ABSTRACT

HU, HAO. A Study of Log-concave Mixture Models. (Under the direction of Yichao Wu.)

Mixture models are widely used when data are from a number of different components. Traditional parametric mixture models can be estimated via the expectation-maximization algorithm (known as the EM-algorithm) based on their parametric assumptions. However, these assumptions are sometimes too restrictive and the estimation results are biased if the models are misspecified. To relax the parametric assumption, we apply a log-concave shape constraint.

This dissertation analyzes the log-concave mixture models, which are more flexible and general than the traditional parametric mixture models. We developed a nonparametric log-concave maximum likelihood estimator (LCMLE) for the log-concave mixture model. In particular, we investigate the theoretical properties, computational algorithms and applications in clustering. We also develop the computational algorithms for the log-concave mixtures of regression model and its extension.

A Study of Log-concave Mixture Models

by
Hao Hu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2016

APPROVED BY:

Weixin Yao	Howard Bondell
Wenbin Lu	Jessie Jeng

Yichao Wu
Chair of Advisory Committee

DEDICATION

To my loving family and friends.

BIOGRAPHY

The author was born in Maanshan, Anhui, China in 1988. In 2006, he was admitted by College of Economics at Zhejiang University of China (ZJU) and graduated with a Bachelor degree is Finance and Economics in 2010. He earned a Master of Science degree in Economics from North Carolina State University in 2012. He is currently a PhD candidate in statistics at North Carolina State University and expects to complete his dissertation in 2016. With the valuable instruction and guidance from his advisors Dr Yichao Wu and Dr Weixin Yao, he focuses on the research of mixture models and log-concave shape constraint.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincerest gratitude and appreciation to my advisors Dr Yichao Wu and Dr Weixin Yao, who proposed the topics of this dissertation, provided me insightful instruction and guided me with extraordinary patience throughout my research. Without their guidance, I would not be able to accomplish this work.

I would also like to thank Dr Jessie Jeng, Dr Howard Bondell, and Dr Wenbin Lu for taking precious time to be on my committee and providing valuable suggestions. I also thank Dr Sylvia Blankenship in the Department of Horticultural Science for attending my oral exams as the graduate representative.

I also want to thank Dr Hua Zhou for serving as my academic advisor for the first two years of my graduate study, and Dr Howard Bondell for serving as the director of graduate programs in the department. Their suggestions helped me adapt to the life as a graduate student more smoothly.

I also want to thank my internship supervisor. Dr. Ram Valluru is the best manager I've ever had, who provided me with valuable industrial works experience.

I thank all the faculty members and staffs in the department. Throughout my four years of study, I attended their lectures, listened to their seminars, served as their teaching assistants, and sought help from them. All of these experiences are valuable to me and will serve me well in the future. I also thank my fellow friends in the department. It is your intelligence that stimulates and inspires me to conquer more difficulties.

Last but not least, I would like to express my deepest gratitude to my family. To my beloved wife, Fan Deng, and my great parents Jibing Hu and Hua Zou, your unconditional support always encourages me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

A finite mixture model assumes that observations come from several different subpopulations without indicating their true labels. It provides a flexible methodology to cluster data and analyze the feature within each group. Finite mixture models are extended to finite mixture of regression (FMR) models when the components' densities depend on specific covariates. FMR models make it possible to analyze linear relations between the response and covariates when the observations belong to different groups as well.

Finite mixture models and finite mixture of regression models are widely used in econometrics, biology, genetics, and engineering; see, e.g. Frühwirth-Schnatter (2001), Grün and Hornik (2012), Liang (2008), and Kostantinos (2000). For this reason, there is a rich history of studying these models both theoretically and practically, e.g. McLachlan and Peel (2000), Mcnicholas and Murphy (2008), and Lindsay (1995).

