{(n_1+D+2)\alpha}{n_1}}{\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)}\right)\\
&= \left(\frac{\frac{n_2+D+2-n_2\alpha-D\alpha-2\alpha}{n_2}}{\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)}+\frac{\frac{(n_1+D+2)\alpha}{n_1}}{\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)}\right)\\
&= \left(\frac{\frac{n_2}{n_2}+\frac{D}{n_2}+\frac{2}{n_2}-\frac{n_2\alpha}{n_2}-\frac{D\alpha}{n_2}-\frac{2\alpha}{n_2}}{\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)}+\frac{\frac{n_1\alpha}{n_1}+\frac{D\alpha}{n_1}+\frac{2\alpha}{n_1}}{\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)}\right)\\
&= \left(\frac{1+\frac{D}{n_2}+\frac{2}{n_2}-\alpha-\frac{D\alpha}{n_2}-\frac{2\alpha}{n_2}}{\left(1+\frac{\delta(x^*(\alpha),2)}{n_2}\right)}+\frac{\alpha+\frac{D\alpha}{n_1}+\frac{2\alpha}{n_1}}{\left(1+\frac{\delta(x^*(\alpha),1)}{n_1}\right)}\right).
\end{aligned}$$

As D and α are constants and $\delta(x^*(\alpha),1)$ and $\delta(x^*(\alpha),2)$ are a bounded functions we

have

$$\begin{aligned}\lim_{n_1\to\infty,n_2\to\infty} f(n_1,n_2) &= \left(\frac{1+0+0-\alpha-0-0}{(1+0)}+\frac{\alpha+0+0}{(1+0)}\right)\\ &= \left(\frac{1-\alpha}{(1+0)}+\frac{\alpha}{(1+0)}\right)\\ &= 1.\end{aligned}$$

□

Corollary 3.1. *The $k_t(\alpha)$ function for t-mixture is exactly same the $k(\alpha)$ for normal mixture when $n_1 \longrightarrow \infty$ and $n_2 \longrightarrow \infty$.*

Proof. Substituting 1 in place of $f(n_1, n_2)$ in equation (3.3) gives us the curvature function of the normal mixture. □

Exploring the modality of the arbitrary multivariate t-mixture is based on the curvature function of the mixture of multivariate t-densities and is a very difficult undertaking. Previously Alexandrovich et al. (2013) used a transformation similar to Ray and Ren (2012) to examine the modes of an arbitrary mixture of multivariate t-distribution. Alexandrovich et al. (2013) showed that the number of modes for the mixture of the multivariate t-distribution $g(\boldsymbol{x};\pi,\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,n_1,n_2)$ is the same as the number of modes for $g(\boldsymbol{x};\pi,\boldsymbol{\mu}^*,0,\boldsymbol{\Sigma}^*,I_D,n_1,n_2)$, where $\boldsymbol{\mu}^*$ and $\boldsymbol{\Sigma}^*$ is chosen to make one of the components identity variance and the other component to have a diagonal variance covariance matrix. Alexandrovich et al. (2013) rewrote the equation of the curvature function $k(\alpha)$ 3.3 as follows:

$$\delta(x^*(\alpha),1) = \alpha^2 \sum_{i=1}^{D} \frac{c_i^2\lambda_i^2}{(1+(\lambda_i-1)\alpha)^2}.$$

$$\delta(x^*(\alpha), 2) = (1-\alpha)^2 \sum_{i=1}^{D} \frac{c_i^2 \lambda_i^2}{(1+(\lambda_i-1)\alpha)^2}.$$

$$p(\alpha) = \sum_{i=1}^{D} \frac{c_i^2 \lambda_i^2}{(\alpha(\lambda_i-1)+1)^3},$$

where c_i's are functions of the mean and λ_i's are the eigen values of $\mathbf{\Sigma}^*$.

By using Lemma 3 in Ray and Ren (2012), the number of modes for $\alpha \in [0,1]$ is exactly same as the number of non-negative solutions $s \in [0,\infty)$, where $\alpha = \dfrac{1}{1+s}$ and the number of modes of $s \in [0,\infty)$ corresponds to $\alpha \in [0,1]$. Alexandrovich et al. (2013) showed that the upper bound of the number of modes for the mixture of the multivariate t-densities with two components, in D-dimensions can be given by the following equation:

$$k(s) = 1 - s(s+1) \sum_{i=1}^{D} \frac{\boldsymbol{\mu}_i^2 \lambda_i^2}{(s+\lambda_i)^3} \left(\frac{(n_2+D+2)s}{n_2 \left(1 + \frac{s^2}{n_2} \sum_{i=1}^{D} \frac{\boldsymbol{\mu}_i^2 \lambda_i^2}{(s+\lambda_i)^2}\right)} + \frac{(n_1+D+2)}{n_1 \left(1 + \frac{1}{n_1} \sum_{i=1}^{D} \frac{\boldsymbol{\mu}_i^2 \lambda_i}{(s+\lambda_i)^2}\right)} \right). \quad (3.5)$$

For simplification we write 3.5 in the following form

$$k(s) = 1 - s(s+1) \sum_{i=1}^{D} \frac{\boldsymbol{\mu}_i^2 \lambda_i^2}{(s+\lambda_i)^3} f(s). \quad (3.6)$$

where

$$f(s) = \left(\frac{(n_2+D+2)s}{n_2 \left(1 + \dfrac{s^2}{n_2} \sum_{i=1}^{D} \dfrac{\boldsymbol{\mu}_i^2 \lambda_i^2}{(s+\lambda_i)^2}\right)} + \frac{(n_1+D+2)}{n_1 \left(1 + \dfrac{1}{n_1} \sum_{i=1}^{D} \dfrac{\boldsymbol{\mu}_i^2 \lambda_i}{(s+\lambda_i)^2}\right)} \right). \quad (3.7)$$

By comparing equation 3.6 for finding the number of modes for the mixture of two multivariate normal densities with equation 2.12 for finding the number of modes for the mixture of two multivariate normal densities, we can see that equation 3.6 reduced to equation 2.12 when $f(s) = 1$, and the number of modes is the same as the normal.

3.4 Conclusion

In this chapter we have explored the upper bound for the number of modes for the mixture of two multivariate t-densities in two ways, namely, graphically and analytically. Graphically, we have explored the number of modes for the mixture of two t-components in the case of using of two and three dimensions. We have given examples to show the effect of changing the degrees of freedom for fixed component parameters on the number of modes, and the maximum number of modes that has been found for the two t-component mixture with two dimensions is three. In this chapter we also explored the number of modes for high dimensions, for the mixture of two t-components, and for three dimensions in order to check the effect of changing the degrees of freedom for fixed parameters of two components. We found that the degrees of freedom do indeed affect the number of modes for the mixture of t-components. The maximum number of modes that we achieved for two t-components in three dimensions was four. In these dimensions we also observed that when the two components have the same degrees of freedom we could get only two or four modes, but when the two degrees are different we can get two, three or four modes. Graphically, the maximum number of modes that we activated by exploring the curvature function $k(\alpha)$ of the mixture of multivariate t-distribution was four.

In order to study the number of modes for the mixture of two multivariate t-densities, we explored the curvature function $k(\alpha)$ analytically . We found that when the degrees of freedom for both of the components went to infinity, then the equation of the curva-

ture function of the mixture of two multivariate t-densities was exactly the same as the equation of the curvature function of the mixture of two multivariate normal densities, which means the number of modes for both of them in this case was the same. By exploring the equation of the curvature function of the mixture of two multivariate of t-densities, we found that it is difficult to simplify the $f(s)$ equation in the case of the n_1, and n_2 not being equal to infinity. The current results of exploring the equation of the curvature function of the mixture of two multivariate t-densities support our argument that the maximum number of modes for the mixture of two multivariate t-densities is the same as the maximum number of modes for the mixture of two multivariate normal densities. This is an interesting open problem.

In the process of doing this study on the effect of the number of degrees of freedom on the number of modes for the mixture of two multivariate t-densities, and on the maximum number of modes that can be got from mixing two t-components, we have come to believe that these results could be used for a wide range of statistical purposes. The results on number of modes for the mixture of two multivariate densities will help us decide if it is possible to merge the components of the mixture of the multivariate t-distribution. Finally, these results provide a clear understanding of the effect of the degrees of freedom on the number of modes for the mixture of two multivariate t-densities, and this might be of help in exploring the number of the modes for the mixture of multivariate skewed t-densities.

Chapter 4

Modality of Skew Normal Densities

In this chapter we will first give an overview of the univariate and multivariate skew normal densities. The rest of the chapter will focus on understanding the behaviour of mode of a one component skew-normal distribution. It should be noted that unlike normal or t-densities the modes of skew-normal is not equal to the mean. Furthermore the mode cannot be expressed in closed form and needs to be calculated numerically. We will later use these results to build a framework to understand the modes and modal behaviour of a mixture of skew-normals.

NOTE: This chapter is also adapted from: **Alruwaili B. and Ray S.** Topography of skew normal distributions and skew normal mixtures. *Under preparation* (2019).

4.1 Introduction

The normal distribution is the most popular distribution in Statistics for modelling continuous data that are symmetric. But when we encounter data that are asymmetric, the normal distribution may not be the most suitable distribution for modelling. To model skewed data, the density distribution should show more flexibility. Azzalini (1985) extended the normal distribution to build in more flexibility to model datasets

that are skewed by introducing a skewness parameter. This new class of distributions is referred to as the skew normal distribution. Furthermore, Azzalini (1985) presented the basic properties of skew normal distribution and represented the means and variances in closed form. He later introduced more results for skew normal distributions which provided the skewness parameters of the density that shape the fixable range for the indices of skewness and kurtosis. Following Azzalani, Henze (1986) provided a neat probabilistic representation for the skew normal in terms of a normal random variable and a truncated normal random variable. This representation provided a clearer understanding of the structure of the skew normal class.

In further research exploring and using the concept of skew normal distribution, Pewsey (2000) considered the problems of inference relating to Azzalini's skew normal distribution; he used the method of moments and maximum likelihood estimation for the centre parameterization instead of using direct parameterization for estimation. By using different values for the skewness parameters of univariate skew normal Gupta and Chen (2001) produced a table for the cumulative distribution function (cdf) of skew normal distribution and used it to examine of the quality of the fit tests for skew normal distributions. More theoretical literature on the skew normal distribution and its related families were provided by Azzalini (2005), Arellano-Valle and Azzalini (2006),and Azzalini (2013).

For the multivariate case of the skew normal distribution, Azzalini and Dalla Valle (1996) extended the concept of the univariate skew normal distribution to introduce the multivariate skew normal. They showed the properties of the multivariate skew by focusing on the bivariate case of the skew normal distribution. Azzalini and Capitanio (1999) introduced further probabilistic properties for the multivariate skew normal distribution by focusing on aspects of statistical relevance, and introducing some applications to discuss some multivariate statistical problems. Genton et al. (2001) provided

the moments of random vectors of the distribution of multivariate skew normal and discussed specific applications in the area of spatial statistics and time series. Capitanio et al. (2003) studied the importance of the distribution of the multivariate skew normal in graphical models. Gupta and Chen (2004) introduced some new possibilities for extending the univariate skew univariate model to multivariate skew normal models.

In the years since these initial studies, a number of other results relating to the extension of the univariate skew normal to the multivariate skew normal have been published. In a particularly notable study, Branco and Dey (2001) generalized the multivariate skew normal of Azzalini and Dalla Valle (1996) to apply to the skew elliptical distribution.

Further studies into the extension of the multivariate skew normal were carried out by Lachos et al. (2010). Readers interested in the details of extensions of the multivariate skew normal can consult the review paper by Lee and Mclachlan (2013) which provides a concise overview of the extension and classification of existing multivariate skew normal distributions, and multivariate t-distributions. In this chapter we will develop modal tools and explore their properties for the class of the univariate skew normal distribution introduced by Azzalini (1985) and their original extension to the multivariate skew normal distribution case (Azzalini and Dalla Valle, 1996).

4.2 Properties of the univariate skew normal distribution

In this section we explore the mode of the univariate skew normal distribution. The skew normal distribution was introduced by Azzalini (1985) as an extension of normal distribution with an extra parameter regarding the skewness of the density shape. This skewness parameter allowed more flexibility for this distribution to skew to the right

side when the skewness parameter was greater than zero and skew to the left side when the skewness parameter was less than zero. When the skewness parameter was zero, then the skewness disappeared, and the distribution of the skew normal distribution became the normal distribution. The density of the skew normal distribution converges to a half-normal when the skewness parameter goes to infinity. We start with the expression of the density of skew normal according to Azzalini (1985).

Definition 4.1. *Let X be a univariate random variable following a standard skew normal distribution, then the probability density function is given by:*

$$f_{sn}(x) = 2\phi(x)\Phi(\lambda x), \quad (-\infty < x < \infty), \lambda \in \mathbb{R}, \tag{4.1}$$

where λ is the skewness parameter, and $\phi(x)$ and $\Phi(x)$ are the standard normal probability density function (pdf) and the cumulative distribution function (cdf) of the skew normal distribution, respectively. Recall that standard normal density function $\phi(x)$ is defined by:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty. \tag{4.2}$$

The cumulative distribution function $\Phi(\lambda x)$ of the univariate skew normal distribution is given by:

$$\Phi(\lambda x) = \int_{-\infty}^{\lambda x} \phi(t)dt. \tag{4.3}$$

Figure 4.1 show the density of the univariate skew normal distribution with different values of the skewness parameters λ.

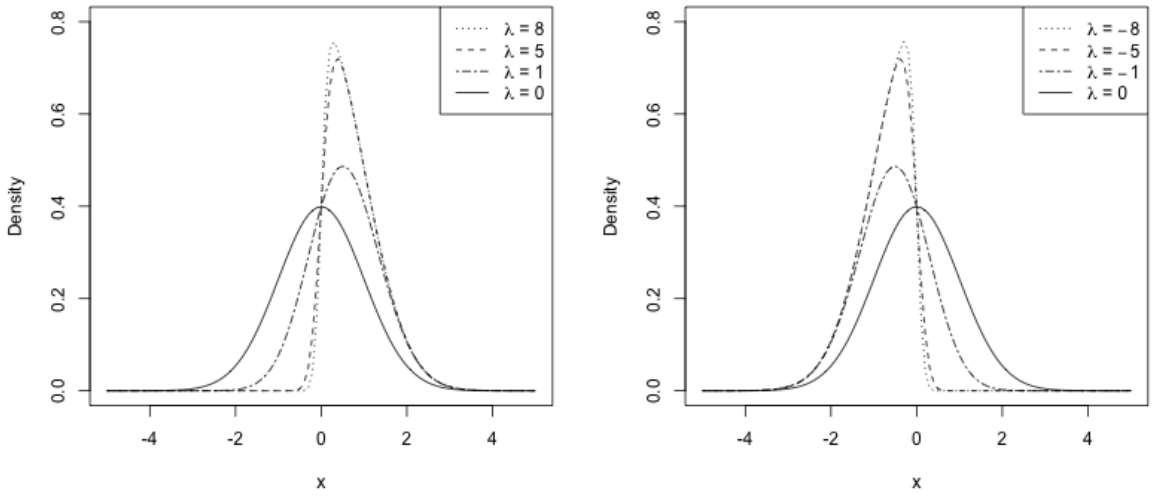


Figure 4.1: Plots of the density function of univariate skew normal distribution with different values of λ.

The left panel shows the densities of the skew normal distribution for $\lambda > 0$, resulting in density shapes that are skewed to the right which makes the tails on the right heavier than those on the left. In contrast, the right-panel in Figure 4.1 plots the densities of the skew normal distribution that are skewed to the left side, as a results of the skewness parameter λ being negative. The density of the skew normal is the same as the normal density when $\lambda = 0$.

To understand the properties of a general skew-normal distribution, the location and scale parameters are introduced through an affine transformation. The affine transformation is defined by

$$Y = \xi + \omega X, \tag{4.4}$$

where ξ is the location parameter, and ω is the scale parameter $\omega > 0$. As X follows a skew normal distribution, Y also follows a skew normal distribution but with location

parameter μ and scale parameter ω and can be written as:

$$f_{sn}(y;\xi,\omega,\lambda) = \frac{2}{\omega}\phi\Big(\frac{y-\xi}{\omega}\Big)\Phi\Big(\lambda\Big(\frac{y-\xi}{\omega}\Big)\Big), \tag{4.5}$$

and is compactly denoted by

$$Y \sim SN(\xi,\omega^2,\lambda).$$

Azzalini (1985) computed the mean and variance of skew normal distribution by introducing the moment-generating function of the univariate skew normal distribution. The moment-generating function of the univariate skew normal distribution is defined as

$$M_x(t) = 2e^{t^2/2}\Phi(\delta t), \qquad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}.$$

The mean of skew normal distribution with location 0 and scale 1 is given by the first moment of the moment generating function of the univariate skew normal distribution, which simplifies to

$$E(X) = \sqrt{\frac{2}{\pi}}\delta, \qquad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{4.6}$$

Using the affine transformation in (4.4) the mean of a skew normal distribution after including the location and scale parameters can be written as:

$$E(Y) = \xi + \omega\sqrt{\frac{2}{\pi}}\delta. \tag{4.7}$$

By using the first and second moments of the moment generating function of uni-

variate skew normal, the variance of skew normal distribution is given by:

$$Var(X) = 1 - \frac{2}{\pi}\delta^2, \qquad \delta^2 = \left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2. \tag{4.8}$$

The variance of skew normal distribution after including the location and scale parameters can now be written as:

$$Var(Y) \;=\; \omega^2\Big(1 - \frac{2}{\pi}\delta^2\Big), \qquad \delta^2 = \left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2. \tag{4.9}$$

4.2.1 Mode of the univariate skew normal distribution

In this section we explore the mode of the univariate skew normal distribution. Recall, that the mode is defined as the local maximum of the probability density. There are two ways to locate the modes, namely graphically and analytically. Graphically, the peak of the density curve gives the mode of the distribution. Analytically, the mode of the distribution is given by the value of x where the first derivative is equal to zero, and the second derivative is less than zero. In this section, we will first show that the mode of skew-normals cannot be written in closed form. Then we will derive the implicit equation to calculate the modes of univariate skew normal distributions and explore the location of modes graphically. Recall, that the probability density function of the univariate standard skew normal distribution is given by:

$$f_{sn}(x) = 2\phi(x)\Phi(\lambda x), \quad (-\infty < x < \infty), \lambda \in \mathbb{R}.$$

Theorem 4.1. *The mode of a univariate normal can be calculated by solving the implicit equation*

$$h(x) = \lambda\phi(\lambda x) - x\Phi(\lambda x) = 0. \tag{4.10}$$

Proof. The mode of the univariate standard skew normal distribution is the solution to $f_{sn}^{'}(x) = 0$, and $f_{sn}^{''}(x) < 0$. To obtain the mode of the univariate skew normal, we first calculate the derivatives

$$f_{sn}'(x) = \frac{\partial}{\partial_x} f(x) \quad = \quad 2\phi'(x)\Phi(\lambda x) + 2\phi(x)\Phi'(\lambda x).$$

We first calculate $\phi'(x)$ and $\Phi'(\lambda x)$.

$$\begin{aligned}
\text{For} \quad \phi(x) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\
\phi'(x) &= \frac{-2x}{2\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\
&= \frac{-x}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \\
&= -x\phi(x). \\
\text{As} \quad \Phi(\lambda x) &= \int_{-\infty}^{\lambda x} \phi(t)dt \\
\Phi'(\lambda x) &= \lambda\phi(\lambda x).
\end{aligned}$$

Putting the expression for $\phi'(x)$ and $\Phi'(\lambda x)$ we get

$$\begin{aligned}
f_{sn}'(x) &= 2(-x)\phi(x)\Phi(\lambda x) + 2\phi(x)\lambda\phi(\lambda x) \\
&= 2\phi(x)\Big[(-x)\Phi(\lambda x) + \lambda\phi(\lambda x)\Big].
\end{aligned}$$

As $\phi(x)$ is the normal density function, $\phi(x) > 0$, and so the mode of the univariate skew normal distribution can be obtained by solving the implicit equation:

$$h(x) = \lambda\phi(\lambda x) - x\Phi(\lambda x) = 0. \tag{4.11}$$

Now we calculate the second derivatives of $f_{sn}(x)$ to show that $f_{sn}^{''}(x) < 0$

$$
\begin{aligned}
f'_{sn}(x) &= 2\Big[(-x)\phi(x)\Phi(\lambda x) + \phi(x)\lambda\phi(\lambda x)\Big] \\
f''_{sn}(x) &= 2\Big[(-1)\phi(x)\Phi(\lambda x) + (-x)(-x\phi(x))\Phi(\lambda x) + (-x)\phi(x)(\lambda\phi(\lambda x) \\
&\quad +(-x\phi(x)(\lambda\phi(\lambda x) - \phi(x)(\lambda^3 x)\phi(\lambda x)\Big] \\
f''_{sn}(x) &= 2\phi(x)\Big[(-1)\Phi(\lambda x) + (-x)(-x)\Phi(\lambda x) + (-x)(\lambda\phi(\lambda x) \\
&\quad +(-x)(\lambda\phi(\lambda x) - (\lambda^3 x)\phi(\lambda x)\Big]
\end{aligned}
$$

as $\lambda\phi(\lambda x) = x\Phi(\lambda x)$ we can rewrite $f'_{sn}(x)$ as follows

$$
\begin{aligned}
f''_{sn}(x) &= 2\phi(x)\Big[(-1)\Phi(\lambda x) + x^2\Phi(\lambda x) + (-x)(x\Phi(\lambda x)) \\
&\quad -x(x\Phi(\lambda x)) - (\lambda^2 x)x\Phi(\lambda x)\Big] \\
&= 2\phi(x)\Big[(-1)\Phi(\lambda x) + x^2\Phi(\lambda x) + (-x^2)(\Phi(\lambda x)) \\
&\quad -x^2(\Phi(\lambda x)) - (\lambda^2 x^2)\Phi(\lambda x)\Big] \\
&= 2\phi(x)\Phi(\lambda x)\Big[-1 + x^2 - x^2 - x^2 - (\lambda^2 x^2)\Big] \\
&= -2\phi(x)\Phi(\lambda x)\Big[1 + x^2 + (\lambda^2 x^2)\Big]
\end{aligned}
$$

so $f''_{sn}(x) < 0$ as $\phi(x)\Phi(\lambda x)$ is always positive and each term in $1+x^2+(\lambda^2 x^2)$ is positive

□

To explore the mode of the univariate skew normal distribution graphically, we can plot equation 4.11 and locate its zero crossing to calculate the mode of the univariate of skew normal distribution. Alternatively one can use root-solving methods and calculate the mode numerically. The next few plots shows the modes of skew normal distribution with different parameters. In each case the mode is calculated numerically by solving the implicit equation given in theorem 4.1 by using the `sn.mode` function from the github repository `skewed-normal-codes` (Alruwaili and Ray, 2019).

Example 4.1. *Mode of the univariate skew normal distribution with the following parameters:* $\xi = 0$, $\omega = 1$, *and different values of skewness parameters* $\lambda = 0, 1,$ *and* -1.

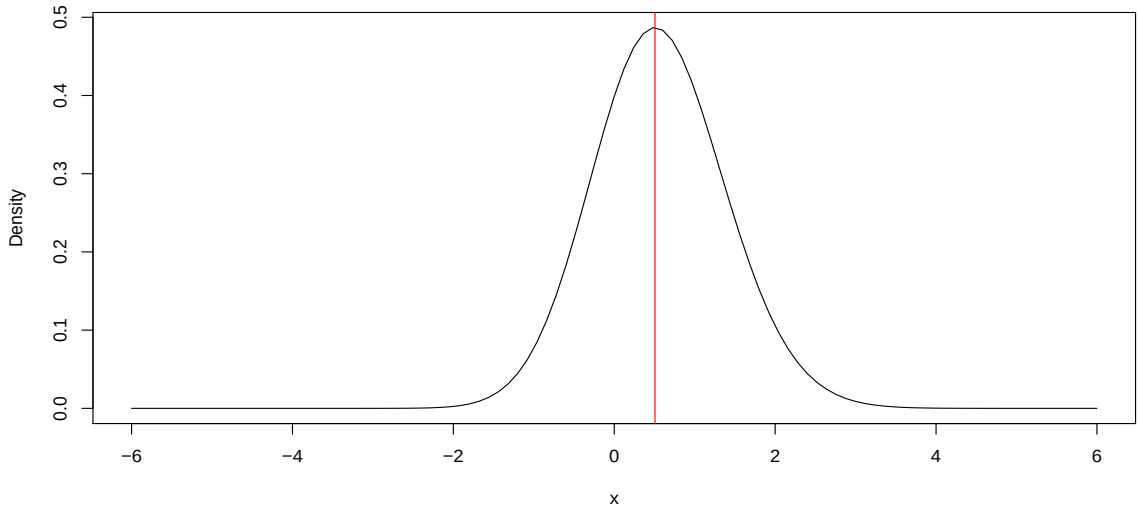


Figure 4.2: Plot of the density function of skew normal parameters given in Example 4.1 where the mode is located at $x = 0.506033$ when $\lambda = 1$.

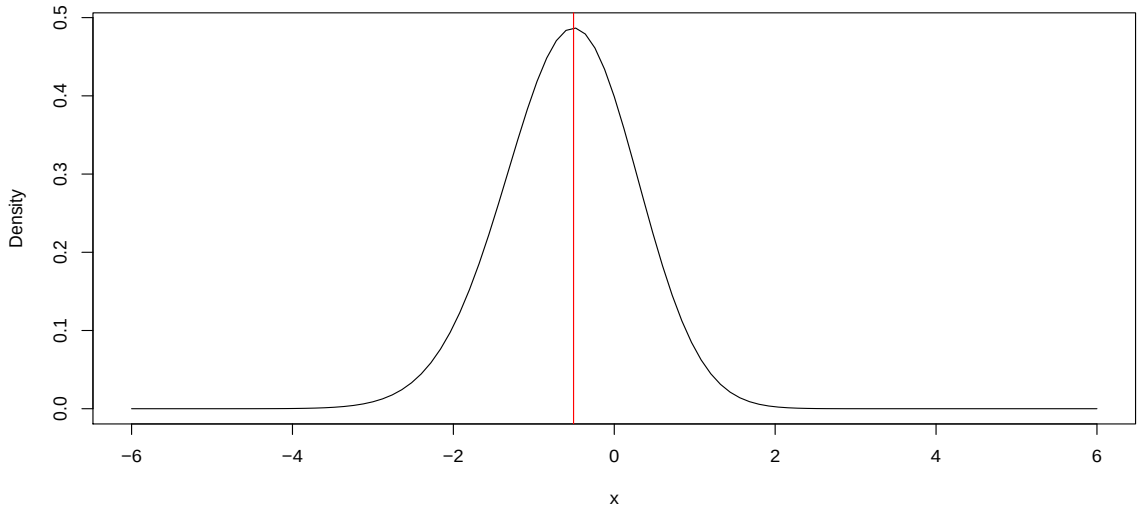


Figure 4.3: Plot of the density function of skew normal parameters given in Example 4.1 where the mode is located at $x = -0.506033$ when $\lambda = -1$.