1.1 Finite Mixture Models

The finite mixture model has the density of the form:

$$f(\mathbf{x}_i|\boldsymbol{\psi}) = \sum_{j=1}^{K} \lambda_j g_j(\mathbf{x}_i;\theta_j), \quad \mathbf{x}_i \in \mathbb{R}^p, \tag{1.1}$$

where $\theta_j \in \mathbb{R}^{q_j}$, $\lambda_1, \ldots, \lambda_K$ are the mixing proportions, $\lambda_j \in (0,1)$ for $j = 1, \ldots, K$, $\sum_{j=1}^{K} \lambda_j = 1$, $\boldsymbol{\psi} = (\lambda_1, \ldots, \lambda_{K-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) \in \mathbb{R}^{\sum_{j=1}^{K} q_j + k - 1}$ and $g_j(x;\theta_j)$'s are component densities.

A traditional parametric mixture model assumes that each g_j belongs to a certain parametric family. A typical example would be a mixture of normal densities shown in Figure 1.1.

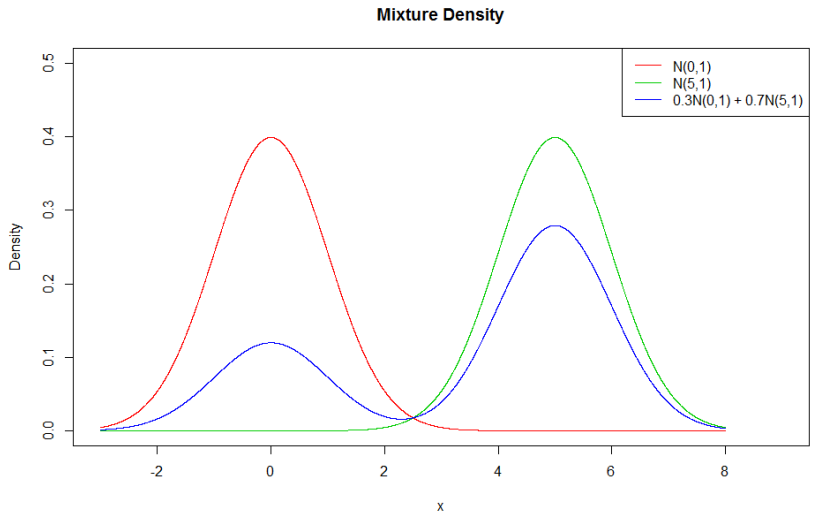


Figure 1.1: A mixture model: $0.3N(0,1) + 0.7N(5,1)$. The blue line represents the mixture density of $N(0,1)$ (red line) and $N(5,1)$ (green line).

Example 1.1 (Old Faithful Data). The dataset consists of the waiting length and eruption length (minutes) of 272 Old Faithful Geyser's eruptions in Yellowstone National Park. The data, which is shown in Figure 1.2, lies within a bimodal distribution. A multivariate normal distribution does not describe the data well. Alternatively, a mixture of multivariate normal distributions might be appropriate.

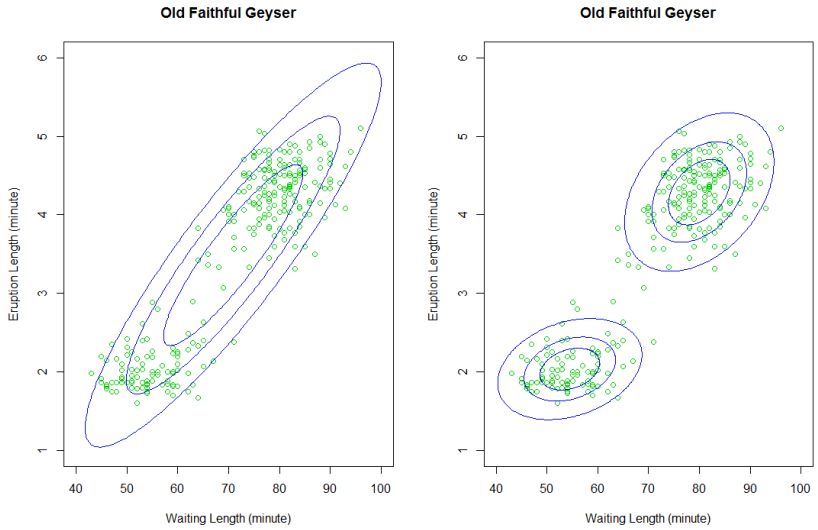


Figure 1.2: Old faithful data: Waiting length vs Eruption length (in minute). Modeling the data with a 2-dimensional multi-normal distribution (left figure) is not appropriate. Alternatively, a 2-component mixture of 2-dimensional normal distribution can characterize this joint distribution well.