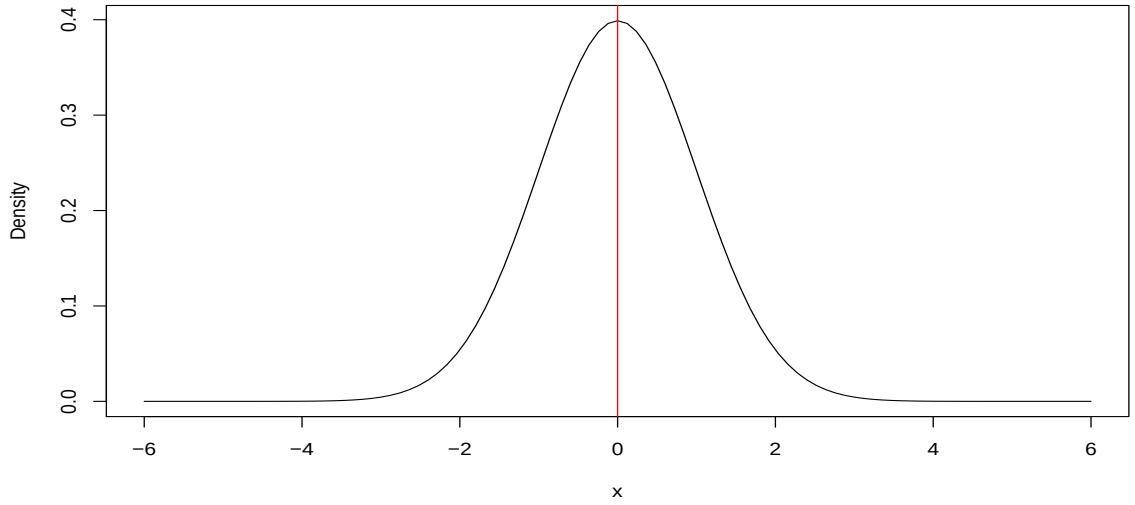


Figure 4.4: Plot of the density function of skew normal parameters given in Example 4.1 where the mode is located at $x = 0$ when $\lambda = 0$.

Example 4.2. *Mode of the univariate skew normal distribution with the following parameters:* $\xi = 2$, $\omega = 1$, *and different values of skewness parameters* $\lambda = 0, 1,$ *and* -1.

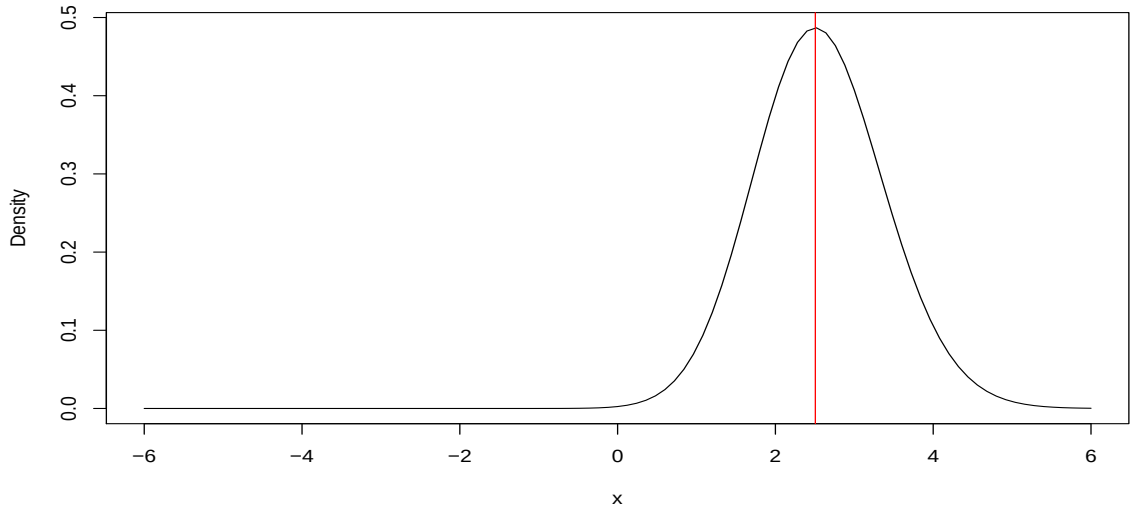


Figure 4.5: Plot of the density function of skew normal parameters given in Example 4.2 where the mode is located at x=2.506033 when $\lambda = 1$.

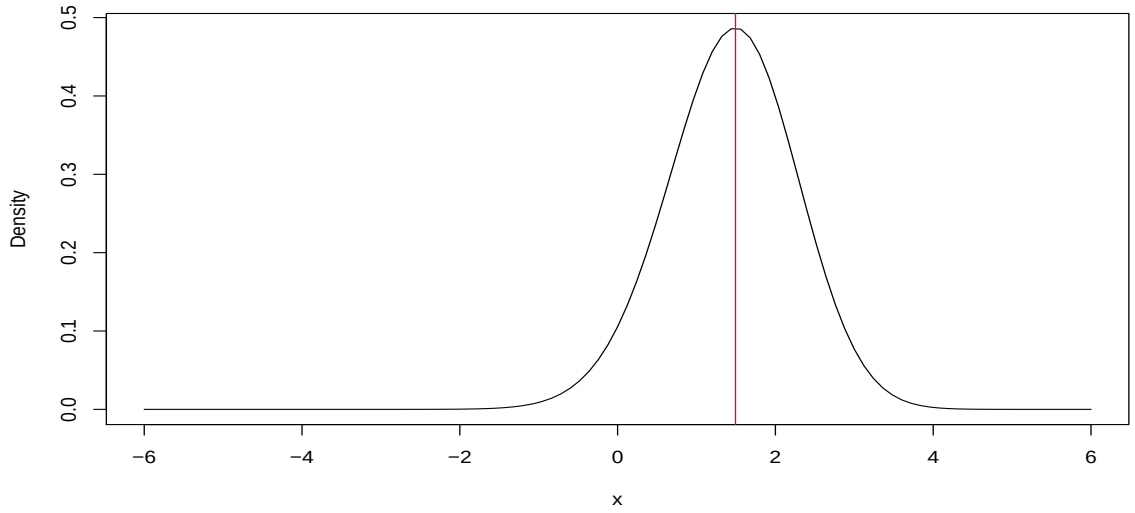


Figure 4.6: Plot of the density function of skew normal parameters given in Example 4.2 where the mode is located at $x = 1.493967$ when $\lambda = -1$.

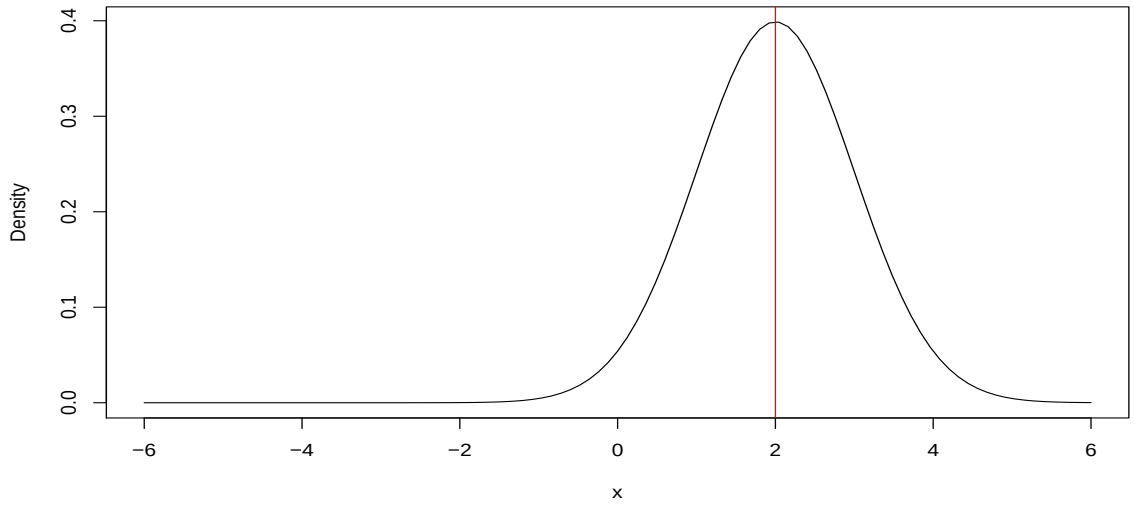


Figure 4.7: Plot of the density function of skew normal parameters given in Example 4.2 where the mode is located at $x = 2$ when $\lambda = 0$.

The plots corresponding to examples 4.1, and 4.2 show the density shapes of the univariate skew normal distribution with the location of the modes for the various different parameters and different values of the skewness parameters. For $\lambda > 0$, i.e

when the density shape of the univariate skew normal distribution is skewed to the right side the mode is larger than the location parameter ξ. In contrast, when $\lambda < 0$, i.e. when the density shape of the univariate skew normal distribution is skewed to the left the modes of the univariate skew normal distribution are located to the left of the location parameter ξ. And finally it is easy to check that when the skewness parameters is zero, the skewness disappears, and the density of the skew normal distribution becomes normal density, and the mode is same as the location parameter.

4.3 Relating the mean and the mode of the univariate skew normal

In the last section we have seen specific examples of the effect of the skewness parameters on the density shape of the skew normal distribution, and the location of the mode. In this section we will provide a comprehensive study of the relationship of the mean and mode of the skew normal distribution as a function of the skewness parameter λ and also explore the limiting cases when the skewness parameter $\lambda \to \pm\infty$.

4.3.1 Mean and variance of the skew normal distribution as a function of skewness parameter

To explore the effect of skewness parameters on the mean and variance of the skew normal distribution, we introduce the mean and variance of the univariate skew normal distribution, which was given by Azzalini (1985). The mean of skew normal distribution, in terms of the location and scale parameters, can be written as:

$$E(Y) = \xi + \omega\sqrt{\frac{2}{\pi}}\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right).$$

The variance of skew normal distribution, in terms of the location and scale parameters, can be written as:

$$Var(Y) \;=\; \omega^2\Big(1-\frac{2}{\pi}\delta^2\Big), \qquad \delta^2=\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2.$$

Lemma 4.2. *The mean of a skewed normal* $E(Y)=\xi+\omega\sqrt{\frac{2}{\pi}}\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)$ *converges to* $\xi\pm\omega\sqrt{\frac{2}{\pi}}$ *as* $\lambda\to\pm\infty$.

Proof. For $\lambda\to+\infty$ we have:

$$\begin{aligned}
\lim_{\lambda\to+\infty}\frac{\lambda}{\sqrt{1+\lambda^2}} &= \lim_{\lambda\to+\infty}\frac{\frac{\lambda}{\sqrt{\lambda^2}}}{\sqrt{\frac{1+\lambda^2}{\lambda^2}}}\\
&= \lim_{\lambda\to+\infty}\frac{\frac{\lambda}{\lambda}}{\sqrt{\frac{1}{\lambda^2}+\frac{\lambda^2}{\lambda^2}}}\\
&= \lim_{\lambda\to+\infty}\frac{1}{\sqrt{\frac{1}{\lambda^2}+1}}\\
&= \frac{1}{\sqrt{0+1}}\\
&= 1.
\end{aligned}$$

For $\lambda\to-\infty$, we suppose that $\lambda=-\theta$ than we have

$$\begin{aligned}
\lim_{\lambda\to-\infty}\frac{\lambda}{\sqrt{1+\lambda^2}} &= \lim_{\theta\to\infty}\frac{-\theta}{\sqrt{1+(-\theta)^2}}\\
&= -\Big(\lim_{\theta\to\infty}\frac{\theta}{\sqrt{1+(-\theta)^2}}\Big)\\
&= -1.
\end{aligned}$$

Thus, the mean of univariate skew normal distribution converges to $\xi\pm\omega\sqrt{\frac{2}{\pi}}$ as the skewness parameters $\lambda\to\pm\infty$. □

In particular, the mean of standard univariate skew normal distribution converges to

$E(X) = \pm\sqrt{\frac{2}{\pi}}$ as skewness parameters of univariate skew normal distribution $\lambda \to \pm\infty$. Now we calculate the limiting value of the variance of skew normals.

Lemma 4.3. *The variance of a skewed normal* $Var(Y) = \omega^2(1 - \frac{2}{\pi}\delta^2)$ *converges to* $\omega^2(1 - \frac{2}{\pi})$ *as* $\lambda \to \pm\infty$, *where* $\delta^2 = \left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2$.

Proof. When the skewness parameter $\lambda \to \pm\infty$ we have:

$$\begin{aligned}
\lim_{\lambda\to\pm\infty}\left(\frac{\lambda}{\sqrt{1+\lambda^2}}\right)^2 &= \lim_{\lambda\to\pm\infty}\frac{\lambda^2}{1+\lambda^2} \\
&= \lim_{\lambda\to\pm\infty}\frac{\frac{\lambda^2}{\lambda^2}}{\frac{1}{\lambda^2}+\frac{\lambda^2}{\lambda^2}} \\
&= \lim_{\lambda\to\pm\infty}\frac{1}{\frac{1}{\lambda^2}+1} \\
&= \frac{1}{0+1} \\
&= 1
\end{aligned}$$

Thus, the variance of univariate skew normal distribution converges to $\omega^2(1 - \frac{2}{\pi})$. □

In particular, the variance of a standard univariate skew normal distribution converges to $Var(X) = (1 - \frac{2}{\pi})$ when skewness parameters of univariate skew normal distribution $\lambda \to \pm\infty$.

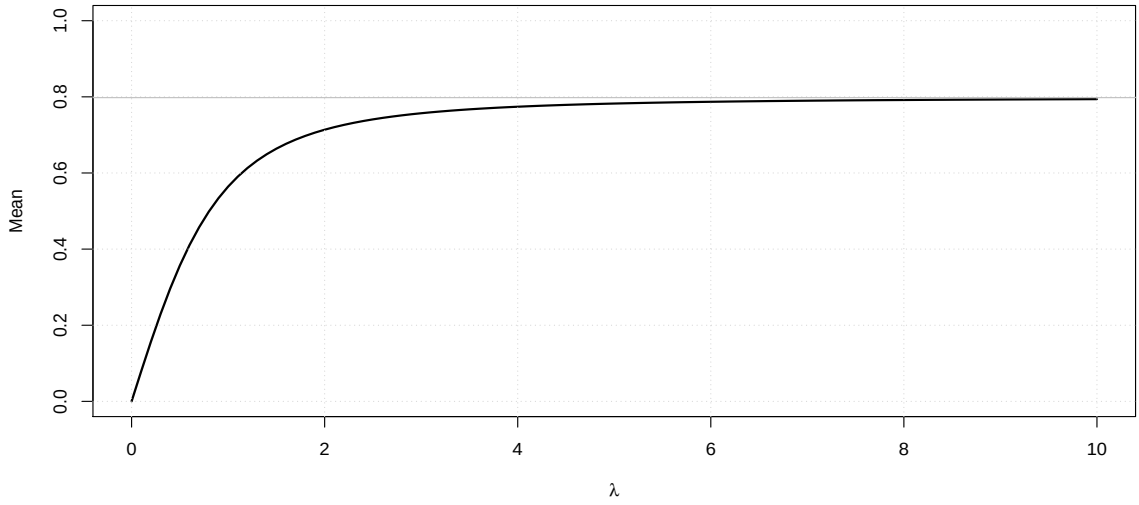


Figure 4.8: Mean of the standard skew normal distribution as a function of λ, for $\lambda > 0$. The gray line represents the asymptotic value $\sqrt{\frac{2}{\pi}}$.

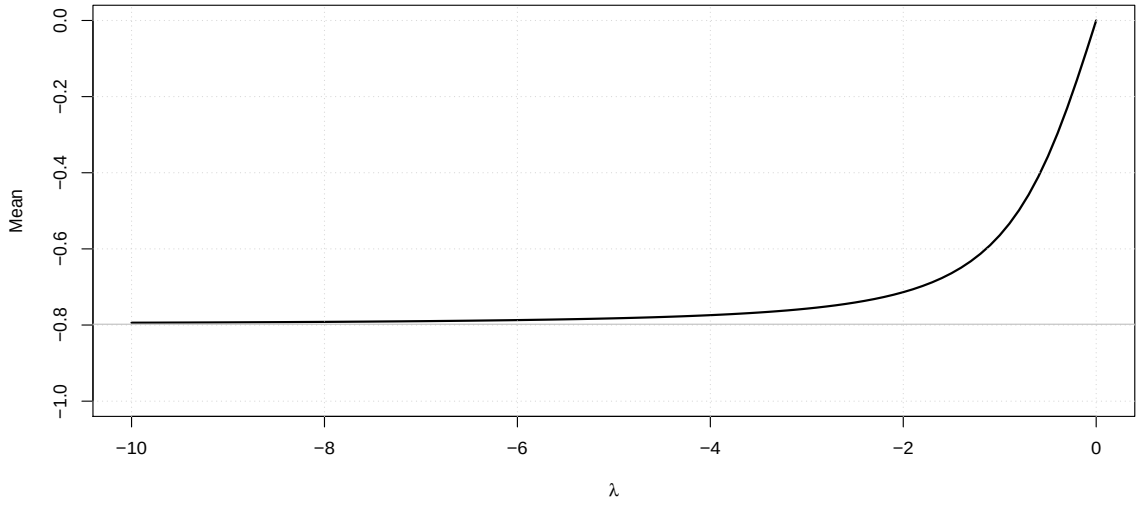


Figure 4.9: Mean of the standard skew normal distribution as a function of λ, for $\lambda < 0$. The gray line represents the asymptotic value $-\sqrt{\frac{2}{\pi}}$.

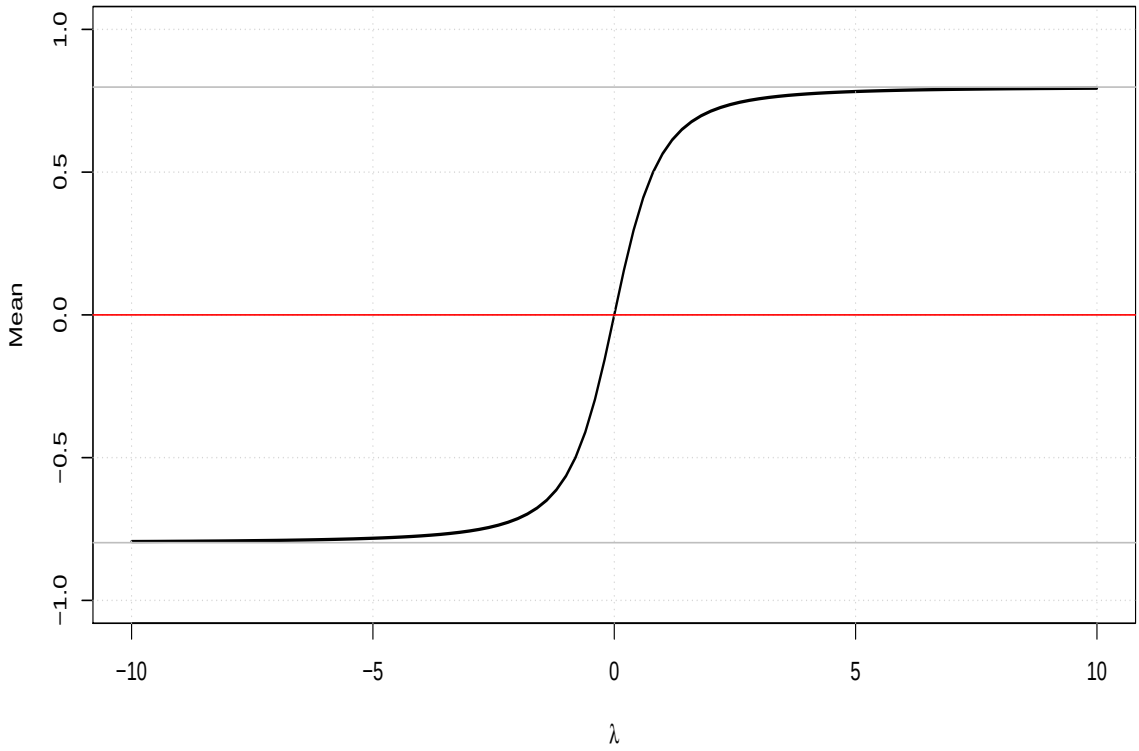


Figure 4.10: Mean of the standard skew normal distribution as a function of λ. Here the two asymptotic plots are given by the two gray line.

Figures 4.8 and 4.9 show the means of univariate skew normal distributions $SN(0, 1, \lambda)$, for positive and negative $\lambda's$ with the asymptotic values. Figure 4.10 show the mean for a large range of λ, which clearly shows the mean convergence to the asymptotic values as λ becomes large.

From corollary 4.2 and corollary 4.3 we found that the mean and variance of skew normal distribution when the skewness parameters $\lambda \to \pm\infty$ are the same as the mean and variance of the half-normal distribution (Wallner, 2016) given by pdf

$$\sqrt{\frac{2}{\pi\sigma^2}}\exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Note that the mean of half-normal distribution is given by

$$\sqrt{\frac{2}{\pi}}.$$

and the variance of half-normal distribution is given by

$$\Big(1-\frac{2}{\pi}\Big).$$

4.3.2 Effect of skewness parameter on the mode of the skew normal distribution

In this section, we explore the modes of univariate skew normal distributions as a function of λ. We first find the limiting value of the mode when the skewness parameter $\lambda \to \pm\infty$. Unlike the mean and variance of skew normal distribution, we do not have a closed form expression of the mode. So we first plot the numerical solution of the modes of standard skew normals as a function of λ.

Theorem 4.4. *The mode of standard a skewed normal distribution $SK(0,1,\lambda)$ converges to 0 as $\lambda \to \pm\infty$.*

Proof. The mode of univariate the skew normal distribution can be calculated by the following equation:

$$h(x) = \lambda\phi(\lambda x) - x\Phi(\lambda x) = 0. \tag{4.12}$$

To show that the mode of skew normal distribution converges to zero, we consider the standard univariate skew normal distribution $SK(0,1,\lambda)$ where the skewness parameters $\lambda \to \pm\infty$. First we will focus on the part I of the equation (4.12) which is

$\lambda\phi(\lambda x)$

$$\begin{aligned}\lim_{\lambda\to+\infty}\lambda\phi(\lambda x) &= \lim_{\lambda\to+\infty}\frac{\lambda}{\sqrt{2\pi}}\exp\left(-(\lambda x)^2/2\right)\\ \lim_{\lambda\to\infty}\lambda\phi(\lambda x) &= \lim_{\lambda\to+\infty}\frac{\lambda}{\sqrt{2\pi}\exp\left((\lambda x)^2/2\right)}\\ &= \lim_{\lambda\to+\infty}\frac{1}{\sqrt{2\pi}}\frac{\lambda}{\exp\left((\lambda x)^2/2\right)}.\end{aligned}$$

We will now use l'Hôpital's rule to calculate the limit. For the numerator $\frac{d}{d\lambda}\lambda = 1$ and for the denominator $\frac{d}{d\lambda}\exp\left((\lambda x)^2/2\right) = \lambda x^2\exp\left((\lambda x)^2/2\right)$ and that gives

$$\begin{aligned}\lim_{\lambda\to+\infty}\lambda\phi(\lambda x) &= \frac{1}{\sqrt{2\pi}}\frac{\left(\lim_{\lambda\to+\infty}1\right)}{\left(\lim_{\lambda\to+\infty}\lambda x^2\exp\left((\lambda x)^2/2\right)\right)}\\ &= 0 \qquad \text{as } \lim_{\lambda\to\pm\infty}\lambda x^2\exp\left((\lambda x)^2/2\right) = \pm\infty.\end{aligned}$$

Now we focus on the part II of equation (4.12) which is $-x\Phi(\lambda x)$. For part II when the skewness parameters $\lambda\to+\infty$ we have that:

$$\begin{aligned}\lim_{\lambda\to+\infty}x\Phi(\lambda x) &= x\lim_{\lambda\to+\infty}\Phi(\lambda x)\\ &= x, \text{as } \lim_{\lambda\to+\infty}\Phi(\lambda x) = 1, \text{ where } \Phi \text{ being a CDF for } x > 0.\end{aligned}$$

As $\lambda\phi(\lambda x) = 0$ and $x\Phi(\lambda x) = x$ for $SN(0,1,\lambda\to+\infty)$ the mode of the skew normal distribution converges to 0, as shown below:

$$\begin{aligned}\lim_{\lambda\to\infty}h(x) &= 0 - x\\ &\Rightarrow \quad x = 0 \text{ which the mode of } SN(0,1,\lambda\to+\infty).\end{aligned}$$

For $SN(0, 1, \lambda \to -\infty)$, the mode of the skew normal distribution converges to 0, as the mode of the $SN(0, 1, \lambda \to -\infty)$ is the mirror image of the mode of the skew normal distribution $SN(0, 1, \lambda \to \infty)$ as shown in figures 4.11 and 4.12. □

Figure 4.11 show the modes of univariate skew normal distribution $SN(0, 1, \lambda)$ as a function of λ, for $\lambda > 0$. Note that, unlike the mean, this graph is used by numerically solving $h(x) = 0$ on a grid of λ using the `sn.mode` function from the github repository `skewed-normal-codes`, Alruwaili and Ray (2019).

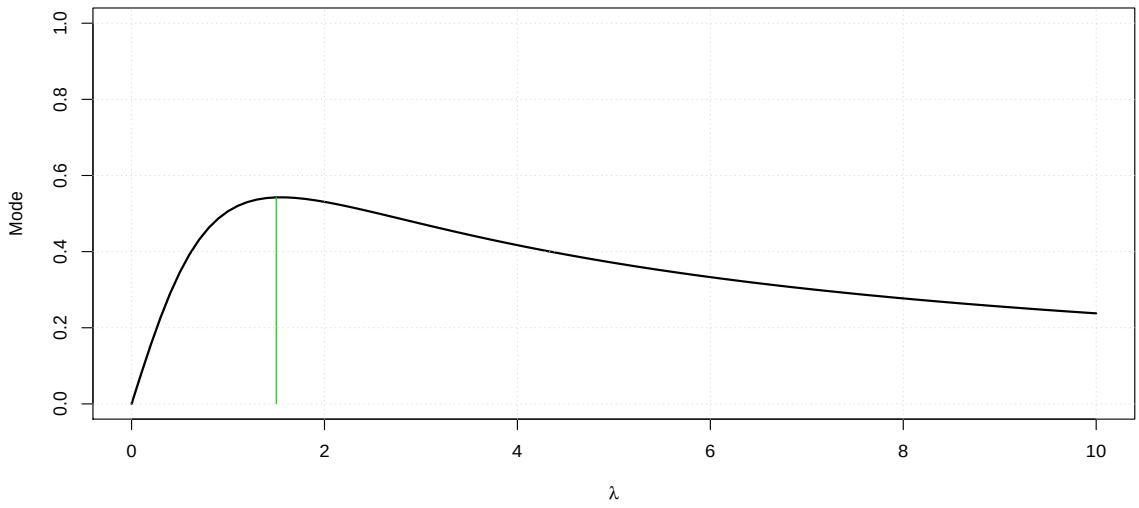


Figure 4.11: Mode of skew normal distributions $SN(0, 1, \lambda)$ for $\lambda > 0$.

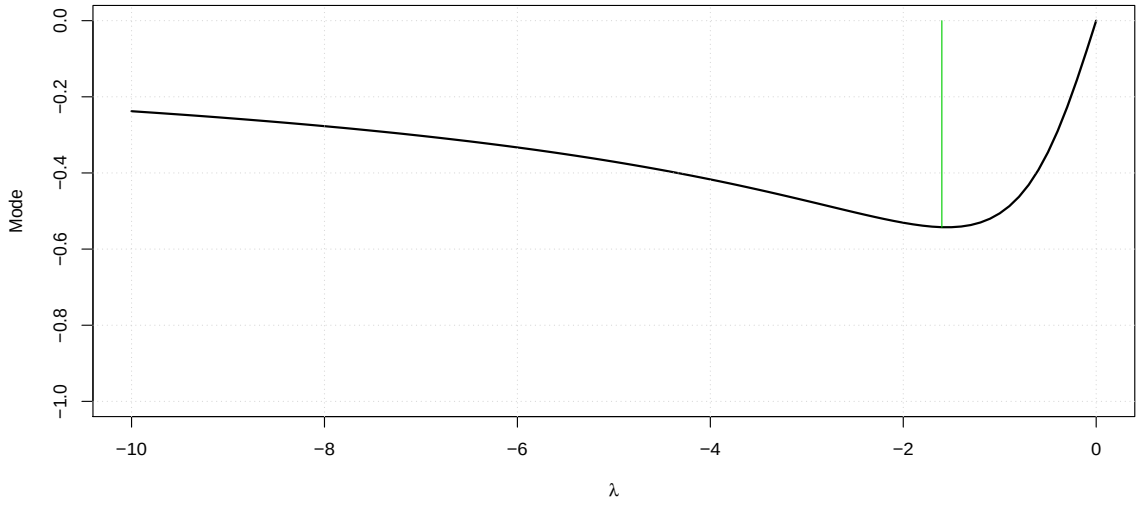


Figure 4.12: Mode of skew normal distributions $SN(0, 1, \lambda)$ for $\lambda < 0$.

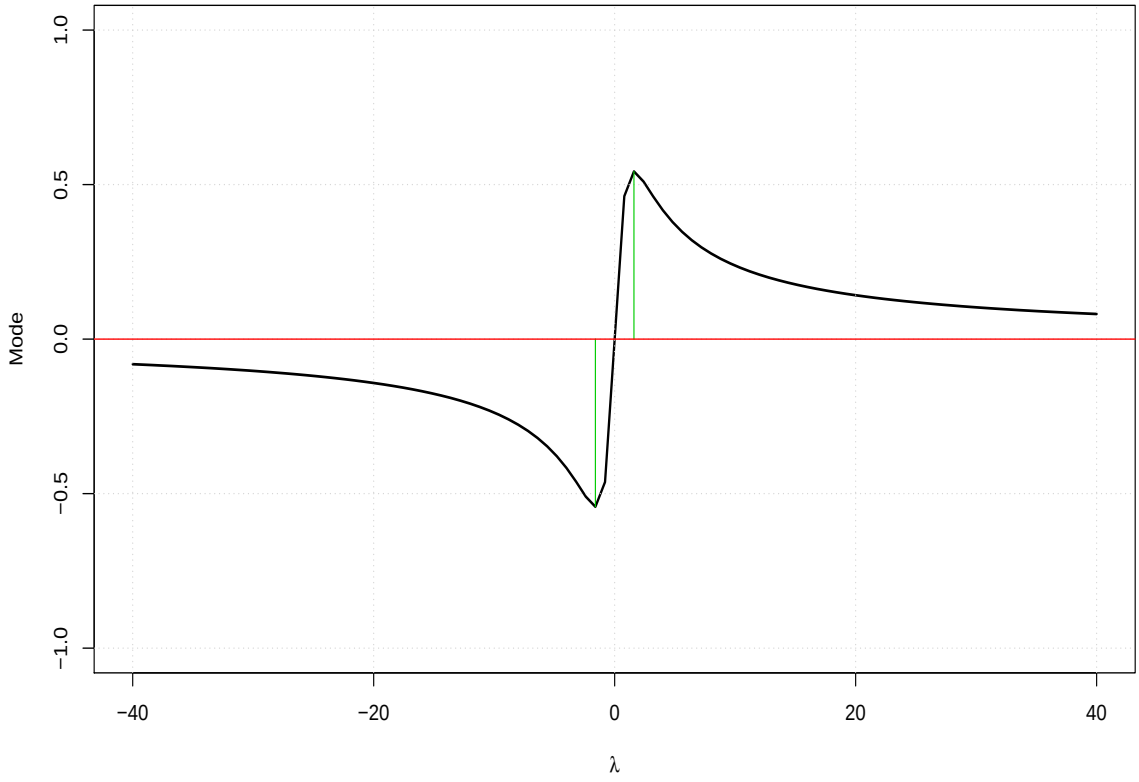


Figure 4.13: Mode of skew normal distributions $SN(0, 1, \lambda)$ for $\lambda > 0$ and $\lambda < 0$.

Figures 4.11 and 4.12 show the modes of univariate skew normal distributions $SN(0, 1, \lambda)$, for positive and negative $\lambda's$. We also observed that the maximum deviation from 0 is 0.53 which occurs for $\lambda = 1.5$ and -0.53 which occurs for $\lambda = -1.5$. Figure 4.13 show the value of the modes for a large range of λ, which clearly shows the convergence to 0 as λ becomes large.

Now, we plot the mean and the mode of the univariate skew normal skew normal distribution $SN(0, 1)$ for a range of λ along with their asymptotic values.

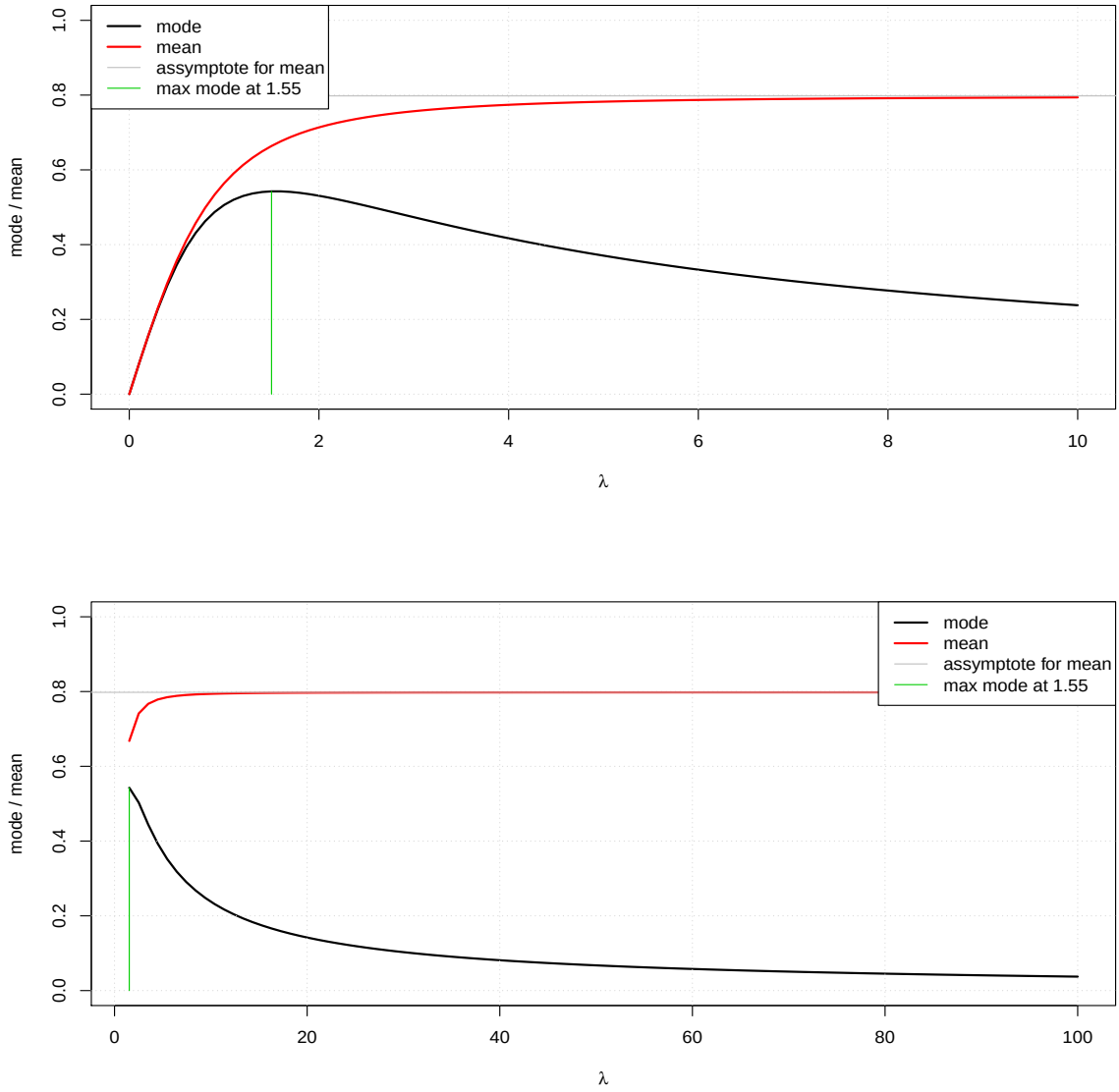


Figure 4.14: Mode and mean of skew normal distribution $SN(0, 1, \lambda)$ where $\lambda \to +\infty$.

Figure 4.14 shows the mean and the mode of the univariate skew normal distribution $SN(0, 1, \lambda)$ for $\lambda \in (0, 10)$ and $\lambda \in (0, 100)$. The plots show that the mode of the skew normal distribution is less than the mean of the skew normal distribution, and the maximum deviation for $x = 0$ when $\lambda \to \infty$ is 0.53 and this occurs at $\lambda = 1.55$. Furthermore it shows that the mode converges to 0, which the mean converges to asymptotic value $\sqrt{\frac{2}{\pi}}$.

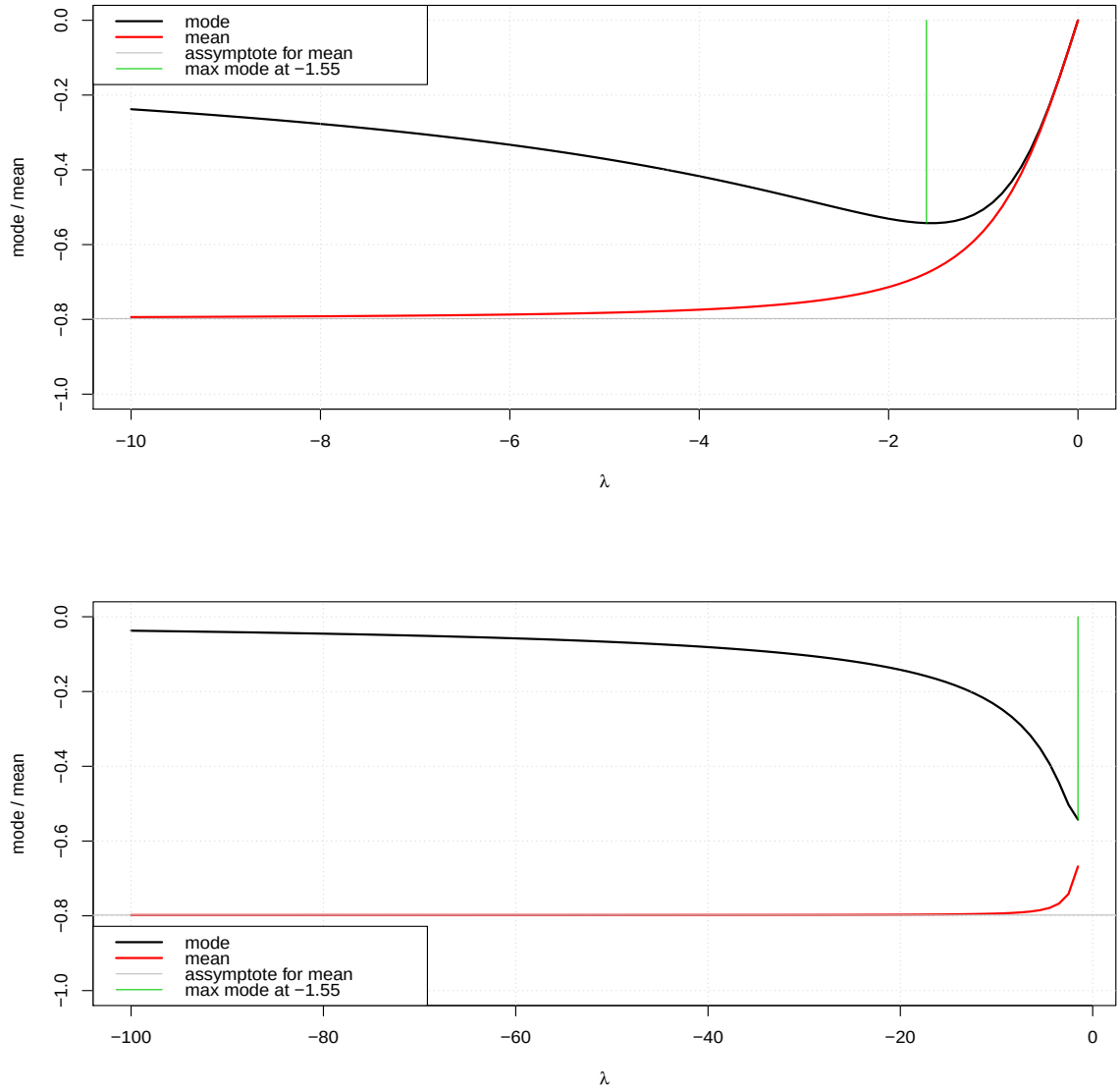


Figure 4.15: Mode and mean of skew normal distribution $SN(0, 1, \lambda)$ where $\lambda \to -\infty$.

Figure 4.15 shows the mean and the mode of the univariate skew normal skew normal distribution $SN(0, 1, \lambda)$ for $\lambda \in (0, -10)$ and $\lambda \in (0, -100)$. The plots show that the mean of the skew normal distribution is less than the mode of the skew normal distribution, and the maximum deviation of mode occurs at $\lambda = -1.55$.

Remark 1 (Search range for the numerical optimization of the mode)**.** *Exploring the mean and the mode of skew normal distribution when* $\lambda \to \pm\infty$ *enabled us with a*

good understanding of the effect of the skewness parameters on the mean and mode of the skew normal distribution. By comparing the mean and mode of the skew normal distribution when skewness parameters $\lambda \to \pm\infty$*, we found that the absolute value of the mode of the skew normal distribution was less than the absolute value of the mean of the skew normal distribution. This relationship allowed us to define the search range for the numerical optimization for the calculation of the mode. For example, let us consider the case where the skewness parameters are* $\lambda > 0$*. Without the above relationship one can use the search range* $(0, \lambda)$ *as at* $x = 0$*,* $h(x) > 0$ *and at* $x = \lambda$*,* $h(\lambda) < 0$*. But exploiting the relationship of the mean and mode we can now narrow down the search range between 0 and the mean* $x_0 = \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}$*, as* $h(x_0) < 0$*.*

4.3.3 Plotting the mode and mean of the skew normal distribution

In this section we first provide some numerical examples to identify the location of the mode and mean for univariate skew normal distributions. We will then comment on the overall relationship of the mean and mode of univariate skew normals.

Example 4.3. *Mode and mean of the univariate skew normal distribution with the following parameters:* $\xi = 0$*,* $\omega = 1$*, and different values of skewness parameters* $\lambda = 0$*,* 4*, and* -4*.*

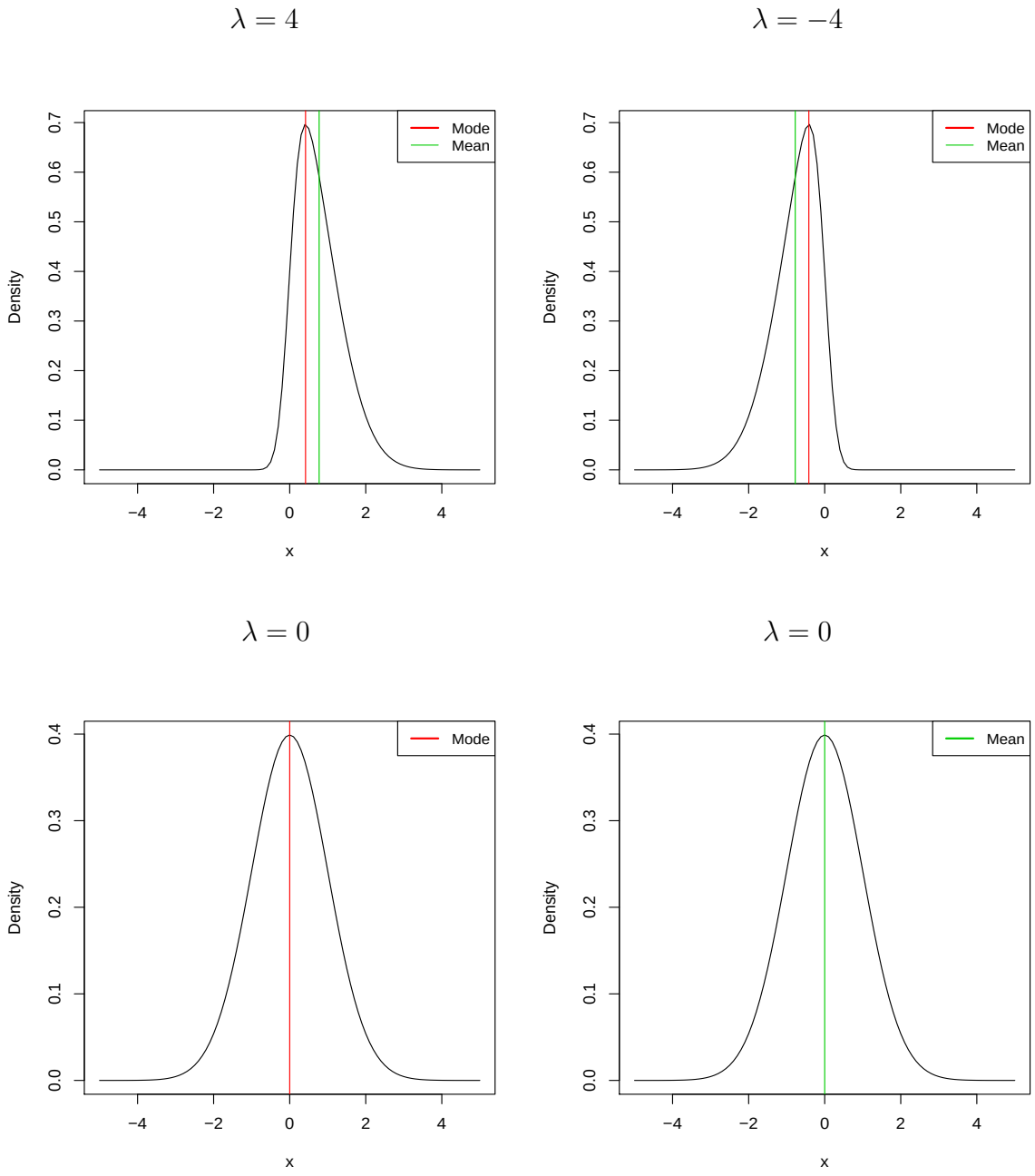


Figure 4.16: Mode and mean of skew normal distributions $SN(0, 1, \lambda)$ where λ=0, 4, and -4.

Figure 4.16 shows the mode and the mean of skew normal distributions with location parameter $\xi = 0$ and scale parameter $\omega = 1$ with different values of skewness parameters $\lambda = 0$, 4 and -4. The top plot in the left side shows the mode and mean of skew normal distribution with skewness parameter $\lambda = 4$ where the mean of skew normal is located

right the mode of skew normal distribution. Where the mode of skew normal is 0.4169 and the mean is 0.774. The top plot in the right side shows the mode and mean of skew normal distribution with skewness parameter $\lambda = -4$ where the mean of skew normal is located to the left of the mode of the skew normal distribution, and the mode of the skew normal is -0.4169 and the mean is $= -0.774$. The two plots in the bottom shows that the density of the skew normal distribution with $\lambda = 0$ becomes the normal density, where the mean and mode are same.

4.4 Maximum value of the mode of a standard skew normal distribution

Exploring the mode and mean of the skew normal distribution helps to understand that the absolute value of the mode of the skew normal distribution never exceeds the absolute value of the mean of the skew normal distribution. In the previous section we have also found that the mean increases with λ, but the mode initially increases but then decreases and finally converges to 0. We have observed that the mode never exceeds 0.53. In this section we will provide a proof for this phenomenon.

Recall, that for a given λ the mode of a skew normal distribution can be calculated by using the following equation:

$$h(x) = \lambda\phi(\lambda x) - x\Phi(\lambda x) = 0.$$

To find the solutions of this equation for univariate skew normal distributions for arbitrary λ, we create a function of two variables x and λ as $h(x, \lambda) = 0$, which is the same as $h(x)$, i.e. $h(x, \lambda) = \lambda\phi(\lambda x) - x\Phi(\lambda x)$. $h(x, \lambda) = 0$ representing all the solutions of the skew normal distribution for any arbitrary value of λ.

The solutions are given by solving $h(x, \lambda) = 0$.

For $h(x,\lambda) = 0$ we have that:

$$\begin{aligned} h(x,\lambda) &= \lambda\phi(\lambda x) - x\Phi(\lambda x) = 0 \\ &\Rightarrow \lambda x\phi(\lambda x) - x^2\Phi(\lambda x) = 0 \\ &\Rightarrow -\phi^{'}(\lambda x) - x^2\Phi(\lambda x) = 0 \\ &\Rightarrow -\phi^{'}(\lambda x) = x^2\Phi(\lambda x) \\ &\Rightarrow x^2 = \frac{-\phi^{'}(\lambda x)}{\Phi(\lambda x)} \\ x &= \pm\sqrt{\frac{-\phi^{'}(\lambda x)}{\Phi(\lambda x)}}. \end{aligned}$$

where $\lambda x\phi(\lambda x) = -\phi^{'}(\lambda x)$ as shown in the proof of lemma 4.5

Lemma 4.5. *For a normal density* $\phi(.)$

$$\lambda x\phi(\lambda x) = -\phi^{'}(\lambda x)$$

Proof.

$$\begin{aligned} \text{For} \quad \phi(\lambda x) &= \frac{1}{\sqrt{2\pi}}\exp^{-(\lambda x)^2/2} \\ \phi^{'}(\lambda x) &= \frac{-2(\lambda x)}{2}\frac{1}{\sqrt{2\pi}}\exp^{-(\lambda x)^2/2} \\ \phi^{'}(\lambda x) &= -\lambda x\frac{1}{\sqrt{2\pi}}\exp^{-(\lambda x)^2/2} \\ \phi^{'}(\lambda x) &= -\lambda x\phi(\lambda x) \\ -\phi^{'}(\lambda x) &= \lambda x\phi(\lambda x) \end{aligned}$$

□

For easy representation we define $\pm\sqrt{\frac{-\phi^{'}(\lambda x)}{\Phi(\lambda x)}}$ by $g(x,\lambda)$. Thus the solution to $h(x,\lambda) = 0$ is given by the implicit equation $x = g(x,\lambda)$. Plotting the two func-

tion $z = x$ and $z = g(x, \lambda)$ for different values of λ, the solution to $h(x, \lambda) = 0$ is given by the intersection of the function $z = x$ and $z = g(x, \lambda)$.

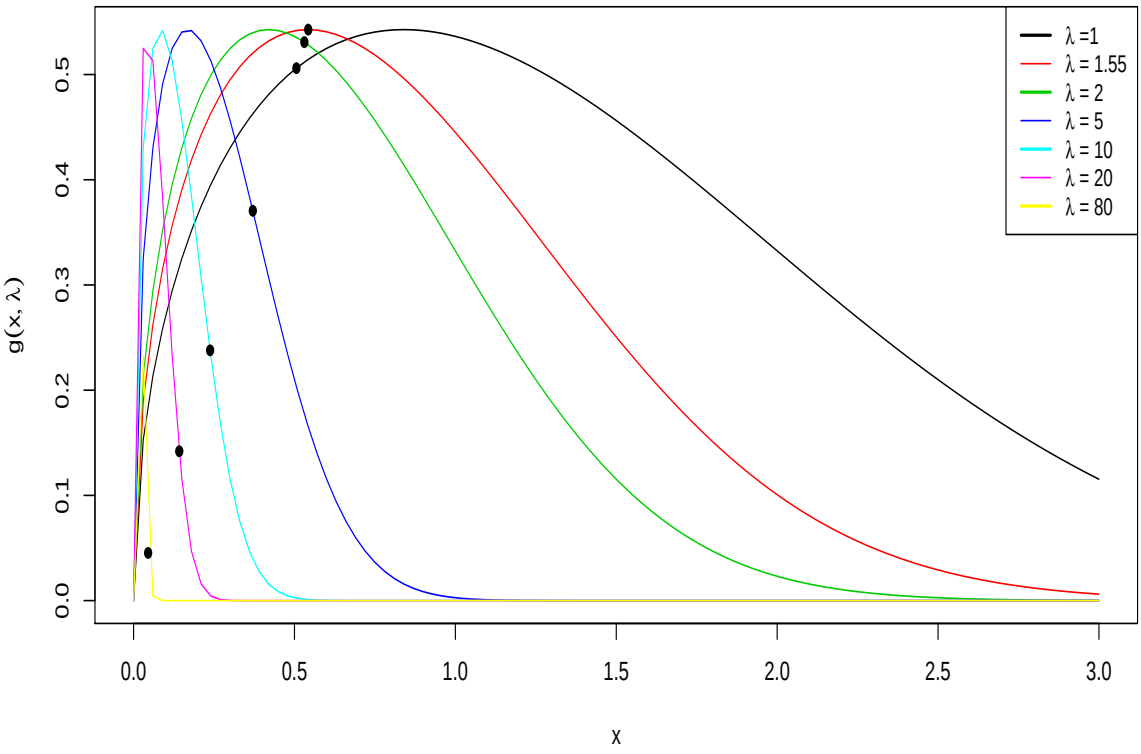


Figure 4.17: Plot of $g(x, \lambda)$ for different values of λ with their intersection with $z = x$.

Figure 4.17 shows the curves of $g(x, \lambda)$ for different values of λ with their intersection with $z = x$. For example, the red curve represents the curve of $g(x, \lambda = 1.55)$ and it intersects with the line $z = x$ at 0.53. Another example is for the large value of λ in yellow curve, where the yellow curve shows the curve of $g(x, \lambda = 80)$ and it intersects with $z = x$ at 0.045. The curves of the $g(x, \lambda)$ show that even when the value of λ is very high they intersect at a value lower than 0.53.

In the next graph we will implement a slight change of variable to provide a clearer picture of the solution given by the implicit equation. We define $x = \frac{y}{\lambda}$, then we can write

$$\begin{aligned} h(y,\lambda) &= y\phi(y) - \left(\frac{y}{\lambda}\right)^2 \Phi(y) = 0 \\ &= -\phi^{'}(y) - \left(\frac{y}{\lambda}\right)^2 \Phi(y) = 0 \\ &= -\phi^{'}(y) = \left(\frac{y}{\lambda}\right)^2 \Phi(y) \\ \Rightarrow \left(\frac{y}{\lambda}\right)^2 &= \frac{-\phi^{'}(y)}{\Phi(y)} \\ \Rightarrow \left(\frac{y}{\lambda}\right) &= \pm\sqrt{\frac{-\phi^{'}(y)}{\Phi(y)}}. \end{aligned}$$

For easy representation we define $\pm\sqrt{\frac{-\phi^{'}(y)}{\Phi(y)}}$ by $g^*(y)$. Thus the solution $h(y,\lambda) = 0$ is given by the implicit equation $\frac{y}{\lambda} = g^*(y)$. Thus the modes are now given by the intersection of the line $z = \frac{y}{\lambda}$ and the curve $g^*(y)$. Figure 4.18 gives the curve of $g^*(y)$ and the lines $z = \frac{y}{\lambda}$ for $\lambda = (1, 1.55, 2, 5, 10, 20)$. Though the lines intersects at the different values of the curve all the solutions lies on the curve. Thus the maximum value of the mode is equal to the maximum value of the curve $g^*(y)$, and it can be numerically shown that the maximum value of the curve $g^*(y)$ is 0.53.

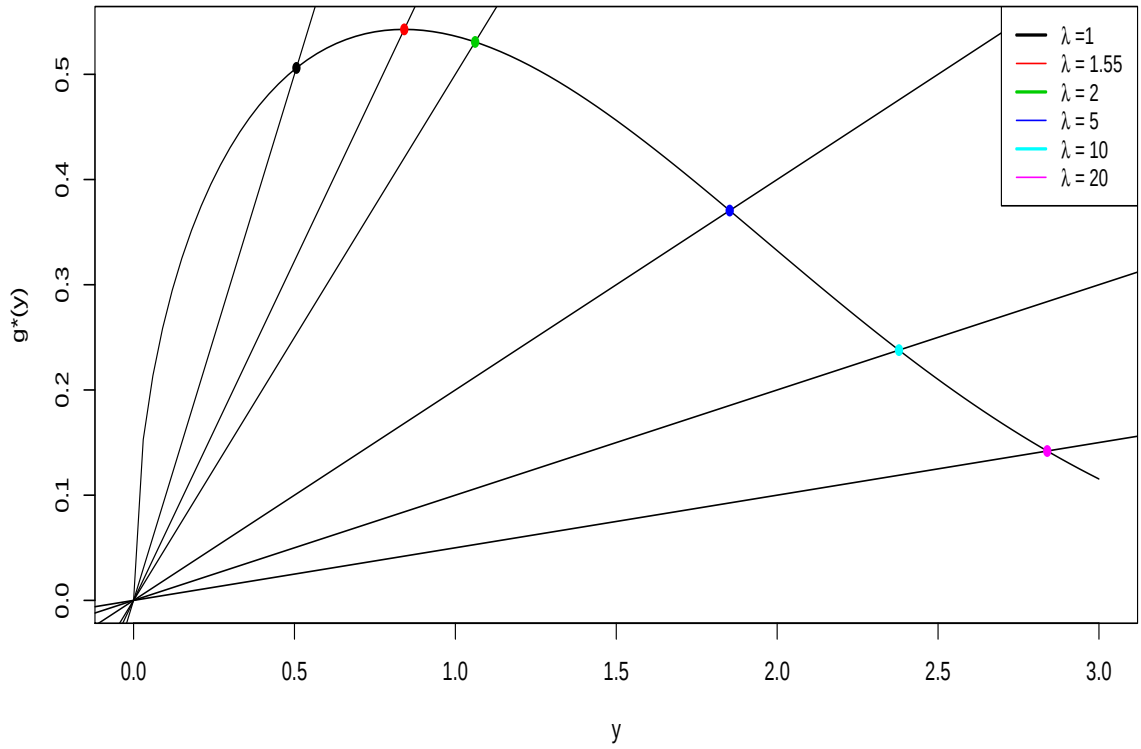


Figure 4.18: Plot of the fixed curve $g^*(y)$ along with varying lines $z = \frac{y}{\lambda}$ and their points of intersection.

4.5 Multivariate skew normal distribution

In this section we explore the mode of the multivariate skew normal distribution. Azzalini and Dalla Valle (1996) extended the concept of the univariate skew normal to introduce the multivariate skew normal distribution. The multivariate skew normal is defined as an extension of the multivariate normal distribution with additional parameters regarding the skewness of the density shape. The density of the skew normal distribution becomes the multivariate normal density when the skewness parameters are zeros. The multivariate skew normal distribution is introduced in the next definition.

Definition 4.2. *(Azzalini and Dalla Valle, 1996) Suppose $\boldsymbol{x}$ is a D-dimensional random vector, the probability density of the multivariate skew normal with mean 0 and skewness vector $\boldsymbol{\lambda}$ is given by:*

$$f_{sn}(\boldsymbol{x}; 0, \boldsymbol{\Omega}, \boldsymbol{\lambda}) = 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}), \tag{4.13}$$

where $\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})$ is the probability density function (pdf) of multivariate normal distribution with positive-definite $D \times D$ correlation matrix Ω, and $\Phi(.)$ is the cumulative distribution function (cdf) of a multivariate standard normal distribution, evaluated at $\boldsymbol{\lambda}^T\boldsymbol{x}$.

In this section we use the same notation as the univariate skew normal distribution for the multivariate skew normal distribution, namely, $\boldsymbol{x}$ denotes the vector observations of skew normal distribution, and $\boldsymbol{\lambda}$ denotes the vector of the skewness parameters of multivariate skew normal distribution.

Recall, that the probability density function (pdf) of multivariate normal with mean 0 and covariance matrix Ω is given by:

$$\phi_D(\boldsymbol{x}; 0, \boldsymbol{\Omega}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Omega}^{-1}\boldsymbol{x})\right)}, \quad -\infty < \boldsymbol{x} < \infty. \tag{4.14}$$

and the cumulative distribution function $\Phi(\boldsymbol{x})$ of multivariate skew normal distribution is defined by:

$$\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) \quad = \quad \int_{-\infty}^{\boldsymbol{\lambda}^T\boldsymbol{x}} \phi_D(t)dt \tag{4.15}$$

In order to use the multivariate skew normal distribution with real data, researchers

usually consider an affine transformation. The affine transformation is defined by

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

where $\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_D)^T$ are the location parameters, and $\boldsymbol{W} = diag(\omega_1, \omega_2, ..., \omega_D) > 0$ are the scale parameters of the multivariate skew normal distribution. The multivariate skew normal distribution, including the location and scale parameters, can be written as:

$$\boldsymbol{Y} \sim SN_D(\boldsymbol{\xi}, \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}, \boldsymbol{\lambda}), \tag{4.16}$$

where $\bar{\boldsymbol{\Omega}} = \boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}$, and $\boldsymbol{\xi} \in \mathbb{R}^D$.

Azzalini and Dalla Valle (1996) derived the mean and variance of the multivariate skew normal distribution by introducing the moment generating function of the multivariate skew normal distribution. The moment generating function of the multivariate skew normal distribution is given as follows:

$$M_{\boldsymbol{x}}(t) = 2\exp\left(\frac{t^T\boldsymbol{\Omega}t}{2}\right)\Phi(\delta^T t),$$

where $\delta = \frac{\boldsymbol{\Omega}\boldsymbol{\lambda}}{\sqrt{(1+\boldsymbol{\lambda}^T\boldsymbol{\Omega}\boldsymbol{\lambda})}}, \qquad t \in \mathbb{R}^D.$

The mean of skew normal distribution is given by the first moment of the moment generating function of multivariate skew normal distribution. The mean of the multivariate skew normal distribution, after including the location and scale parameters, is given by:

$$E(\boldsymbol{Y}) = \boldsymbol{\xi} + \boldsymbol{W}\sqrt{\frac{2}{\pi}}\left(\frac{\boldsymbol{\Omega}\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\Omega}\boldsymbol{\lambda}}}\right).$$

By using the first and second moments of the moment generating function of the

multivariate skew normal, the variance of multivariate skew normal distribution, after including the location and scale parameters, can be written as:

$$\begin{aligned} Var(\boldsymbol{Y}) &= \boldsymbol{\Omega} - \frac{2}{\pi}\left(\frac{\boldsymbol{\Omega\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\Omega\lambda}}}\right)\left(\frac{\boldsymbol{\Omega\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\Omega\lambda}}}\right)^T \\ &= \boldsymbol{\Omega} - \frac{2}{\pi}\delta\delta^T. \end{aligned}$$

where $\delta = \frac{\boldsymbol{\Omega\lambda}}{\sqrt{(1+\boldsymbol{\lambda}^T\boldsymbol{\Omega\lambda})}}$.

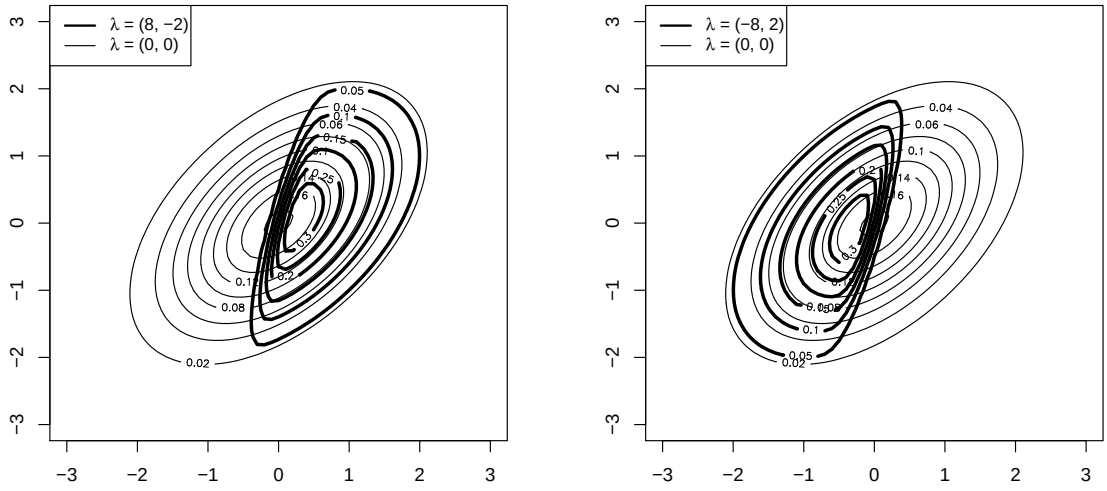


Figure 4.19: Contour plots of bivariate normal densities with different skewness parameters.

Figures 4.19 shows the contour plots for the densities of the bivariate normal distribution with different skewness parameters. The plot in the left side show the densities of the bivariate skew normal distribution with skewness parameters $\boldsymbol{\lambda} = (8, -2)$ where the density of the bivariate skew normal is skewed to the right side. The plot in the right side show the densities of the bivariate skew normal distribution with skewness parameters $\boldsymbol{\lambda} = (-8, 2)$ where the density of the bivariate skew normal is skewed to the

left side.

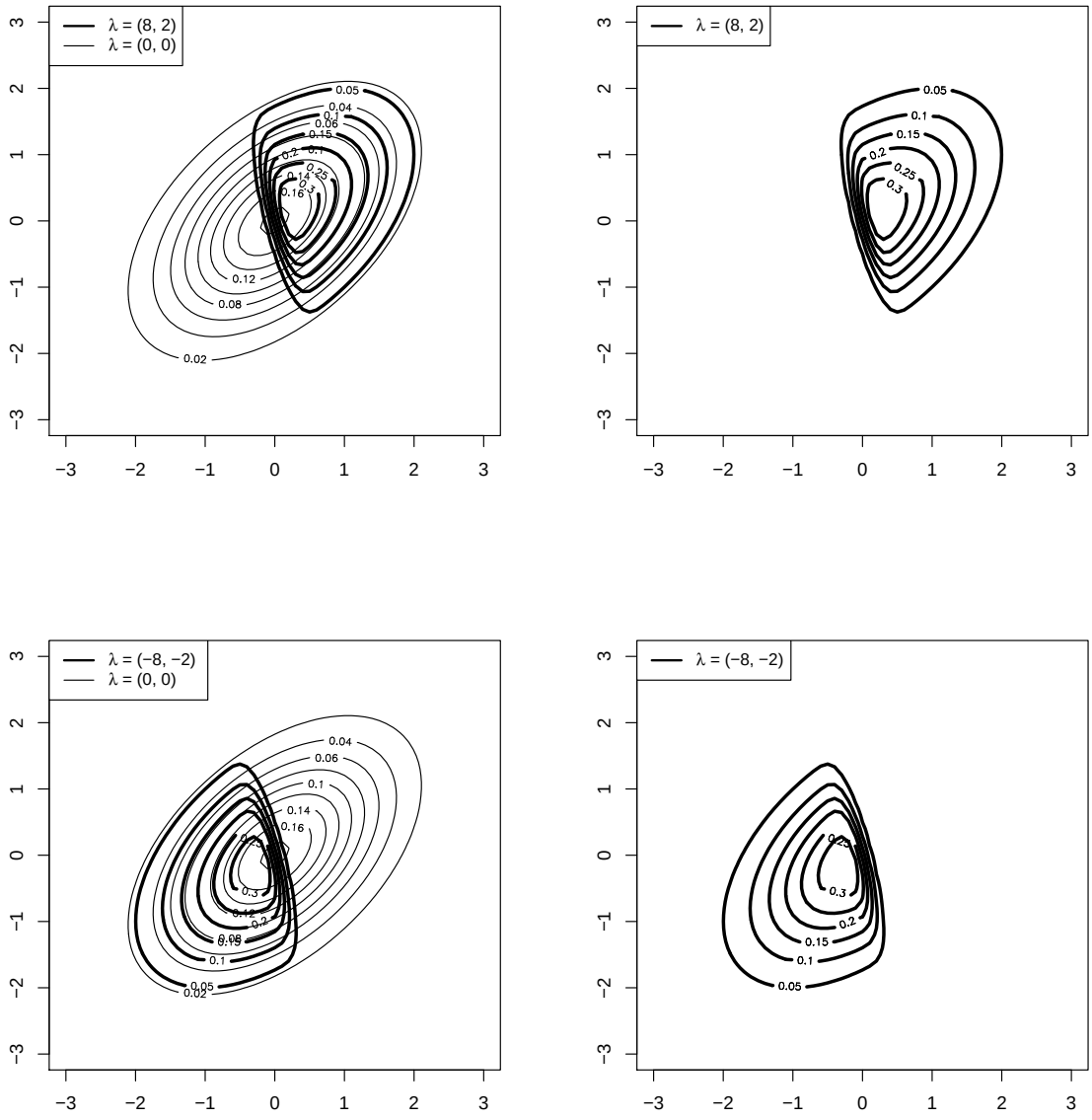


Figure 4.20: Contour plots of bivariate normal densities with different skewness parameters.

Figures 4.20 shows the contour plots for the densities of the bivariate normal distribution with different skewness parameters. The top two plots show the densities of the bivariate skew normal distribution with positive skewness parameters in both

dimensions, where the density of the bivariate skew normal is skewed towards the upper diagonal for positive skewness parameters. The light colour contour plot shows the densities of the bivariate normal in the absence of skewness. The two plots at the bottom shows the density of the bivariate skew normal distribution with the skewness parameters $\boldsymbol{\lambda} = (-8, -2)$, where the density of bivariate skew normal is skewed towards the lower diagonal.

4.5.1 Mode of multivariate skew normal distribution

Similar to the univariate case, the mode of a multivariate skew normal distribution cannot be calculated in closed form. So we need to introduce the implicit equation and provide a numerical solution to calculate the mode vector. Similar to the univariate case the mode of the multivariate skew normal distribution is the local maximum of the probability density of multivariate skew normal distribution, given in (4.13). In particular, the mode of the multivariate skew distribution is the solution to $f'_{sn}(\boldsymbol{x}) = 0$, and $f''_{sn}(\boldsymbol{x})$ is negative definite.

To obtain the mode of the multivariate skew normal distribution, we calculate the derivative of the probability density function of the multivariate skew normal distribution (4.13) as follows:

For $f_{sn}(\boldsymbol{x}) = 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi_D(\boldsymbol{\lambda}^T\boldsymbol{x})$ we can compute $f^{'}(\boldsymbol{x})$ as below:

$$\begin{aligned}
\frac{\partial}{\partial_{\boldsymbol{x}}} f_{sn}(\boldsymbol{x}) &= \frac{\partial}{\partial_{\boldsymbol{x}}} 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) \\
&= 2\phi_D^{'}(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi^{'}(\boldsymbol{\lambda}^T\boldsymbol{x}) \\
\text{For : } \phi_D(\boldsymbol{x}; \boldsymbol{\Omega}) &= \frac{1}{2\pi^{\frac{D}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(\boldsymbol{x}^T\boldsymbol{\Omega}^{-1}\boldsymbol{x})\right)} \\
\phi_D^{'}(\boldsymbol{x}; \Omega) &= (-\boldsymbol{\Omega}^{-1}\boldsymbol{x})\phi_D(\boldsymbol{x}; \boldsymbol{\Omega}) \\
\text{For : } \Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) &= \int_{-\infty}^{\boldsymbol{\lambda}^T\boldsymbol{x}} \phi_D(t)dt \\
\Phi^{'}(\boldsymbol{\lambda}^T\boldsymbol{x}) &= \boldsymbol{\lambda}\phi(\boldsymbol{\lambda}^T\boldsymbol{x}).
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{\partial}{\partial_{\boldsymbol{x}}} f(\boldsymbol{x}) &= \frac{\partial}{\partial_{\boldsymbol{x}}} 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) \\
&= 2\phi_D^{'}(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi^{'}(\boldsymbol{\lambda}^T\boldsymbol{x}) \\
&= 2(-\boldsymbol{\Omega}^{-1}\boldsymbol{x})\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\boldsymbol{\lambda}\phi_D(\boldsymbol{\lambda}^T\boldsymbol{x}) \\
&= 2[(-\boldsymbol{\Omega}^{-1}\boldsymbol{x})\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + \phi_D(\boldsymbol{x}, \boldsymbol{\Omega})\boldsymbol{\lambda}\phi_D(\boldsymbol{\lambda}^T\boldsymbol{x})] \\
&= 2\phi_D(\boldsymbol{x}, \boldsymbol{\Omega})[(-\boldsymbol{\Omega}^{-1}\boldsymbol{x})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + \boldsymbol{\lambda}\phi_D(\boldsymbol{\lambda}^T\boldsymbol{x})]
\end{aligned}$$

As $\phi_d(\boldsymbol{x}, \boldsymbol{\Omega})$ is the density function of normal distribution and $\phi_d(\boldsymbol{x}, \boldsymbol{\Omega}) \neq 0$, the mode of the multivariate of skew normal with mean 0 can be obtained by solving the implicit equation:

$$h(x) = (-\boldsymbol{\Omega}^{-1}\boldsymbol{x})\Phi(\boldsymbol{\lambda}^T\boldsymbol{x}) + \boldsymbol{\lambda}\phi_D(\boldsymbol{\lambda}^T\boldsymbol{x}) = 0. \tag{4.17}$$

The next few examples illustrate the location of the modes of bivariate skew normal distributions under different parameters. The `snm.mode` function from the github repository `skewed-normal-codes` (Alruwaili and Ray, 2019) can be used to cal-

culate the numerical solution of (4.17) for any $\boldsymbol{\lambda}$, giving us the corresponding mode vector.

Example 4.4. *Mode of bivariate skew normal distribution with the following parameters:* $\boldsymbol{\xi} = (0, 0)$, *and* $\boldsymbol{\Omega} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

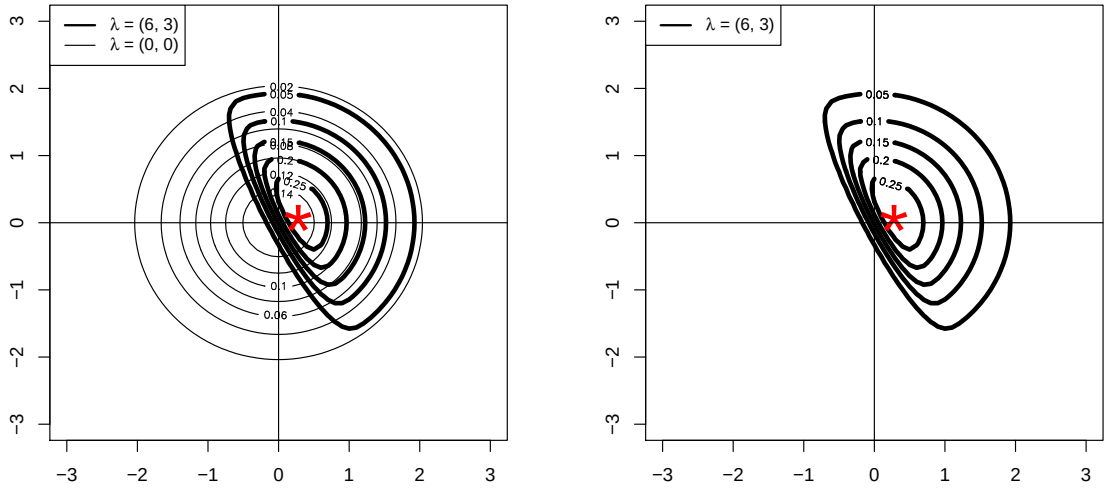


Figure 4.21: Contour plots with modes of bivariate normal density with different skewness parameters $\boldsymbol{\lambda} = (6, 3)$.

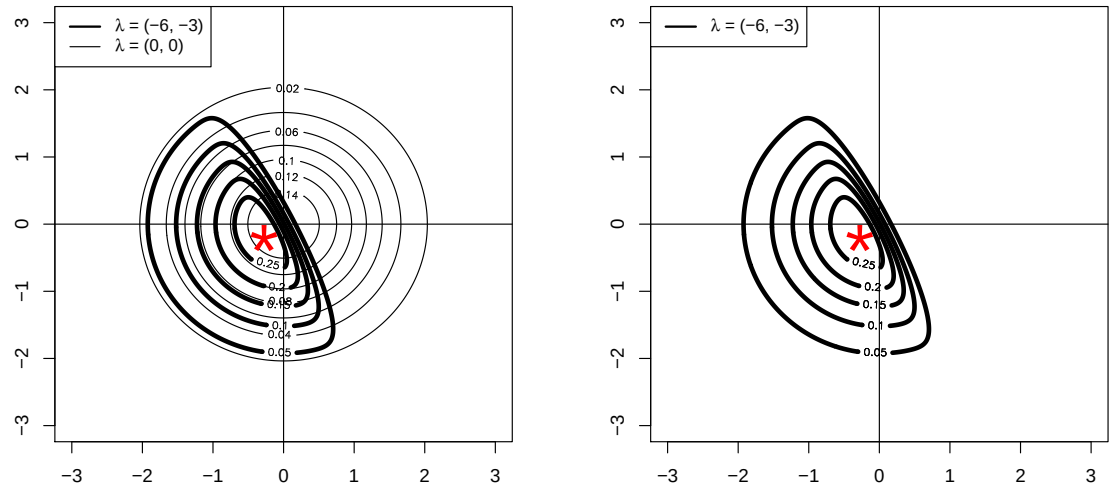


Figure 4.22: Contour plots and modes of bivariate normal density with different skewness parameters $\boldsymbol{\lambda} = (-6, -3)$.

Figures 4.21 and 4.22 show the contour plots for the parameters of example 4.4. Figures 4.21 show the density of bivariate of skew normal distribution is skewed to the right side where the skewness parameter is greater than zero, and the red asterix indicates the mode of bivariate of skew normal distribution. Figure 4.22 shows the densities of bivariate of the skew normal distribution are skewed to the left side where the skewness parameter is less than zero. The standard bivariate normal contour plots are given for comparison. In our R-package we provide a function that allows users to calculate this mode numerically.

4.6 Conclusion

In this chapter we explored the modality of univariate and multivariate skew normal distribution. For the univariate case we introduced a new way of calculating the mode of univariate skew normal distribution, and then we showed the mode of univariate skew distribution graphically. We found that the mode lies between 0 and the mean of

skew normal distribution. The mode and mean of the skew normal distribution were the same when the skewness parameter was zero, where the density of the skew normal distribution had become the normal density.

We explored the effect of the skewness parameters on the mean and mode of univariate skew normal distributions when the skewness parameter $\lambda \to \pm\infty$ and this helped us to understand the relationship between the mean and the mode of the skew normal distribution. Exploring the relationship between the mean and the mode of the univariate skew normal distribution helped us to narrow the search range for the numerical optimization for finding the mode. By exploring the mode of the skew normal distribution, we found that the modes of skew normal distributions do not move as much as the mean of the skew normal distribution with a change in the skewness parameters and the maximum values of the mode occurs for $\lambda = 1.55$.

For the multivariate case of the skew normal distribution, we explored the modality by introducing a new format to calculate the modes of multivariate skew normal distributions and to present the location of the modes of multivariate skew normal distributions graphically. Exploring the mode of the univariate and multivariate is important as this mode is one of the measures of central tendency used to describe the dataset.

The task for the immediate future in this area is to find a bound for the mode of a multivariate skew normal distribution, and to generalize the concepts of finding the mode of a skew normal distribution to explore the modality of univariate and multivariate skew $t-$distributions and to study the relationship between the mean and the mode of univariate and multivariate skew $t-$distributions.

Chapter 5

Modality of Mixtures of Multivariate Skew Normals

In this chapter we explore the modality of the multivariate skew normal mixture. Finite mixture models are used as a tool for modelling heterogeneous population, and the concept of mixture models and their application has been introduced in many books, such as McLachlan and Peel (2000) and Titterington et al. (1985). In recent years skew normal mixtures have been used for analysing data ranging from applications in biology and remote sensing data. We will present some fundamental results on the modality of skew normal mixtures and show how they can be used to provide a more robust explanation of clustering by merging components of skew normal fits that are unimodal in the data.

NOTE: This chapter is also adapted from: **Alruwaili B. and Ray S.** Topography of skew normal distributions and skew normal mixtures. *Under preparation* (2019).

5.1 Skew normal mixtures

The skew normal distribution was introduced by Azzalini (1985) as an extension of the normal distribution but with an extra skewness parameter, and this skewness parameter gives the density the flexibility to allow asymmetric distributions. Azzalini and Dalla Valle (1996) generalized the concept of univariate skew normal distributions to multivariate skew normal distributions. The concept of the skew normal distribution has been further generalised to univariate and multivariate skew t-distributions (Jones, 2001; Gupta, 2003).

Due to the skewness parameters, the multivariate skew normal distribution and multivariate skew t-distribution show more flexibility for modelling skewed data-sets than regular multivariate normal and t-distributions. After the concepts of multivariate skew normal and multivariate skew t-densities became available, the concepts of multivariate skew normal and skew t-densities were used with the finite mixture models as tools for modelling heterogeneous data with in asymmetric features (Lee and Mclachlan, 2013). The mixture of skew normal densities and the mixture of skew t-densities are applied in different areas of research, such as medicine, finance, sciences, and biology (Pyne et al., 2009; Soltyk and Gupta, 2011; Riggi and Ingrassia, 2013; Contreras-Reyes and Arellano-Valle, 2012).

In this section we will explore only the modality of the mixture of skew normal densities. We will first give an overview of the previous research on the mixture of skew normal densities. Lin et al. (2007) started with using maximum likelihood and Bayesian methods to estimate the parameters of the mixture of skew normals. Lin (2009) developed an EM algorithm to find the maximum likelihood estimate for the mixture of multivariate normal densities parameters. More theoretical results on the shape mixtures of the multivariate skew normal and the scale mixtures of the skew normal densities were given by (Vernic (2005); Arellano-Valle et al. (2009); Basso et al.

(2010); Kim and Zhao (2018) and (Arellano-Valle et al., 2018)).

Frühwirth-Schnatter and Pyne (2010) developed a new method based on Bayesian inference for univariate and multivariate skew normal mixtures. Cabral et al. (2012) presented some results on exploring the multivariate mixture models including the skew normal independent densities. Prates et al. (2013) provided an R-package named by `mixsmsn` package for fitting the univariate and multivariate finite mixtures of skew normal densities. Following this Lin et al. (2014) provided a new mixture model by using the multivariate skew t-normal densities. Lee et al. (2018) introduced another approach, using an EM algorithm to fit mixtures of multivariate skew normal densities and the mixture of multivariate t-densities. More results on exploring the theoretical mixture of the multivariate skew normal densities are summarised in Bernardi (2013) and Arslan (2015).

To our knowledge none of the papers discussed in the previous paragraph deals with the modes of mixtures of skew normal densities. In this chapter we introduce new tools to explore the modes of the mixture of multivariate skew normal densities. These tools are built on the concept of the ridgeline manifold and the height of the ridgeline manifold that were developed for the mixtures of normals and t's discussed in chapter 2 and 3.

5.2 Ridgeline function for the mixture of multivariate skew normal densities

In this section we generalize the results of Ray and Lindsay (2005) for exploring the modality of the mixture of multivariate skew normal densities. We will generalize the concepts of the ridgeline function and the height of ridgeline function for the mixture of multivariate skew normal densities. We start by introducing the ridgeline function

and the plot of the height of the ridgeline function of the mixture of multivariate skew normal densities. The ridgeline function and the height of the ridgeline function will help us to calculate the number of modes for the mixture of multivariate skew normal densities. By using the general definition of the mixture model (1.1), the probability density function of the mixture of multivariate skew normal densities can be given as:

$$g(\boldsymbol{x}) = \sum_{i=1}^{K} \pi_k \, f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\big), \qquad \boldsymbol{x} \in \mathbb{R}^D, \tag{5.1}$$

where K is the number of the components in the mixture of D-dimensional skew normal densities, π_i are the mixing proportions for the K components, where $\pi_i \in [0, 1]$, $\pi_1 + \ldots + \pi_K = 1$, and $f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\big)$ is the density function of a multivariate skew normal distribution with location $\boldsymbol{\xi}_i$, and variance $\boldsymbol{\Omega}_i$ and skewness vector $\boldsymbol{\lambda}_i$. For ease of notation we call $f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\big)$ by $f_{sn}(\boldsymbol{x})$.

In this section we will focus on the modality of a two component mixture of skew normals, whose density can be written as:

$$g(\boldsymbol{x}) = \pi_1 \, f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_1, \boldsymbol{\Omega}_1, \boldsymbol{\lambda}_1\big) + \pi_2 \, f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_2, \boldsymbol{\Omega}_2, \boldsymbol{\lambda}_2\big), \qquad \boldsymbol{x} \in \mathbb{R}^D \tag{5.2}$$

To introduce the ridgeline function of the mixture of two multivariate skew normals, we generalize the concept of the ridgeline function of two multivariate normal densities (Ray and Lindsay, 2005). The ridgeline of the two component mixture of multivariate skew normal densities is a curve that starts at the mode of the first component and ends at the mode of the second component, and this curve provides all important information relating to the mixture density, such as the critical points (i.e. the maximum and minimum points, and the saddle point), the location and number of modes. For K-components, the ridgeline function is given by the first derivative of the mixture of K

multivariate skew normal distribution (5.2) and can be written as:

$$\nabla\, g(\boldsymbol{x}) \;=\; \sum_{i=1}^{K} \pi_i\, \nabla f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big). \tag{5.3}$$

$$\begin{aligned} \text{As}\quad \nabla \log f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_k,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big) &= \frac{1}{f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)} \nabla f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big) \\ \text{we have}\quad \nabla f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big) &= f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\, \nabla \log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\Big). \end{aligned}$$

Now, the derivative of $\nabla\, g(\boldsymbol{x})$ given in (5.3) can be written as follows

$$\nabla\, g(\boldsymbol{x}) \;=\; \sum_{i=1}^{K} \pi_i\, f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\, \nabla \log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\Big). \tag{5.4}$$

or equivalently

$$\sum_{i=1}^{K} \pi_i\, \frac{f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)}{g(\boldsymbol{x})}\, \nabla \log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\Big). \tag{5.5}$$

Recall the simplex defined for normal mixture in Ray and Lindsay (2005). When $K > 2$ we will have a set of $\alpha_i' s$ lying in a K-dimensional simplex, i.e. $\sum_{i=1}^{K} \alpha_i = 0$, where $\alpha \in [0, 1]$ is the range of the ridgeline function. Based on the same simplex we define the new manifold $\mathcal{M}_{sn}$ by:

$$\mathcal{M}_{sn} = \Big\{\boldsymbol{x} \in \mathbb{R}^D :\ \exists\ \alpha \in S_{K-1} :\ \sum_{i=1}^{K} \alpha_i\, \nabla \log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\Big) = 0\Big\},$$

where $S_{K-1} = \big\{\alpha = (\alpha_1, \ldots, \alpha_i) \in \mathbb{R}^K :\ \alpha_i \geq 0,\ \alpha_1 + \ldots + \alpha_i = 1\big\}$.

Theorem 5.1. *For the mixture of multivariate skew normal densities, $g(\boldsymbol{x})$, all key features of the mixture density $g(\boldsymbol{x})$, such as critical points, saddle points, number and*

location of modes, are points in $\mathcal{M}_{sn}$.

Proof. Similar to proof of the corresponding manifold for normal mixtures given in Ray and Lindsay (2005) and by showing that

$$\sum_{i=1}^{K} \frac{\pi_i f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i)}{g(\boldsymbol{x})} = 1. \tag{5.6}$$

Thus α_is defined by $\alpha_i = \dfrac{\pi_i f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i)}{g(\boldsymbol{x})}$ is an element in the simplex S_{k-1}

□

Based on theorem 5.1 the ridgeline function of the mixture of multivariate skew normal mixture can be written as

$$x^*(\alpha) = \sum_{k=1}^{K} \alpha_i \nabla \log \Big(f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i) \Big), \tag{5.7}$$

where $\alpha \in S_{k-1}$. Note that, for the normal mixture, the ridgeline function had a closed form expression given by (2.5). But for the skew normal mixture we need to numerical obtain $x^*(\alpha)$ for each α as we do not have explicit equation for (5.7). We have created the `ridgeline.skew` function Alruwaili and Ray (2019) to numerically evaluate the $x^*(\alpha)$ for any given set of parameters. For the ridgeline function of the mixture of multivariate skew normal mixture, we first calculate the $\nabla \log \Big(f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i) \Big)$ including the linear transformation (location and scale).

Lemma 5.2. *The derivative of* $\log \Big(f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i) \Big)$ *is given by*

$$\nabla \log \Big(f_{sn}(\boldsymbol{x}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\lambda}) \Big) = -\boldsymbol{\Omega}^{-1}(\boldsymbol{x} - \boldsymbol{\xi}) + \boldsymbol{\lambda} \frac{\phi(\boldsymbol{\lambda}^T(\boldsymbol{x} - \boldsymbol{\xi}))}{\Phi(\boldsymbol{\lambda}^T(\boldsymbol{x} - \boldsymbol{\xi}))}.$$

Proof.

$$\begin{aligned}
\nabla \log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda}\big)\Big) &= \nabla\log(2) + \nabla\log\phi_d(\boldsymbol{x};\boldsymbol{\xi},\boldsymbol{\Omega}) + \nabla\log\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)\\
\text{where } \nabla\log(2) &= 0,\\
\nabla\log\phi_d(\boldsymbol{x};,\boldsymbol{\xi},\boldsymbol{\Omega}) &= \nabla\log\frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Omega}|^{\frac{1}{2}}} + \nabla\log e^{\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\xi})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\right)}\\
&= 0 + \nabla\big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\xi})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi})\big)\\
&= -\boldsymbol{\Omega}^{-1}(\boldsymbol{x}-\boldsymbol{\xi}),\\
\text{and}\quad \nabla\log\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big) &= \frac{\nabla\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}{\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}\\
&= \boldsymbol{\lambda}\frac{\phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}{\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}.
\end{aligned}$$

Then we get

$$\nabla\log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda}\big)\Big) = -\Omega^{-1}(\boldsymbol{x}-\boldsymbol{\xi}) + \boldsymbol{\lambda}\frac{\phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}{\Phi\big(\boldsymbol{\lambda}^T(\boldsymbol{x}-\boldsymbol{\xi})\big)}.$$

□

Using lemma 5.2 for fixed $\alpha \in S_{K-1}$, the ridgeline function of the mixture of multivariate skew normal densities including the location and scale parameters is given as:

$$\begin{aligned}
x^*(\alpha) &= \sum_{i=1}^{K}\alpha_i\nabla\log\Big(f_{sn}\big(\boldsymbol{x};\boldsymbol{\xi}_i,\boldsymbol{\Omega}_i,\boldsymbol{\lambda}_i\big)\Big)\\
&= \sum_{i=1}^{K}\alpha_i\Big(-\boldsymbol{\Omega}_i^{-1}(\boldsymbol{x}-\boldsymbol{\xi}_i) + \boldsymbol{\lambda}_i\frac{\phi\big(\boldsymbol{\lambda}_i^T(\boldsymbol{x}-\boldsymbol{\xi}_i)\big)}{\Phi\big(\boldsymbol{\lambda}_i^T(\boldsymbol{x}-\boldsymbol{\xi}_i)\big)}\Big).
\end{aligned}$$

In particular, the ridgeline function of the mixture of univariate skew normal distribution reduces to

$$x^*(\alpha) = \sum_{i=1}^{K} \alpha_i \Big(-\frac{x-\xi_i}{\sigma_i} + \lambda_i \frac{\phi\big(\lambda_i(x-\xi_i)\big)}{\Phi\big(\lambda_i(x-\xi_i)\big)} \Big), \qquad \text{where} \quad \alpha \in S^{K-1}.$$

The next step in our study on the modality of the mixture of skew normal densities is to consider the height of the ridgeline function.

Height of the ridgeline function

Similar to the definition for the normal mixtures in Chapter 2, the height of the ridgeline function is defined as the height of the density along the $x^*(\alpha)$ curve and is given by:

$$h(\alpha) = g(x^*(\alpha)). \tag{5.8}$$

Similar to normal mixture, the height plot will be a useful tool to determine the number of modes in any dimensions, even if the contour plots are not available.

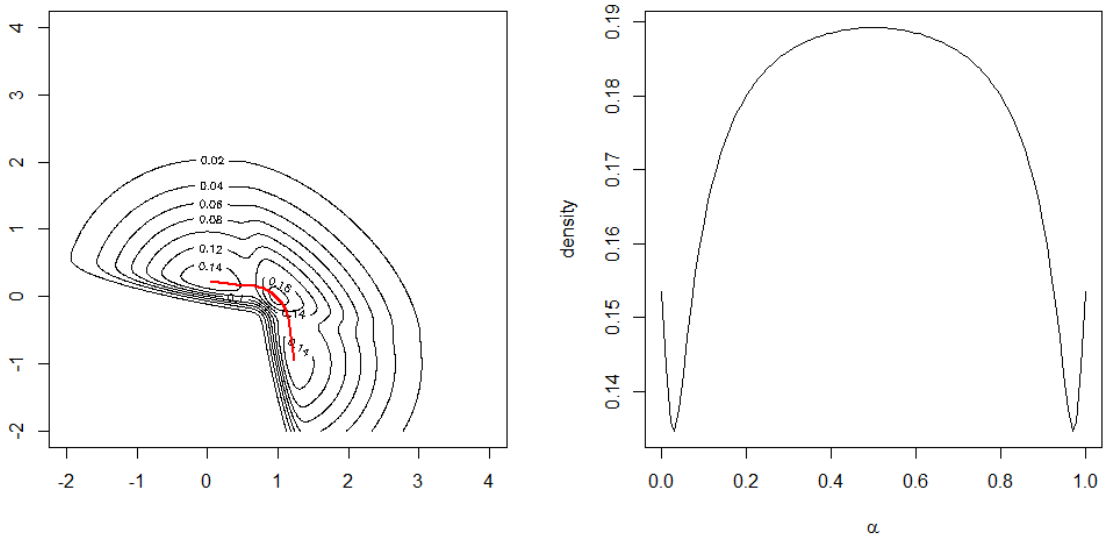
5.3 Plotting the ridgeline function and the height of the ridgeline function

In this section we provide some illustrative examples for exploring the modes of the mixture of two multivariate skew normal densities by using the ridgeline function and the height of the ridgeline function. The aim of providing these examples is to illustrate how the ridgeline function and the height of the ridgeline function can be used as an important tool to explore the number of modes for the mixture of two components with particular focus on the skewness parameters. The ridgeline function in each case is numerically calculated using the `ridgeline.skew` function given in our github repository `skewed-normal-codes` (Alruwaili and Ray, 2019).

Examples of mixing two skew normal components, in two dimensions giving three modes:

Example 5.1. *(Two components, equal variance, three modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 2 \\ 10 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



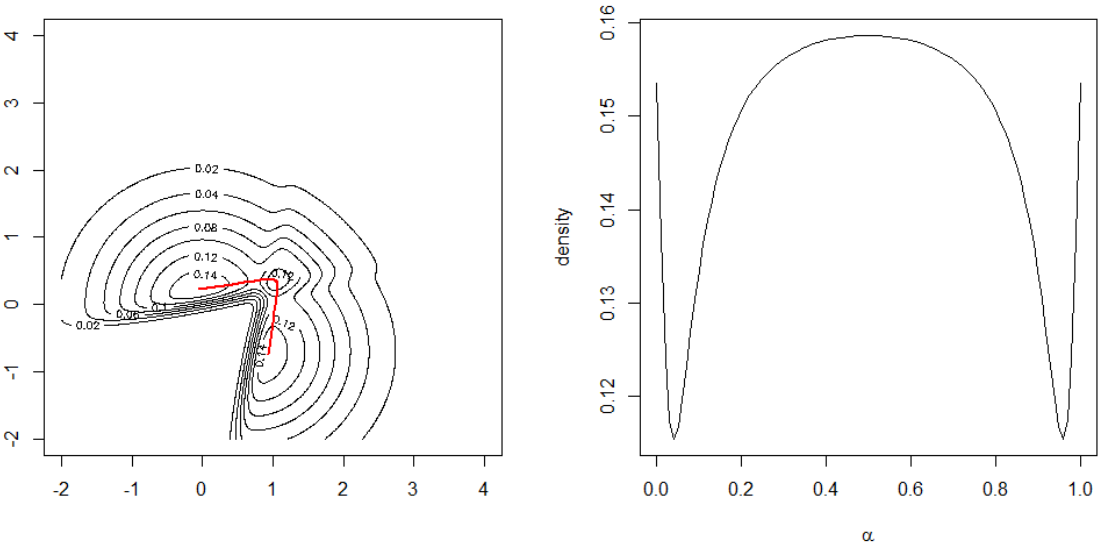
(a) (b)

Figure 5.1: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.1. (b) The height of the ridgeline function for the same mixture density.

Figure 5.1 shows that there are three modes visible in the contour plot located at $x = (0, 0)$, $x = (1, 0)$, and $x = (1, -1)$, and these three modes are present in the height plot at $\alpha = 0$, 0.5, and 1.

Example 5.2. *(Two components, equal variance, three modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 0.7 \\ -0.7 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} -2 \\ 10 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



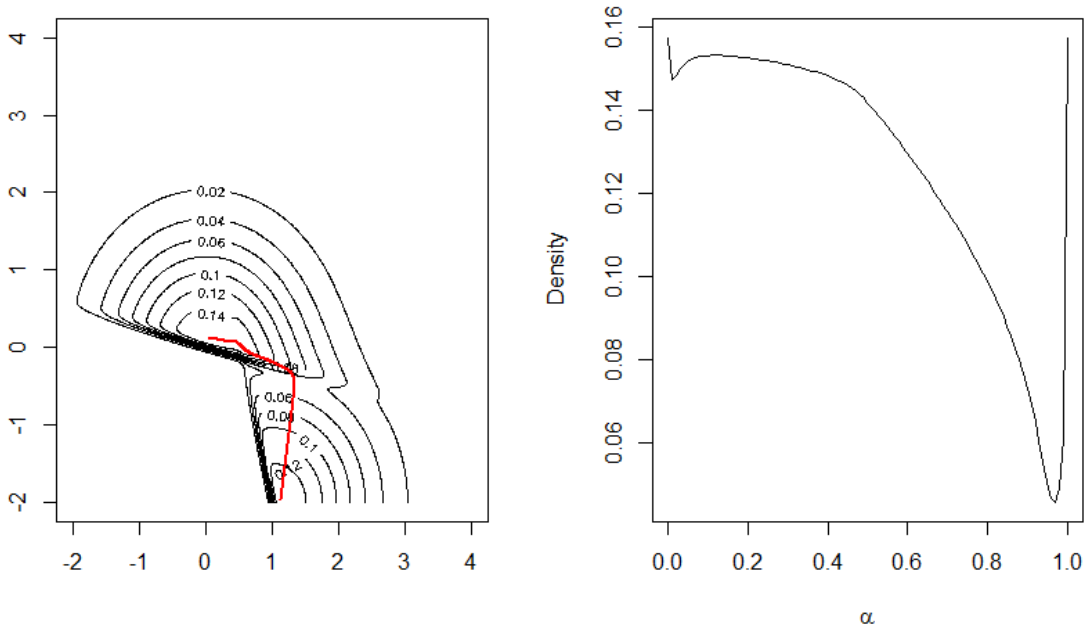
(a) (b)

Figure 5.2: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.2. (b) The height of the ridgeline function for the same mixture density.

Figure 5.2 shows that there are three modes visible in the contour plot, and the ridgeline height plot has three local maxima's at $\alpha = 0$, 0.5, and 1.

Example 5.3. *(Two components, equal variance, three modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 5 \\ 20 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



(a) (b)

Figure 5.3: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.3. (b) The height of the ridgeline function for the same mixture density.

Figure 5.3 shows that there are two modes visible in the contour plot, but the third mode is not clearly visible. The third modes is visible in the ridgeline height plot near $\alpha = 0, 15$.

Explanation of the Figures for Examples 5.1, 5.2, and 5.3:

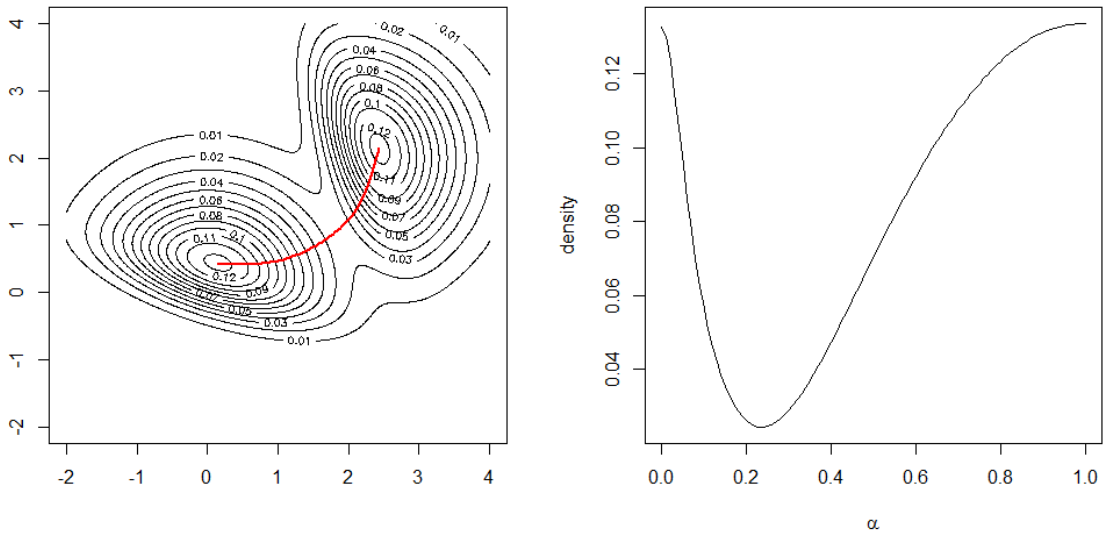
For examples 5.1, 5.2, and 5.3 plots (a) show the contour plots with the curve of the ridgeline function of these examples. Plots (b) show the height of the ridgeline function of the examples. For example, the plots of the ridgeline function and the height of the ridgeline function for the mixture density, given in Example 5.1, shows that: (a) the contour plot with the ridgeline curve shows that there are three modes and the curve of the ridgeline function passes from the mode of the first component to the mode of the second component through the third mode; (b) the plot of the height of the ridgeline function shows that α is one dimensional. There are three local maxima, giving the three modes of mixture density in this plot. These modes are visible near $\alpha = 0$, 0.5, and 1. The mode at $\alpha = 0$ corresponds to the mode at (0, 0), the mode at α =0.5 corresponds to the mode at the intersection of the two mixtures at (1, 0) and finally, the mode at α =1 corresponds to the mode at the other mean at (1, -1).

By exploring the mixtures in examples 5.1, 5.2, and 5.3 with different values of skewness parameters, we found that the total number of modes for these examples was three modes.

Examples on mixing two skew normal components, in two-dimensions giving two modes:

Example 5.4. *(Two components, equal variance, two modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



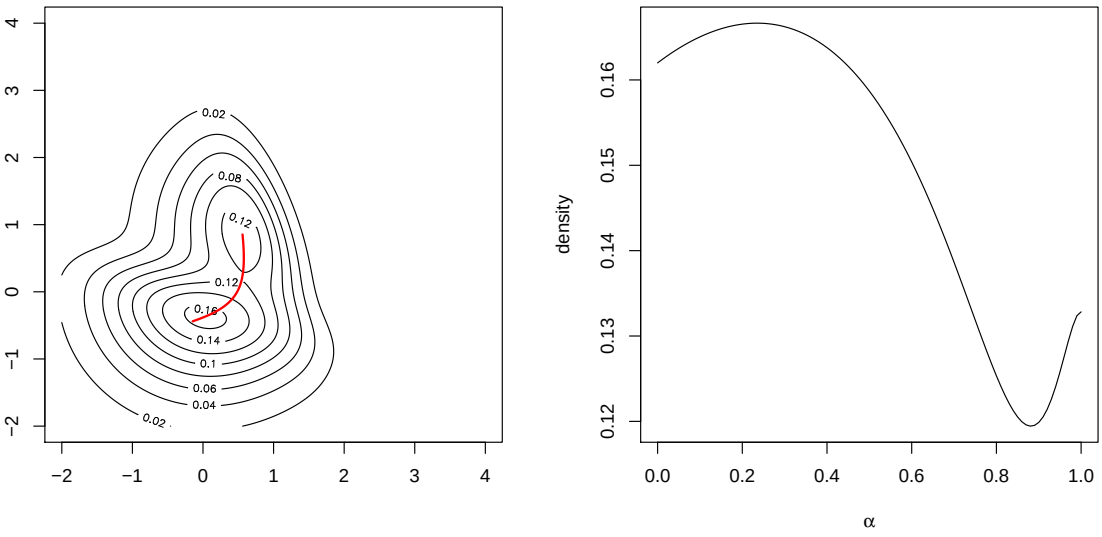
(a) (b)

Figure 5.4: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.4. (b) The height of the ridgeline function for the same mixture density.

Figure 5.4 shows that there are two modes visible in the contour plot at $x = (0, 0)$, and $x = (2.5, 2.5)$, and these two modes are present in the height plot at $\alpha = 0$ and 1.

Example 5.5. *(Two components, equal variance, two modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} -1 \\ -3 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



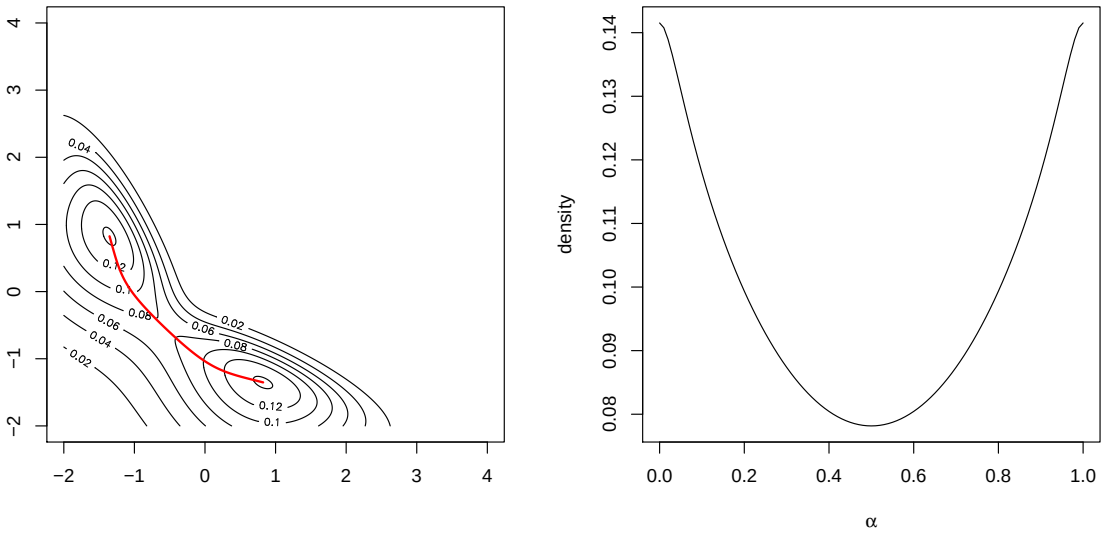
(a) (b)

Figure 5.5: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.5. (b) The height of the ridgeline function for the same mixture density.

Figure 5.5 shows there are two mode visible in the contour plot, and these two mode are located in the height ridgeline plot at α=0 and 1.

Example 5.6. *(Two components, equal variance, two modes)By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} -2 \\ -4 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$



(a) (b)

Figure 5.6: (a) The contour plot with the ridgeline curve for the mixture density given in Example 5.6. (b) The height of the ridgeline function for the same mixture density.

Figure 5.6 shows that there are two modes visible in the contour plot, and these two modes are located in the height ridgeline plot at α=0 and 1.

Explanation of the Figures of Examples 5.4, 5.5, and 5.6:

The figures of Examples 5.4, 5.5, and 5.6 show the ridgeline function and the height of the ridgeline function for the mixture densities of these examples. Plots (a) show the contour plots with the curve of the ridgeline function of the mixture densities of

these examples, and plots (b) the height of the ridgeline function of the examples. For example, the plots of the ridgeline function and the height of the ridgeline function for the mixture density given in Example 5.4 shows that (a) the contour plot with the ridgeline curve shows that there are two modes and the curve of the ridgeline function passes from the mode of the first component to the mode of the second component, and (b) the plot of the height of the ridgeline function shows that α is one-dimensional. There are two local maxima, giving the two modes of mixture density in this plot. These modes are visible near $\alpha = 0$ and 1. The mode at $\alpha = 0$ corresponds to the mode at (0,0), the mode at α=1 corresponds to the mode at the intersection of the two mixtures at (2.5,2.5).

Examples of mixing two skew normal components, in three-dimensions giving four modes: Unlike the two dimensional examples, we do not have a contour plot any more. We need to based our decision on the ridgeline elevation.

Example 5.7. *(Two components, equal variance, four modes) By considering the mixture of multivariate skew normal densities with $D = 3$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ 1.559 \\ 1.559 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} -2.2 \\ -6.5 \\ 10 \end{pmatrix},$$

$\pi_1 = \pi_2 = 0.5$.

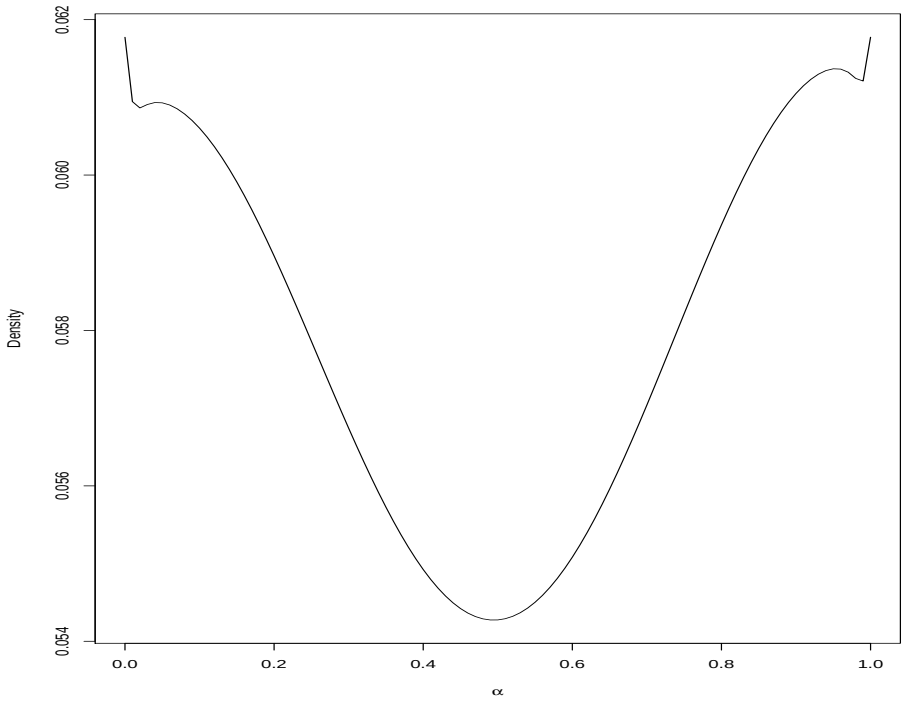


Figure 5.7: Plot of the height of the ridgeline function for the mixture density given in Example 5.7.

As the contour plots are not available to present the number of modes for three-dimensional data, the plot of the height of the ridgeline function shows the ability to

present all the modes of the mixture density of Example 5.7 as shown in Figure 5.7 and the total number of modes is four.

Example 5.8. *(Two components, equal variance, four modes) By considering the mixture of multivariate skew normal densities with $D = 3$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} -1 \\ -1.559 \\ -1.559 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \boldsymbol{\lambda} = \begin{pmatrix} 2.2 \\ 6.5 \\ -10 \end{pmatrix},$$

$\pi_1 = \pi_2 = 0.5$.

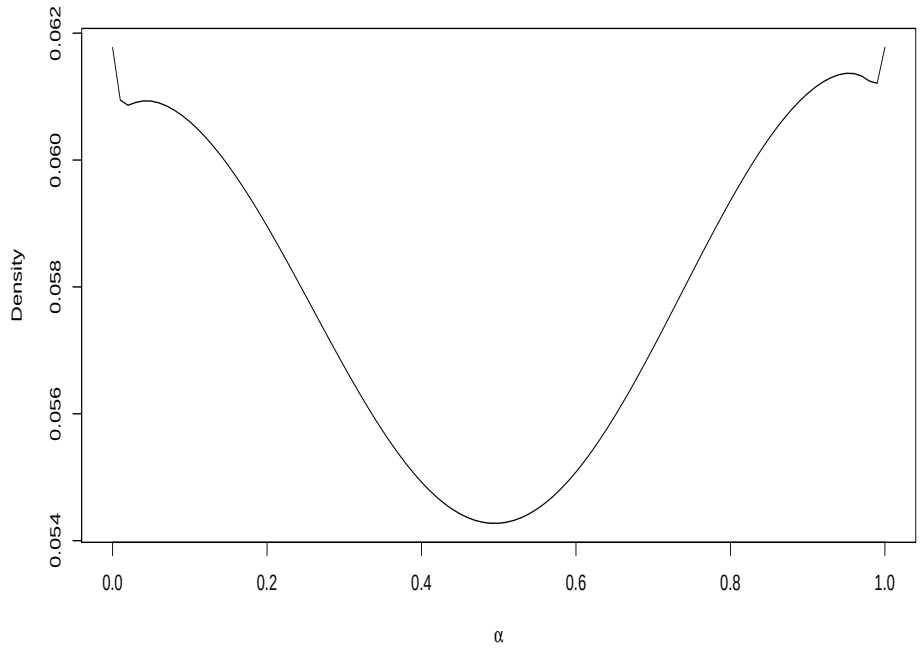


Figure 5.8: Plot of the height of the ridgeline function for the mixture density given in Example 5.8.

Figure 5.8 shows the plot of the height of the ridgeline function for the mixture density given in Example 5.8, and the total number of modes is four.

Explanation of the Figures of Examples 5.7, and 5.8:

As the contour plots are not available to present the number of modes when there are more than two-dimensions, the height of the ridgeline function shows the ability to present the modes of the multivariate of skew normal densities for a higher number of dimensions. The height plots of examples 5.7 and 5.8 show the number of modes for the mixture of skew normal densities with two components and three dimensions. The total number of modes for the density mixtures of examples 5.7, and 5.8 is four.

By exploring the mixture of two components of skew normal densities in two and three dimensions, we found that the ridgeline function and the height of the ridgeline function have the ability to present the number of modes for the mixture of skew normal densities in two and three dimensions and with any values for the skewness parameters. From exploring the modality of the mixture of skew densities with two components in two and three dimensions and different values of skewness parameters, we observe that, the maximum number of modes that we have seen is three modes in two dimensions, and four for three dimensions. These observations led us to ask the following question: Is the maximum number of modes for two components skew normal mixture in any dimensions the same as the maximum number for a two component normal mixture in the same dimensions. This is an open question and we are unaware of any other research addressing this problem. Our conjecture is that like the normal mixture the maximum number of modes for a D-dimensional mixture of two skew normal components will be $D + 1$.

5.4 Systematic study of change in modes with change in skewness parameters

We now provide a systematic study to explore the relationship between the number of mode and the skewness parameters in a two component skew-normal mixtures.

Example 5.9. *(Two components, equal variance) We consider the mixture of multivariate skew normal densities with $D = 2$, and $K = 2$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \pi_1 = \pi_2 = 0.5.$$

Table 5.1 summarises the number of modes for changing skewness parameters of the components, while keeping the means, variances and mixing proportions same. For these set of parameters we notice that we get one or two modes. When the skewness is small and or 0 we get only one mode. But we are able to increase the number of modes to 2 by moving the skewness parameter away from 0 in either of both components.

Table 5.1: Number of modes for the mixture density, given in Example 5.9, for changing skewness parameters of the two components

λ2 \ λ1	5	4	3	2	1	0	-1	-2	-3	-4	-5
5	2	2	2	2	2	2	2	2	2	2	2
4	2	2	2	2	2	2	2	2	2	2	2
3	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	1	2	2	2	2	2
1	2	2	2	2	1	1	2	2	2	2	2
0	2	2	2	1	1	1	1	1	2	2	2
-1	2	2	2	2	2	1	1	2	2	2	2
-2	2	2	2	2	2	1	2	2	2	2	2
-3	2	2	2	2	2	2	2	2	2	2	2
-4	2	2	2	2	2	2	2	2	2	2	2
-5	2	2	2	2	2	2	2	2	2	2	2

5.5 Conclusion

In this chapter we explored the modality of the mixture of multivariate skew normal densities by introducing new tools, namely the ridgeline function and the height of the ridgeline function; these two tools helped to get the number of the modes for the mixture of skew normal distribution. We introduced the ridgeline function of the mixture of multivariate skew normal densities as a curve that passes from the mode of the first component to the mode of the second component through all the available modes, where all the information of the components' mixture, such as the number and location of modes for the density, critical points, and saddle-points, lie on the curve of the ridgeline function.

Furthermore, we also explored the number of the modes for the mixture of multivariate skew normals by introducing the height of the ridgeline function. The height of the ridgeline function plot shows all the modes of the density mixture along the ridgeline curve. The height of the ridgeline function plot also gives the ability to introduce the number of the modes of the mixture of multivariate skew normal densities for more than two dimensions, where the contour plots are not available to present the modes for more than two dimensions.

After we introduced the concepts of the ridgeline function and the height of the ridgeline function for the mixture of multivariate skew normal densities, we examined the ridgeline function and the height of the ridgeline function for different examples and then considered the different values of the skewness parameters. The ridgeline function and the height of the ridgeline function showed the ability to present all the modes for the mixture of multivariate skew normal densities with different values of skewness parameters. Unfortunately, as we do not have a explicit experssion for $x^*(\alpha)$ we cannot explore the number of modes analytically for the multivariate skew mixture.

By using plots of the ridgeline function and the height of the ridgeline function we

explore the number of modes for the mixture of multivariate skew normal densities. We observed that the largest number of modes that we have from mixing two skew components in two dimensions was three, and the largest number of modes for mixing two skew components in three dimensions was four. These results raised the following question in this research area: What is the maximum number of modes that one can get from mixing two skew normal components in D dimensions? In the next chapter we will consider the concepts of the ridgeline function and the height of the ridgeline function to create a framework (R-code) for modelling skewed data, where the concepts of the ridgeline function and the height of the ridgeline function will help to decide when it is possible to merge two components together to give one cluster.

Future work in this area will involve the generalization of this concept of the ridgeline function and the height of the ridgeline function to study the modality of the mixture of multivariate skew t-distribution.

Chapter 6

Tools for Merging Skew Normal Mixture Components

One of the primary applications of mixture models is clustering. The general definition of a cluster is to find homogeneous separated groups in the dataset (Wilks, 2011; Kaufman and Rousseeuw, 2009). In the previous chapter we introduced the concepts of the ridgeline function and the ridgeline elevation plot to explore the modes of the mixture of multivariate skew normal distributions. In this chapter we will use those tools to create a computational framework, which will then be used to decide when it is possible to merge components of multivariate skew normal mixture together to form homogeneous groups based on the number of modes.

NOTE: This chapter is also adapted from: **Alruwaili B. and Ray S.** Tools for merging skew-normal mixtures and its application to flow cytometry data analysis *Under preparation* (2019).

6.1 Motivation for developing a tool for merging components of mixtures

We first present a motivation for developing the tools for merging components resulting from a mixture fit for the purpose of clustering. In this chapter we will refrain from providing a detailed discussion on individual methods for performing cluster analysis. We will only focus on clustering obtained using the mixture of multivariate skew normal densities. Usually the cluster labels are associated with the different components of the mixture, but they may not be separate enough to be labelled as separate clusters. In this chapter we therefore introduce the ridgeline function and the ratio of the ridgeline as tools to find natural groupings in the data by merging components of the mixture of multivariate skew normal densities. In the context of normal mixtures the same technique has been used by Hennig (2010a) based on the ridgeline elevation plots of Ray and Lindsay (2005). Methods for merging gaussian mixtures have also been proposed by Li (2005) and Baudry et al. (2010), and a nice summary for merging gaussian components is available in Hennig (2010b). In particular Baudry et al. (2010) presents a entropy based method for merging gaussian components and provides a piecewise linear regression fit to the rescaled entropy plot for automatic selection of the number of merged components.

We start by illustrating that the ridgeline elevation plot and the ratio of the ridgeline are useful tools for merging the components of the multivariate skew normal mixture. We will first introduce a real dataset on flow cytometry analysis that was introduced in Ray and Pyne (2012). Similar datasets has been previously clustered using mixtures of skew-normals by Pyne et al. (2009), but it remained unclear whether each component corresponded to separate clusters. We will use the ridgeline function and the ratio of the ridgeline on pairs of components to decide when it is possible to merge the components

of this mixture.

6.2 Description of flow cytometry data

Flow cytometry is one of the most common imaging technique used in medical diagnostics and in research labs to identify and diagnose types of cells and their function in a sample of the cell population. Flow cytometry is usually used to diagnose cancer or to observe immune response against pathogens. More details on flow cytometry data and the preprocessing of flow data are given by Pyne et al. (2009), Lo et al. (2009), Scheuermann et al. (2009), Hahne et al. (2009), and Finak et al. (2010).

More details on the biological background of the flow cytometry data that we will use are given in Ray and Pyne (2012). Pyne et al. (2009) and Ray and Pyne (2012) have used the Forward Scatter (FSC) and Side Scatter (SSC) to isolate dead cells from live cells. Figure 6.1 shows the skewed nature of the of two clusters. Though Ray and Pyne (2012) used a non-parametric approach, Pyne et al. (2009) argued that a skew-normal and skew-t mixture might model the clusters accurately. The skewed distribution gives the density shape more flexibility to include the non-ellipsoidal features and the outlier cells of the flow cytometry data, but choosing the number of clusters remained a difficult problem.

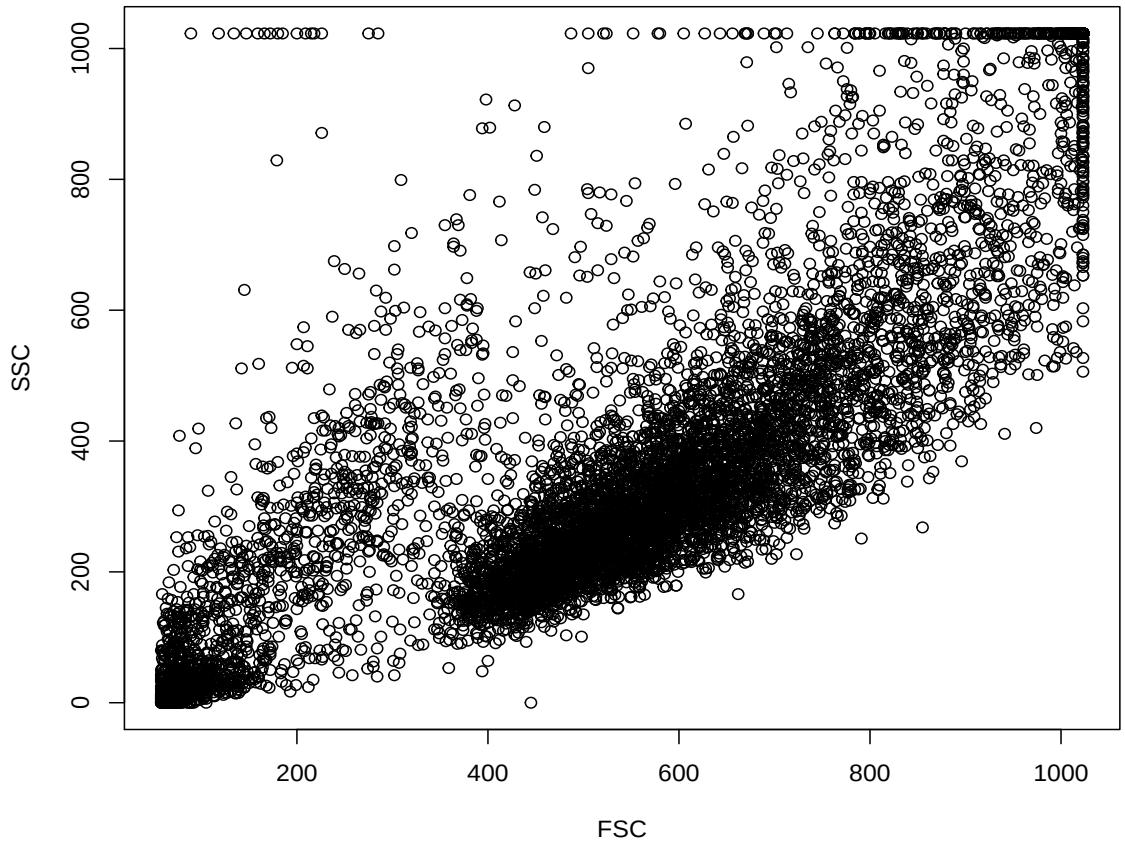


Figure 6.1: Scatter plot of the flow cytometry data.

Figure 6.1 above shows the scatter plot for the flow cytometry data, where the flow cytometry data demonstrates the skewed clusters very clearly. Even if one transforms the data we cannot get rid of the skewness in both directions simultaneously for clusters.

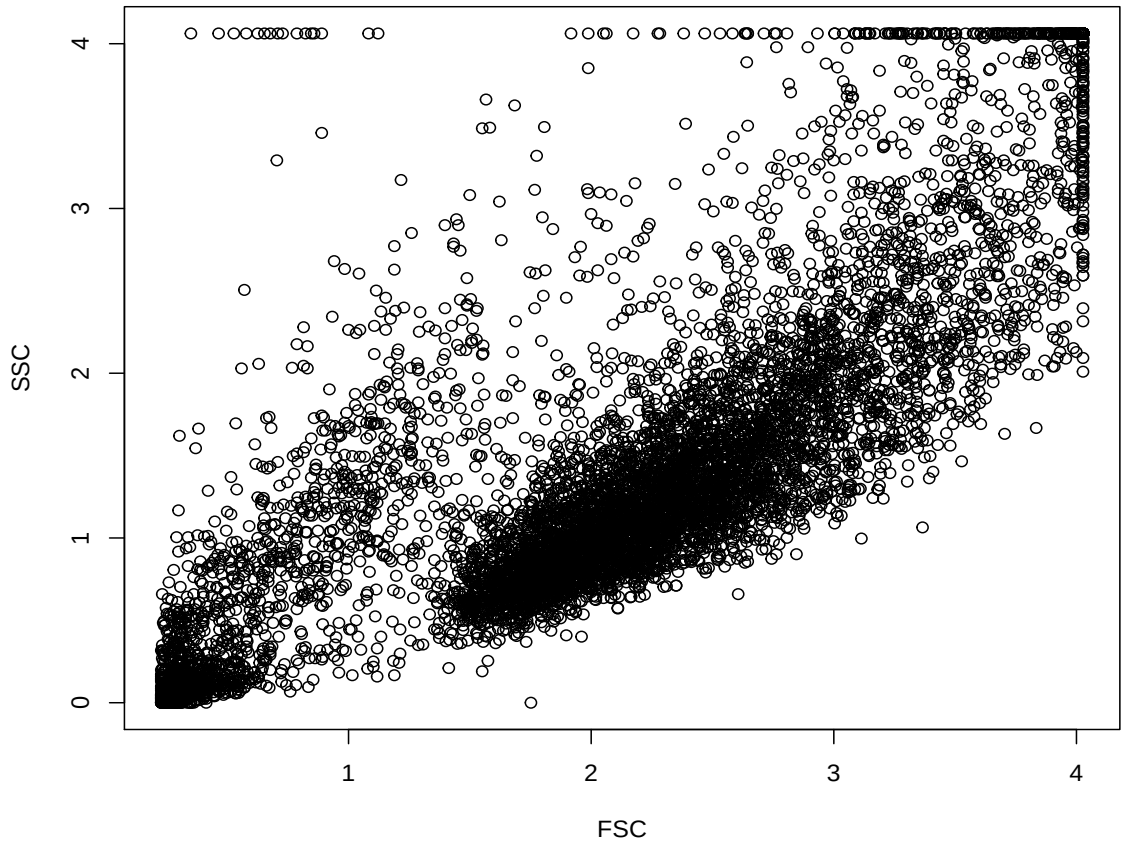


Figure 6.2: Plot of the scaled flow cytometry data.

We will use the `mixsmsn` R-package to fit a mixture of skew normals to the data. As the `mixsmsn` works with scaled data we first scale the original data, which is given in Figure 6.2. However, the scaling of the original flow cytometry data will not affect the decision to merge and group the mixture, and the effective number of groups will remain same. All following plots are given for the scaled measurements.

6.3 Determining the number of groups based on the elevation ridgeline plot

First we will fit a four components mixture to the dataset. We chose four as it was widely used in the literature in particular to capture some of the outlying points. The resulting cluster based on the fit by `mixsmsn` is given by the different coloured points in figure 6.3. Our next task will be to check whether we can merge any pairs of components.

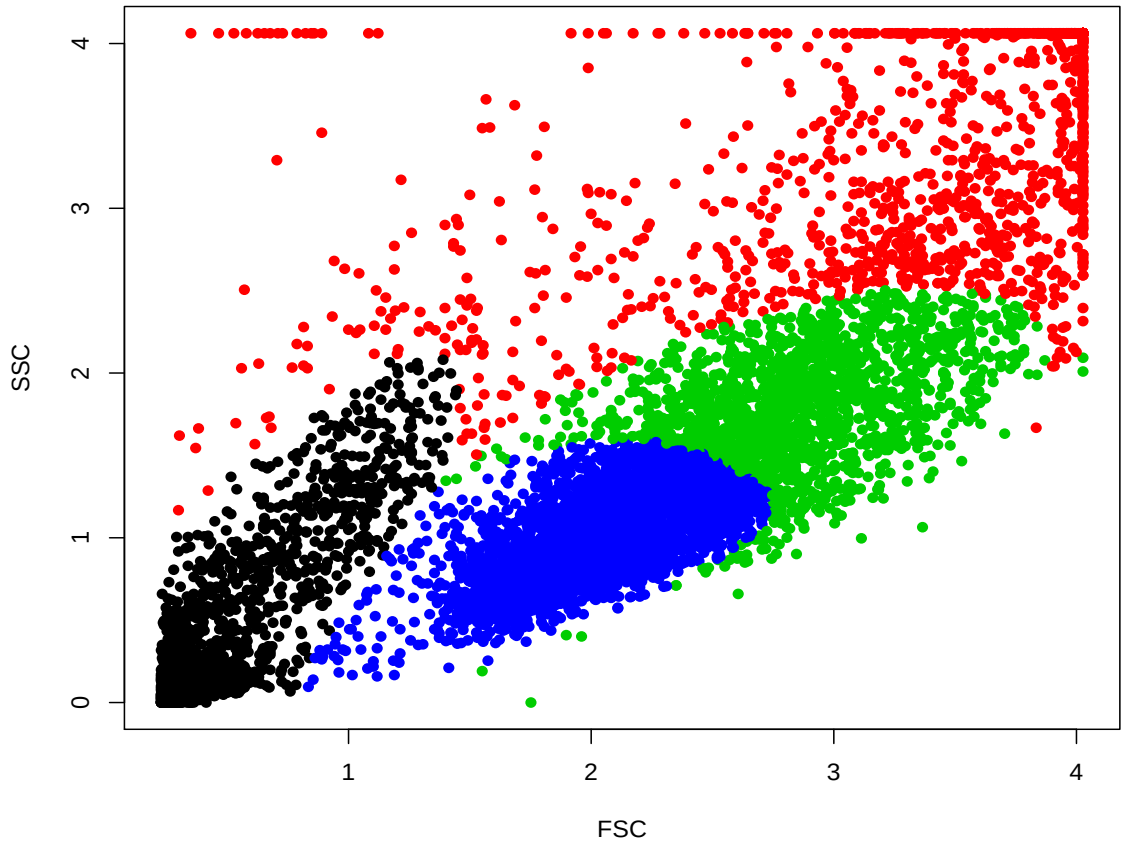


Figure 6.3: Flow cytometric plot with four mixture components of skew normal distribution.

6.4 Ridgeline analysis

We introduced the ridgeline elevation plot for the mixture of multivariate skew normal densities as the curve that passes from the mode of the first component to the mode of the second component. This curve provides important information relating to these components, such as the critical points (i.e. the maximum and lower points, and the saddle-point), the location, and the number of modes. The ridgeline function of the mixture of multivariate skew normal densities is given in the following equation:

$$x^*(\alpha) = \sum_{k=1}^{K} \alpha_i \nabla \log \Big(f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\big)\Big), \qquad \alpha_i \in S^{K-1},$$

where $f_{sn}\big(\boldsymbol{x}; \boldsymbol{\xi}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\big)$ is the probability density function of the multivariate skew normal distribution and α is the range of the ridgeline function of the skew normal mixture. We will use the curve of the ridgeline function to analyse the modality of the four mixture components used with the mixture of multivariate skew normal densities for modelling the flow cytometry data. We used our `ridge.skew.var` function (Alruwaili and Ray, 2019) to perform the analysis and obtain the next set of plots and tables that will help us decide whether to merge the components or not.

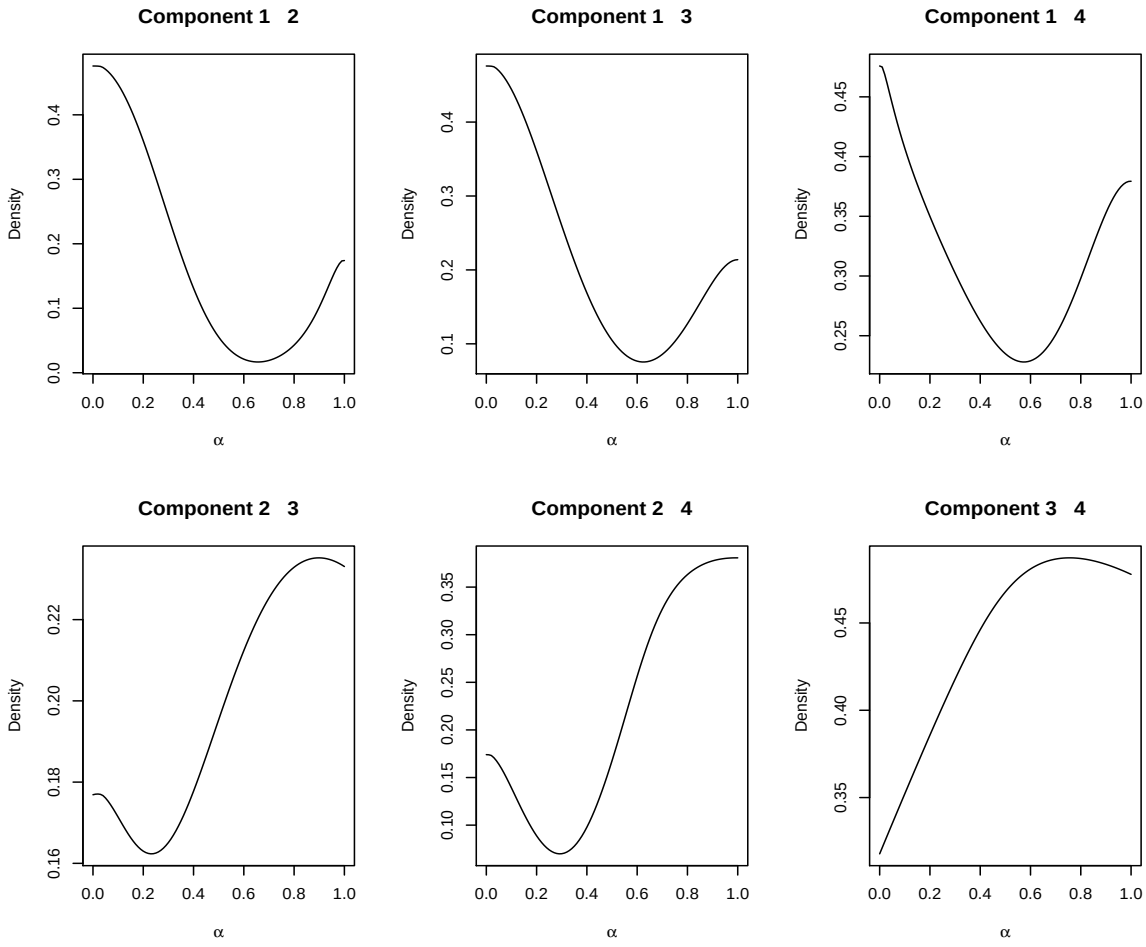


Figure 6.4: Ridgeline analysis for four mixture components of flow cytometry data.

Table 6.1: Ridgeline connection for four mixture components of flow cytometric data.

	Component 1	Component 2	Component 3	Component 4
Component 1	**1**	0	0	0
Component 2	0	**1**	0	0
Component 3	0	0	**1**	**1**
Component 4	0	0	**1**	**1**

Figure 6.4 shows the ridgeline analysis for the four mixture components. The plots show that all the ridgeline elevation plots for these pairs of mixture components display two modes, except in the case of components three and four, where only one mode is displayed. Table 6.1 also shows the ridgeline connection for the four mixture components of the flow cytometry data. Here each mixture of two components displaying one mode is denoted by 1, and each mixture of two components displaying more than one mode is denoted by 0.

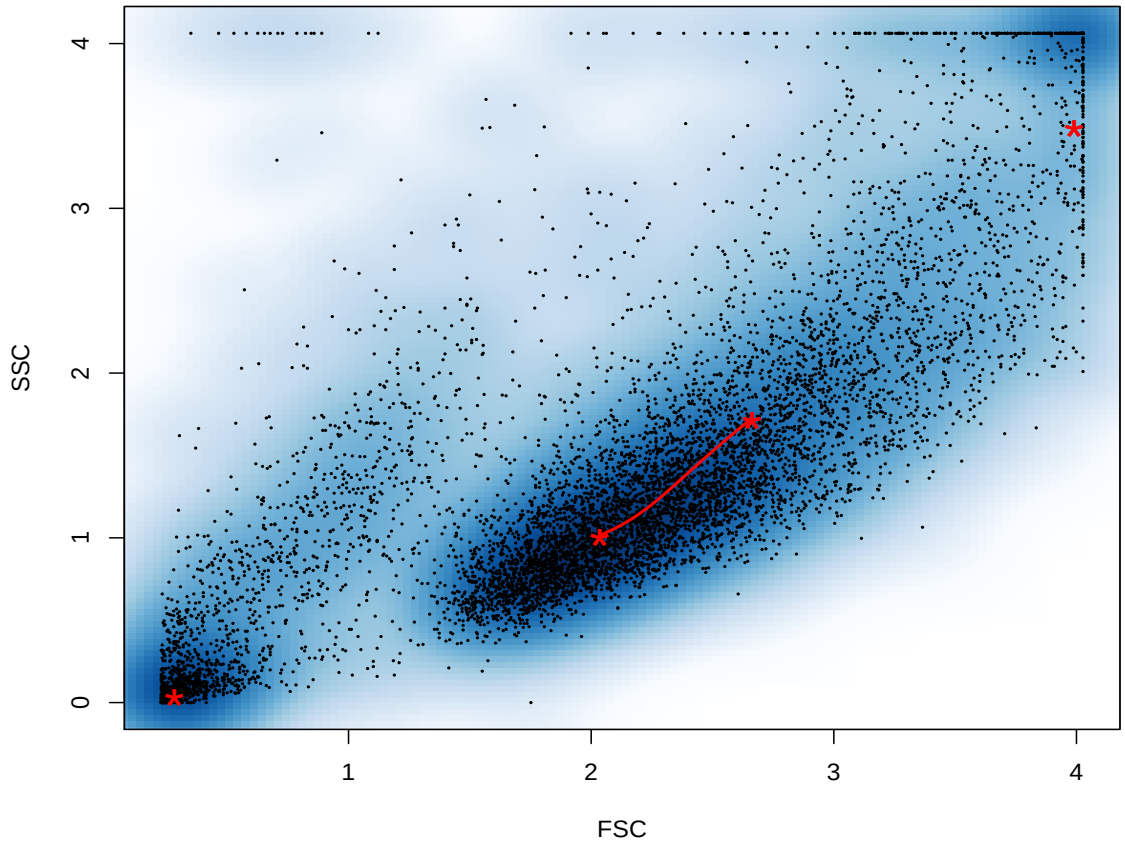


Figure 6.5: Modes of the four components along with the ridgeline curve connecting the merged components.

Figure 6.4 and Table 6.1 enable a clear understanding of which mixture components should be merged together and which mixture components should not be merged to-

gether, based on the number of modes for these pairs of mixture components. Figure 6.5 shows the modes of the four mixture components, with the ridgeline connection for components three and four, when these two components could be merging in one group. The red star marks the mode of each component.

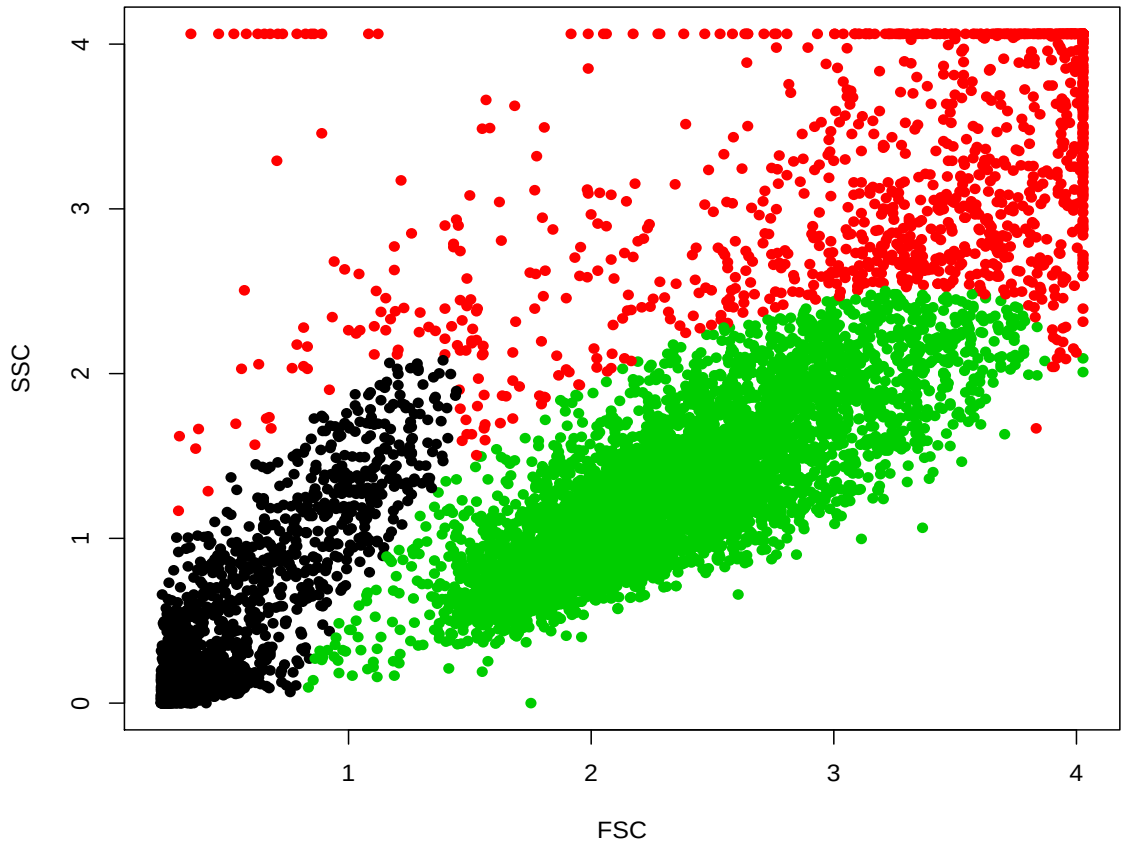


Figure 6.6: Plot of the flow cytometry data after merging the relevant components resulting in three clusters.

Figure 6.6 shows the scatter plot of the flow cytometry data after merging components three and four together. By using the result of the ridgeline analysis and the table of the ridgeline connection, we know which components should be merging together as one cluster, as shown in Figure 6.5.

6.5 Ratio of the ridgeline curve

Another approach to group and merge the components of the mixture of multivariate skew normal densities is the ratio of the ridgeline curve. From the plots of the ridgeline analysis of the mixture components, we observed that sometimes the mixture of the two components display two local maxima and one of them is very close to the minima. The two local maxima gives us the mathematical modes, but are they statistically significant? It is difficult to answer this question as we do not know the distribution of the modal heights. So we will implement a threshold based approach to account for the statistical variability.

Figure 6.7 provides an illustration of the situation where there are two mathematical modes, but the separation is very shallow.

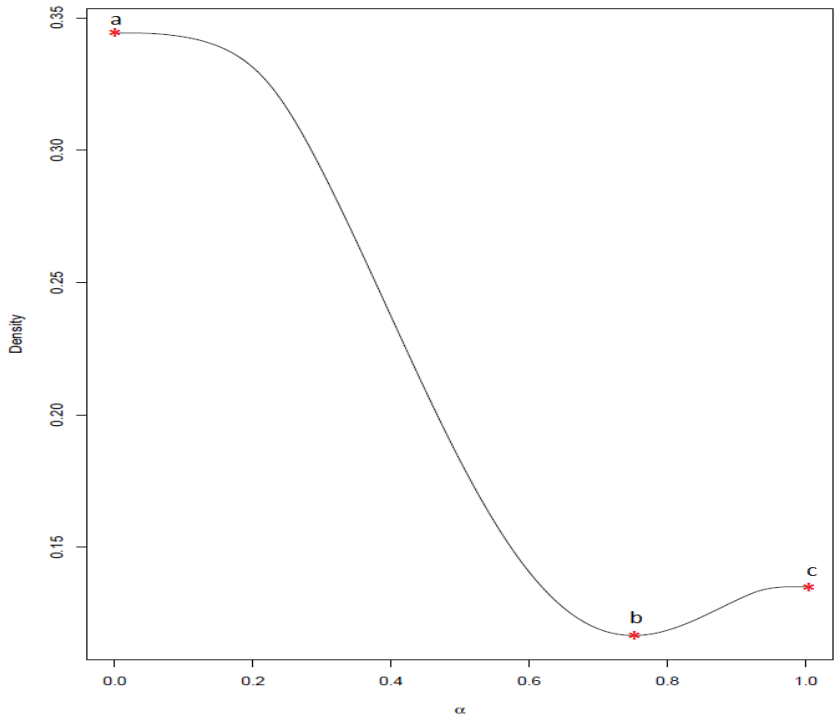


Figure 6.7: Critical points definining the maxima and minima along the ridgeline elevation plot.

Figure 6.7 shows the ridgeline curve with the critical points on the ridgeline. Here

(a) denotes the higher of the two local maxima, and (c) denotes the lower of the two local maxima. The minimum point between the two local maxima is denoted by (b). To design a new approach of merging the mixture components into a single cluster, we considered the ratio of the minimum point between the two local maxima and the lower of the two local maxima and then choose a cut off value to merge these two mixture components into one cluster. The ratio of the ridgeline can be calculated as follows:

$$\frac{\text{The minimum point between the two local maxima}}{\text{The lower of the two local maxima}} = \frac{b}{c}.$$

Based on a predefined cut off value between 0 and 1 we define the rules for merging mode as follows:

- if $\frac{b}{c} >$ (cut off value) then merge the components.
- if $\frac{b}{c} <$ (cut off value) then do not merge the components.
- if the elevation plot has one mode the ratio is one and hence they are merged.

Table 6.2 shows the ratio for the ridgeline of the some four component fit as in section 6.3.

Table 6.2: The ratio matrix of the ridgeline of four components.

	Component 1	Component 2	Component 3	Component 4
Component 1	**1.000**	0.096	0.353	0.601
Component 2	0.096	**1.000**	0.917	0.401
Component 3	0.353	0.917	**1.000**	**1.000**
Component 4	0.601	0.401	**1.000**	**1.000**

Table 6.3: Merging matrix for four components with a cut off value of 0.7

	Component 1	Component 2	Component 3	Component 4
Component 1	**1**	0	0	0
Component 2	0	**1**	**1**	0
Component 3	0	**1**	**1**	**1**
Component 4	0	0	**1**	**1**

Table 6.2 shows the ratio of the ridgeline of the mixture components of the flow cytometry data. Table 6.3 shows the ridgeline connection for the four mixture components of the flow cytometry data when using 0.7 as the cut off point. We used 0.7 as an example,and the user might choose to use a different value between 0 and 1 for specific applications. Table 6.3 shows that there are two extra mixture components that should merge in the case of using 0.7 with the results from the ratio of the ridgeline that were given in Table 6.2.

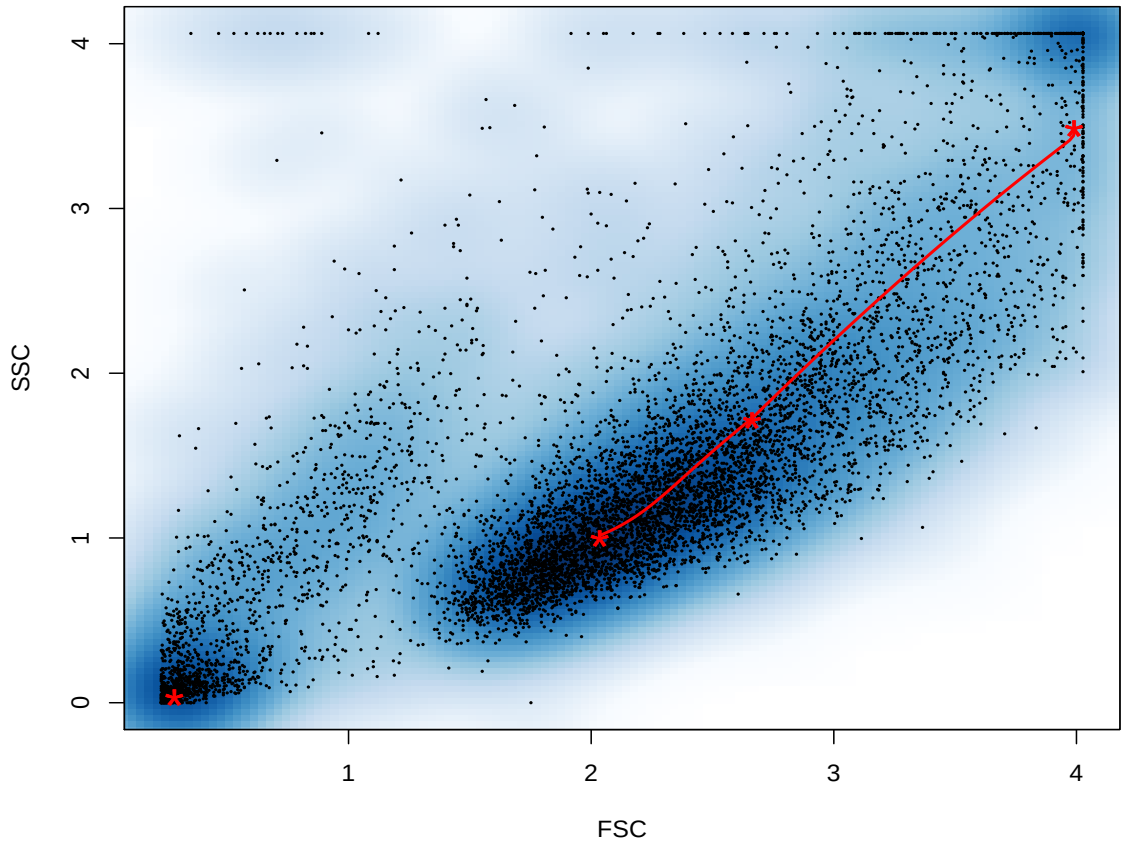


Figure 6.8: Modes of the four components along with the ridgeline curve connecting the merged components using the ratio of the ridgeline criteria.

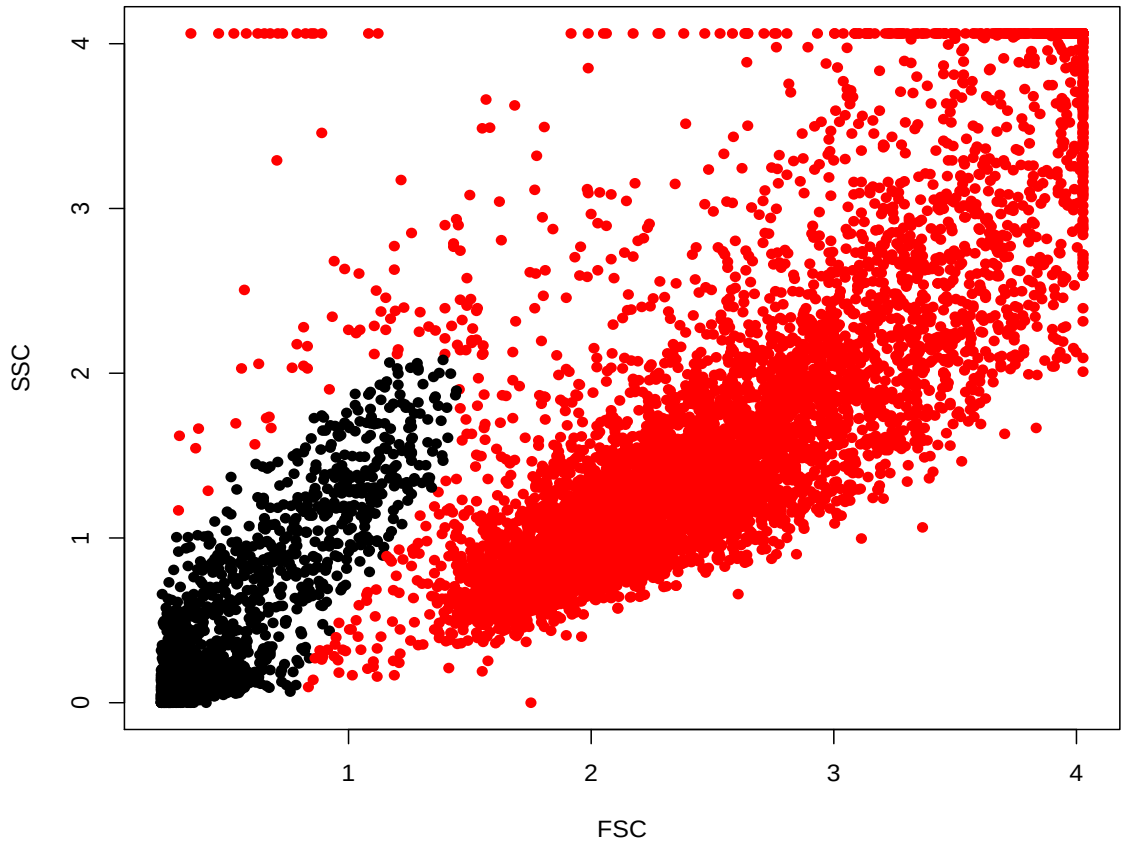


Figure 6.9: Scatter plot of the flow cytometry data after merging the components according to the ratio of the ridgeline.

Figure 6.8 shows the mixture components modes that should be merging in the case of using the ratio of the ridgeline and the cut off point 0.7. Figure 6.9 shows the scatter plot of the flow cytometry data after merging components three and four together and components two and three together in one cluster.

6.6 Simulation data

Now we provide an extensive simulation scheme to evaluate the performance of our mode merging tools. In each case we will simulate data from a mixture of four skew-

normal components with varying number of clusters organised by their modes. We will first use the `snsm.mmix` function to fit a four component skew-normal and then merge those components base on our merging tools. We evaluate the performance of our tool for varying number of samples for 100 simulations under each scenario and sample size.

6.6.1 Four components and three clusters

The first set of parameters incur three modes or clusters as displayed by the contour plot in figure 6.10.

Example 6.1. *(Four components, equal variance two modes) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 4$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \boldsymbol{\xi}_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \boldsymbol{\xi}_4 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$
$$\boldsymbol{\Omega}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Omega}_4 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\lambda}_1 = \begin{pmatrix} -5 \\ 5 \end{pmatrix},$$
$$\boldsymbol{\lambda}_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \boldsymbol{\lambda}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \boldsymbol{\lambda}_4 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25.$$

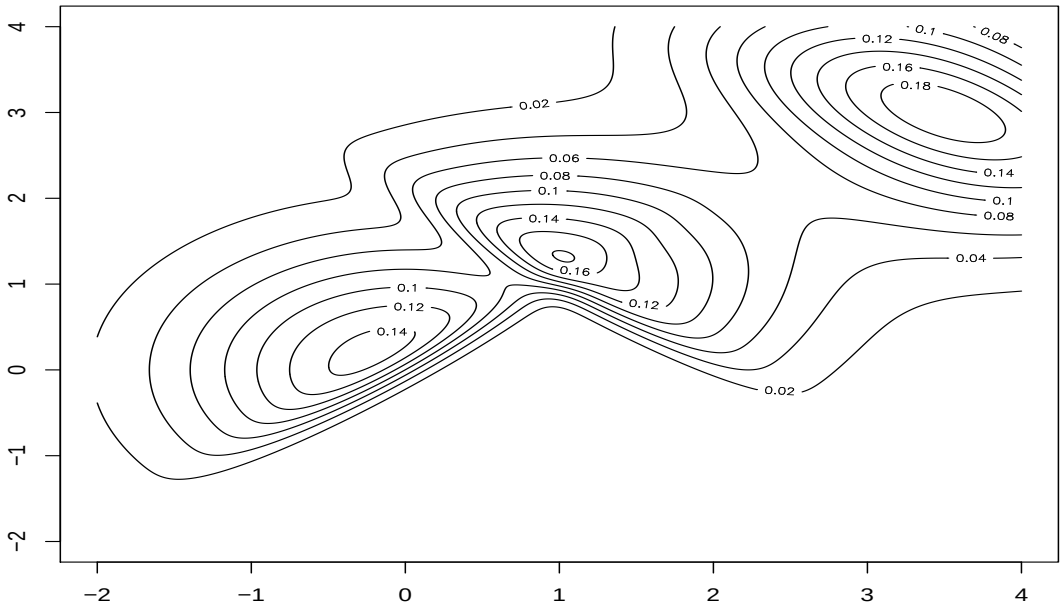


Figure 6.10: Contour plot of the four component mixture given in example 6.1 displaying three clusters.

Table 6.4 and gives the percentage of times the ridgeline analysis concludes one, two, three or four components. In contrast 6.5 gives the result using the ratio of ridgeline analysis with a threshold of 0.7. The different rows gives the results for changing number of samples used in each simulation.

Table 6.4: Clustering the data by using ridgeline analysis.

Number of sample	1 Cluster	2 Clusters	3 Clusters	4 Clusters
400	0 %	4 %	73 %	23 %
800	0 %	3 %	77 %	20 %
2000	0 %	0 %	97 %	3 %

In both cases we can observe that performance of the mode merging algorithm is getting better (i.e. selecting 3 clusters) with the increase in sample size. We can also observe that the performance of the ratio method is not as good as the non-ratio method. The ratio method in general is more conservative erring on the lower number

Table 6.5: Clustering the data by using ratio of ridgeline.

Number of sample	1 Cluster	2 Clusters	3 Clusters	4 Clusters
400	2 %	42 %	55 %	1 %
800	1 %	51 %	48 %	0 %
2000	0 %	65 %	35 %	0 %

of clusters.

6.6.2 Four components and two clusters

In this example the true number of clusters is two as displayed by the contour plot in figure 6.11.

Example 6.2. *(Four components, equal variance, two clusters) By considering the mixture of multivariate skew normal densities with $D = 2$, and $K = 4$ with the following parameters:*

$$\boldsymbol{\xi}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\xi}_2 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, \boldsymbol{\xi}_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \boldsymbol{\xi}_4 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \boldsymbol{\Omega}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$
$$\boldsymbol{\Omega}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \boldsymbol{\Omega}_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \boldsymbol{\Omega}_4 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \boldsymbol{\lambda}_1 = \begin{pmatrix} -5 \\ 5 \end{pmatrix},$$
$$\boldsymbol{\lambda}_2 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \boldsymbol{\lambda}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \boldsymbol{\lambda}_4 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25.$$

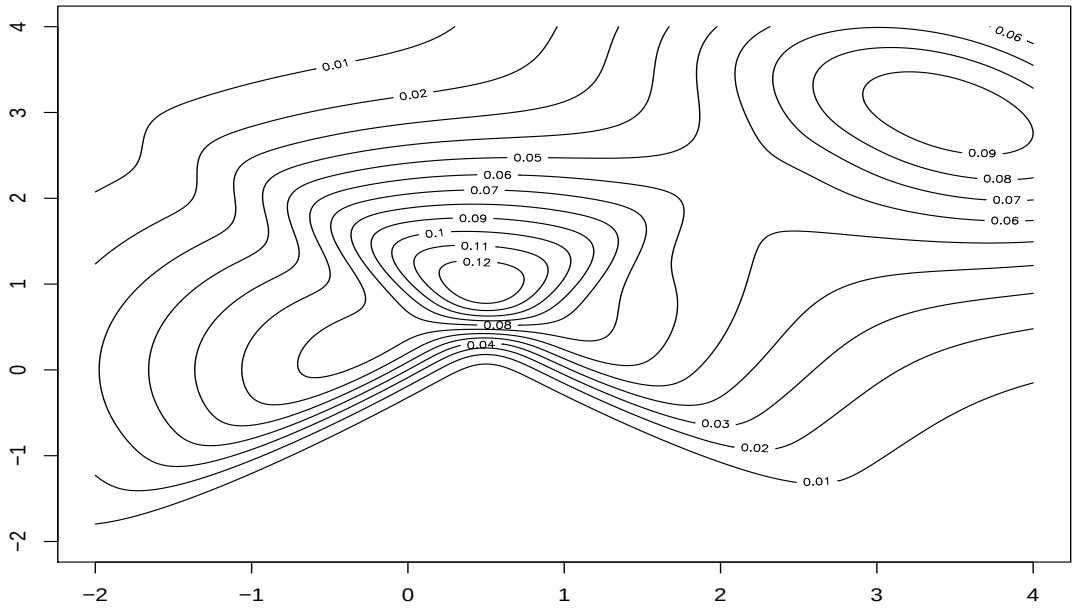


Figure 6.11: Contour plot of the four component mixture given in example 6.2 displaying two clusters.

Table 6.6 gives the percentage of times the ridgeline analysis concludes one, two, three or four components. In contrast 6.7 gives the result using the ratio of ridgeline analysis with a threshold of 0.7. The different rows gives the results for changing number of samples used in each simulation.

Table 6.6: Clustering the data by using ridgeline analysis.

Number of sample	1 Cluster	2 Clusters	3 Clusters	4 Clusters
400	1 %	32 %	47 %	20 %
800	1 %	52 %	37 %	10 %
2000	1 %	82 %	16 %	1 %

Again, in both cases we can observe that performance of the mode merging algorithm is getting better (i.e. selecting 2 clusters) with the increase in sample size. But here We can observe that the performance of the ratio method is better than the non-ratio method. Again we see that the ratio method in general is more conservative which

Table 6.7: Clustering the data by using ratio of ridgeline.

Number of sample	1 Cluster	2 Clusters	3 Clusters	4 Clusters
400	20 %	67 %	12 %	1 %
800	21 %	75 %	4 %	0 %
2000	12 %	88 %	0 %	0 %

might be the reason it is selecting two components more often than less.

6.7 Conclusion

The main contribution in this chapter was to describe the creation of computational (R-code) that can help merge homogeneous skew normal mixture components into one group. We introduced two tools to merge the mixture components together in one cluster. The first tool for merging the mixture components was the ridgeline function, where the ridgeline function is introduced as curve passes from the mode of the first component to the mode of the second component where this curve provides all the important information about these two mixture components, such as other available modes, and critical points. We used the ridgeline function as a tool to merge the homogeneous mixture components in one cluster based on the number of modes for these mixture components. For example, we merged two mixture components together in one cluster if these two mixture components were just displaying one mode.

The second tool for merging the mixture components of the multivariate skew normal densities was the ridgeline ratio. By exploring the number of modes for all the pairs of the components of the skew normal mixture by using the ridgeline curve, we observed that some of the mixture components displaying two local maxima and one of these local maxima was higher than the other, and that gave us the chance to merge these mixture components together in one cluster. In other words, we considered the ratio of the minimum point between the two local maxima and the lower of the two local

maxima to merge the mixture components together in one cluster.

In this chapter we used a real dataset (flow cytometry data) as an example to illustrate the usefulness of using the ridgeline function and the ratio of the ridgeline to merge the mixture components of the multivariate skew normal distribution. The functions for merging the homogeneous skew normal mixture components will be available in the future as an R-package to make it suitable for the other researchers to use.

Chapter 7

Conclusion and Future Work

This chapter summarises the outcome of my dissertation research and suggests a range of unresolved questions that can be solved by extending the current research work. This research focused on studying and exploring the mode (modality) of densities. To date, methodological development on studying the modality of multivariate mixtures has been primarily focused on normal mixtures. But in this era of cheap computation, models beyond normal mixtures are widely used to fit and cluster real datasets if they are deemed to provide a better explanation of the generating process. This dissertation provides the first research work that provides a thorough methodological study of the modal properties of mixture of multivariate t-densities, the univariate of skew normal distribution, multivariate skew normal distribution, and the mixture of multivariate skew normal densities.

In this dissertation we also introduced tools that can be readily used for merging mixture components, based on the number of modes of these mixture components. The new methodological results and the computational tools were detailed in four separate chapters in the dissertation, as summarised below:

Chapter Three: Summary of its content and suggestions for future work.

In Chapter 3 we explored the modality of the mixture of multivariate t-densities to find

the upper-bound of the number of modes for the mixture of multivariate t-densities. We explored the upper bound for the number of modes for the mixture of multivariate t-densities in two ways, namely, graphically and analytically. Graphically, we investigated the number of modes for a two component mixture of multivariate t-densities in two and three dimensions. We introduced some examples to show the possible effect of varying degree of freedom on the number of modes of the mixture components. By using different degrees of freedom for the fixed parameters of the two t-component mixture, we found that the degrees of freedom do indeed affect the number of modes. The maximum number of modes we achieved for the mixture of two t-components in two dimensions was three, and in three dimensions it was four modes. By exploring the effect of the degree of freedom on the number of modes of the mixture of two t-distribution and three dimensions, we noted that the total number of modes that could be derived from using the same degrees of freedom for both of the mixture components was two or four modes. We also observed that this was different in the case of using different degrees of freedom for each component. The total number of modes that could be achieved from using different degrees of freedom was two, three, and four. By exploring the number of modes for the mixture of two t-components graphically, we were unable to get more than $D + 1$ modes for two or three dimensions.

Analytically, by exploring the equation of the curvature function $k(\alpha)$ of the mixture of multivariate t-densities, we found that it was difficult to exploit the curvature function to comment on the number of modes. The current results from exploring the equation of the curvature function $k(\alpha)$ of the mixture of multivariate t-densities support our argument that the upper bound of the number of modes for the mixture of multivariate t-densities is the same as the upper bound of the number of modes for the mixture of multivariate normal densities. By solving $f(s) = 1$, the upper bound of the number of modes for the mixture of multivariate t-densities will then be exactly

the same as the upper bound of the number of modes for the mixture of multivariate normal distribution. This interesting conjecture will be considered in more detail in our future research work. Beyond the conjecture on the upper bound, the results from this dissertation can be readily used for a wide range of statistical purposes, such as:

- Understanding the number of modes for the mixture of multivariate t-densities helps to make a decision about when it is possible to merge the components of the mixture of the multivariate t-densities.

- The results from exploring the number of modes for the mixture of multivariate t-densities provide a clear understanding of the effect of the degrees of freedom on the number of modes for the mixture of multivariate t-densities, and this might be of help in exploring the number of the modes for the mixture of multivariate skewed t-densities.

Chapter Four: Summary of its content and suggestions for future work.
In Chapter 4 we studied and explored the modality of the univariate skew normal distribution and the modality of the multivariate skew normal distribution. For univariate skew normal distributions we provided an implicit equation to calculate the modes of the univariate skew normal distributions along with an R-package to numerically find the mode.

By studying the relationship between the mean and the mode of the skewness parameters, we found that the absolute value of the mode never exceeds the mean of the univariate of skew normal distribution, and this helped us to find the search range for the numerical optimization of the mode.

By exploring the mode of the skew normal distribution, we found that the mode of a standard skew normal is very stable and converges to zero for large λ. In fact it can be shown that the maximum deviation of the mode is 0.53 and this occurs for

$\lambda = 1.55$. We arrived at similar results for multivariate skew mixtures and provided the corresponding function in the R-package.

Open problems in this area include

- Finding the value of the maximum divergence for the mode for multivariate skew normal distributions.
- Studying the relationship between the mean and the mode of the multivariate skew normal distribution.
- Generalizing the concepts for finding the mode of the skew normal distribution to explore the modality of univariate and multivariate skew t-distributions.
- Studying the relationship between the mean and the mode of the univariate and multivariate skew t-distribution.

Chapter Five: Summary of its content and suggestions for future work.

In Chapter 5 we explored the modality of a multivariate skew normal mixture. We introduced the ridgeline function and the height of the ridgeline function as useful tools to explore the modes of the mixture of multivariate skew normal densities. We also introduced the ridgeline function of the mixture of multivariate skew normal densities as a curve that passes from the mode of the first component to the mode of the second component through all the available modes, where all the information of the components' mixture, such as the number and location of modes for the density, critical points, and saddle-points, lie on the curve of the ridgeline function. The height of the ridgeline function was introduced as the height of the ridgeline function along the $x^*(\alpha)$ curve. The height of the ridgeline function allows the number of the modes of the mixture of multivariate skew normal densities to be introduced, where the contour plots are not available to present the modes for more than two dimensions. The ridgeline function and the height of the ridgeline function enable all the number of modes to be presented

for the mixture of multivariate skew normal densities for different values of skewness parameters. The concept of the ridgeline function and the height of the ridgeline function of the mixture of multivariate skew normal densities can be uses as tools to decide when it is possible to merge two mixture skew normal components together in one cluster.

Future work in this area will involve the generalization of the concept of using the ridgeline function and the height of the ridgeline function to study the modality of the mixture of multivariate skew t-densities. From our exploration of the modality of the mixture of multivariate skew normal densities, we observed that the total number of modes that we got from mixing two skew normal components in two dimensions was three and the total number of modes for mixing two skew normal components in three dimensions was four. These results raised the following question in this research area: What is the upper bound for the number of modes that one can get from mixing two skew normal components in any number of dimensions?

Chapter Six: Summary of its content and suggestions for future work

In Chapter 6 we used the concepts of the ridgeline function and the height of the ridgeline function to create a framework (R-code) that can be used to decide when to merge the pairs of the mixture components of multivariate skew normal densities into one group. We introduced two tools to decide when to merge the pairs of the mixture components of multivariate skew normal densities together in one cluster. The first tool for merging the mixture components of multivariate skew normal densities was the ridgeline function; the ridgeline function was introduced as the curve that passes from the mode of the first component to the mode of the second component, where this curve provides all the important information about these two mixture components, such as other available modes and critical points. We used the ridgeline function as a tool to merge the homogeneous mixture components into one cluster, based on the number of modes for these mixture components. For example, the two mixture components

of multivariate skew normal distribution can be merged into one cluster if these two mixture components just display one mode. The second tool for deciding when to merging pairs of the mixture components of the multivariate skew normal densities into one cluster was the ridgeline ratio. We then considered the ratio of the minimum point between the two local maxima and the lower of the two local maxima to merge the mixture components together into one cluster.

In Chapter 6 we also used a real dataset (flow cytometric data) as an example to illustrate the usefulness of using the ridgeline function and the ratio of the ridgeline to merge the mixture components of the multivariate skew normal densities. This framework (R-code) for merging the mixture components of multivariate skew normal densities components will be available in the future as an R-package, which will make it suitable for other researchers to use.

Bibliography

Mohammad Ahsanullah, BM Golam Kibria, and Mohammad Shakil. *Normal and student's t distributions and their applications*, volume 4. Springer, 2014.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, Dec 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.1100705.

G Alexandrovich. *Analytische Eigenschaften von Mischungen elliptischer Verteilungen und deren Anwendung in der Clusteranalyse*. PhD thesis, Philipps Universität Marburg, 2011.

Grigory Alexandrovich, Hajo Holzmann, and Surajit Ray. On the number of modes of finite mixtures of elliptical distributions. In Berthold Lausen, Dirk Van den Poel, and Alfred Ultsch, editors, *Algorithms from and for Nature and Life*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 49–57. Springer International Publishing, 2013. ISBN 978-3-319-00034-3. doi: 10.1007/978-3-319-00035-0_4. URL http://dx.doi.org/10.1007/978-3-319-00035-0_4.

Bader Alruwaili and Surajit Ray. *R package: skewmode*, 2019. URL https://github.com/BaderAlruwaili/skewed-normal-codes.git.

R. B. Arellano-Valle and Adelchi Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006.

Reinaldo B Arellano-Valle, Marc G Genton, and Rosangela H Loschi. Shape mixtures of multivariate skew-normal distributions. *Journal of Multivariate Analysis*, 100(1): 91–101, 2009.

Reinaldo B Arellano-Valle, Clécio S Ferreira, and Marc G Genton. Scale and shape mixtures of multivariate skew-normal distributions. *Journal of Multivariate Analysis*, 166:98–110, 2018.

Olcay Arslan. Variance-mean mixture of the multivariate skew normal distribution. *Statistical Papers*, 56(2):353–378, 2015.

Adelchi Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, pages 171–178, 1985.

Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

Adelchi Azzalini. *The skew-normal and related families*, volume 3. Cambridge University Press, 2013.

Adelchi Azzalini and Antonella Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.

Adelchi Azzalini and Alessandra Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

Rodrigo M Basso, Víctor H Lachos, Celso Rômulo Barbosa Cabral, and Pulak Ghosh. Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis*, 54(12):2926–2941, 2010.

Jean-Patrick Baudry, Adrian E. Raftery, Gilles Celeux, Kenneth Lo, and Raphael Gottardo. Combining mixture components for clustering. *Journal of Computational*

and Graphical Statistics, 19(2):332–353, 2010. doi: 10.1198/jcgs.2010.08111. URL https://doi.org/10.1198/jcgs.2010.08111.

Mauro Bernardi. Risk measures for skew normal mixtures. *Statistics & Probability Letters*, 83(8):1819–1824, 2013.

Márcia D Branco and Dipak K Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79(1):99–113, 2001.

Celso Rômulo Barbosa Cabral, Víctor Hugo Lachos, and Marcos O Prates. Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis*, 56(1):126–142, 2012.

A Capitanio, A Azzalini, and Elena Stanghellini. Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*, 30(1):129–144, 2003.

Miguel Á Carreira-Perpiñán and Christopher KI Williams. On the number of modes of a Gaussian mixture. In *International Conference on Scale-Space Theories in Computer Vision*, pages 625–640. Springer, 2003.

Javier E Contreras-Reyes and Reinaldo B Arellano-Valle. Growth curve based on scale mixtures of skew-normal distributions to model the age-length relationship of cardinalfish (epigonus crassicaudus). *arXiv preprint arXiv:1212.5180*, 2012.

Greg Finak, Juan-Manuel Perez, Andrew Weng, and Raphael Gottardo. Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC bioinformatics*, 11(1):546, 2010.

Sylvia Frühwirth-Schnatter and Saumyadipta Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11(2):317–336, 2010.

Marc G Genton, Li He, and Xiangwei Liu. Moments of skew-normal random vectors and their quadratic forms. *Statistics & probability letters*, 51(4):319–325, 2001.

AK Gupta. Multivariate skew t-distribution. *Statistics: A Journal of Theoretical and Applied Statistics*, 37(4):359–363, 2003.

Arjun K Gupta and John T Chen. A class of multivariate skew-normal models. *Annals of the Institute of Statistical Mathematics*, 56(2):305–315, 2004.

Arjun K Gupta and Tuhao Chen. Goodness-of-fit tests for the skew-normal distribution. *Communications in Statistics-Simulation and Computation*, 30(4):907–930, 2001.

Florian Hahne, Nolwenn LeMeur, Ryan R Brinkman, Byron Ellis, Perry Haaland, Deepayan Sarkar, Josef Spidlen, Errol Strain, and Robert Gentleman. flowcore: a bioconductor package for high throughput flow cytometry. *BMC bioinformatics*, 10(1):106, 2009.

J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *Ann. Statist.*, 13 (1):70–84, 03 1985. doi: 10.1214/aos/1176346577. URL https://doi.org/10.1214/aos/1176346577.

Christian Hennig. Ridgeline plot and clusterwise stability as tools for merging Gaussian mixture components. In *Classification as a Tool for Research*, pages 109–116. Springer, 2010a.

Christian Hennig. Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification*, 4(1):3–34, Apr 2010b. ISSN 1862-5355. doi: 10.1007/s11634-010-0058-3. URL https://doi.org/10.1007/s11634-010-0058-3.

Norbert Henze. A probabilistic representation of the skew-normal'distribution. *Scandinavian Journal of Statistics*, pages 271–275, 1986.

MC Jones. A skew t distribution. *Probability and Statistical Models with Applications*, pages 269–278, 2001.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009.

Christine Keribin. Consistent estimate of the order of mixture models. *Sankhy=a, Series A*, 62:49–66, 01 2000. doi: 10.2307/25051289.

Hyoung-Moon Kim and Jun Zhao. Multivariate measures of skewness for the scale mixtures of skew-normal distributions. *Communications for Statistical Applications and Methods*, 25(2):109–130, 2018.

Victor H Lachos, Pulak Ghosh, and Reinaldo B Arellano-Valle. Likelihood based inference for skew-normal independent linear mixed models. *Statistica Sinica*, pages 303–322, 2010.

Sharon X Lee and Geoffrey J Mclachlan. On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, 7(3):241–266, 2013.

Sharon X Lee, Kaleb L Leemaqz, and Geoffrey J McLachlan. A block EM algorithm for multivariate skew normal and skew t-mixture models. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

Brian G. Leroux. Consistent estimation of a mixing distribution. *Ann. Statist.*, 20(3): 1350–1360, 09 1992. doi: 10.1214/aos/1176348772. URL https://doi.org/10.1214/aos/1176348772.

Jia Li. Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics*, 14(3):547–568, 09 2005. doi: 10.1198/106186005X59586. URL https://doi.org/10.1198/106186005X59586.

Tsung I Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257–265, 2009.

Tsung I Lin, Jack C Lee, and Shu Y Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, pages 909–927, 2007.

Tsung-I Lin, Hsiu J Ho, and Chia-Rong Lee. Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing*, 24(4):531–546, 2014.

Xiao Liu and A Unlü. Multivariate logistic mixtures. *Universal Journal of Applied Mathematics*, 3(4):77–87, 2015.

Kenneth Lo, Florian Hahne, Ryan R Brinkman, and Raphael Gottardo. flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC bioinformatics*, 10(1):145, 2009.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

Geoffrey J. McLachlan and David Peel. *Finite mixture models*. Wiley New York ; Chichester, 2000. ISBN 0471006262. URL http://www.loc.gov/catdir/toc/onix07/00043324.html.

David Peel and Geoffrey J McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

Arthur Pewsey. Problems of inference for azzalini's skewnormal distribution. *Journal of Applied Statistics*, 27(7):859–870, 2000.

Marcos Oliveira Prates, Victor Hugo Lachos, and C Cabral. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54(12):1–20, 2013.

Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I Lin, Lisa M Maier, Clare Baecher-Allan, Geoffrey J McLachlan, Pablo Tamayo, David A Hafler, et al. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524, 2009.

S. Ray and B.G. Lindsay. Model selection in high dimensions: a quadratic-risk-based approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):95–118, 2008. doi: 10.1111/j.1467-9868.2007.00623.x. URL http://eprints.gla.ac.uk/68987/.

Surajit Ray and Bruce G. Lindsay. The topography of multivariate normal mixtures. *The Annals of Statistics.*, 33(5):2042–2065, 10 2005. doi: 10.1214/009053605000000417. URL http://dx.doi.org/10.1214/009053605000000417.

Surajit Ray and Saumyadipta Pyne. A computational framework to emulate the human perspective in flow cytometric data analysis. *PloS one*, 7(5):e35693, 2012.

Surajit Ray and Dan Ren. On the upper bound of the number of modes of a multivariate normal mixture. *Journal of Multivariate Analysis*, 108:41 – 52, 2012. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2012.02.006. URL http://www.sciencedirect.com/science/article/pii/S0047259X12000401.

S Riggi and S Ingrassia. Modeling high energy cosmic rays mass composition data via mixtures of multivariate skew-t distributions. *arXiv preprint arXiv:1301.1178*, 2013.

Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, 6, 12 2012.

Richard H Scheuermann, Yu Qian, Chungwen Wei, and Inaki Sanz. Immport flock:

Automated cell population identification in high dimensional flow cytometry data (42.17), 2009.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics.*, 6(2): 461–464, 03 1978. doi: 10.1214/aos/1176344136. URL https://doi.org/10.1214/aos/1176344136.

S Soltyk and Ritu Gupta. Application of the multivariate skew normal mixture model with the EM algorithm to value-at-risk. *MODSIM2011*, pages 1638–1644, 2011.

D Michael Titterington, Adrian FM Smith, and Udi E Makov. *Statistical analysis of finite mixture distributions*. Wiley,, 1985.

Raluca Vernic. On the multivariate skew-normal distribution and its scale mixtures. 13 (2):83–96, 2005.

Michael Wallner. A half-normal distribution scheme for generating functions. *arXiv preprint arXiv:1610.00541*, 2016.

Daniel S Wilks. Cluster analysis. In *International geophysics*, volume 100, pages 603–616. Elsevier, 2011.