1.2 Finite Mixture of regressions Model

A finite mixture of regression model (FMR) is best suited for random variables with a covariate-dependent finite mixture density. Suppose we observe a univariate response y_i and a p-dimensional covariate $\mathbf{x}_i$. For the j-th component, with probability λ_j, the

response and covariates follow a linear relationship of:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_j, \tag{1.2}$$

where $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,p-1})^T$ and ϵ_j's are independent random variables following distribution g_j. Then, the likelihood of the FMR model can be written as follows:

$$f(y_i, \mathbf{x}_i; \boldsymbol{\psi}) = \sum_{j=1}^{K} \lambda_j g_j(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j), \tag{1.3}$$

where $\boldsymbol{\beta}_j \subseteq \mathbb{R}^p$, $\lambda_j \in (0,1)$, $\sum_{j=1}^{K} \lambda_j = 1$, $\boldsymbol{\psi} = (\lambda_1, \ldots, \lambda_{K-1}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{Kp+K-1}$, and g_j is a parametric distribution function, such as normal, for j-th component, $j = 1, \ldots, K$.

Example 1.2 (Tone Data). The tone dataset (from R package `mixtools`) contains 150 trials from the same musician; see Cohen (1980) for a detailed description. In each trial, a fundamental tone, which was purely determined by a stretching ratio, was first provided to the musician. The musician then tuned the tone one octave above. The tuning ratio, which was measured as the adjusted tone divided by the fundamental tone, was recorded. The purpose of this experiment was to demonstrate the "Two musical perception Theory".

Example 1.3 ($CO_2 - GDP$ Data). The $CO_2 - GDP$ Data includes 172 countries' GDP (in 10,000 dollars) per capita and CO_2 emission (ton) per capita for the year of 2005. (We delete three outliers: Moldova Republic, Luxembourg and Qatar, from the original dataset.) The data is shown in Figure 1.4. Clearly the relationship between GPD and CO_2 has several different patterns, depending on the economic structure of each country.

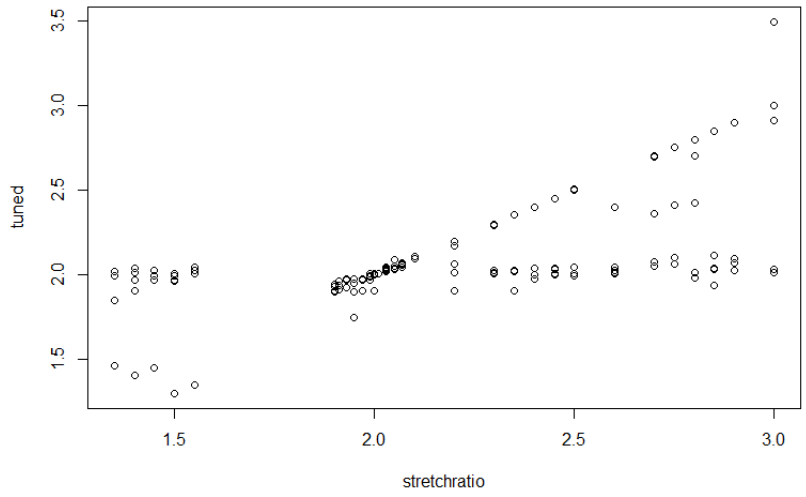


Figure 1.3: Tone data: Stretch ratio vs Tuned ratio. Clearly there are two patterns between those two ratios.

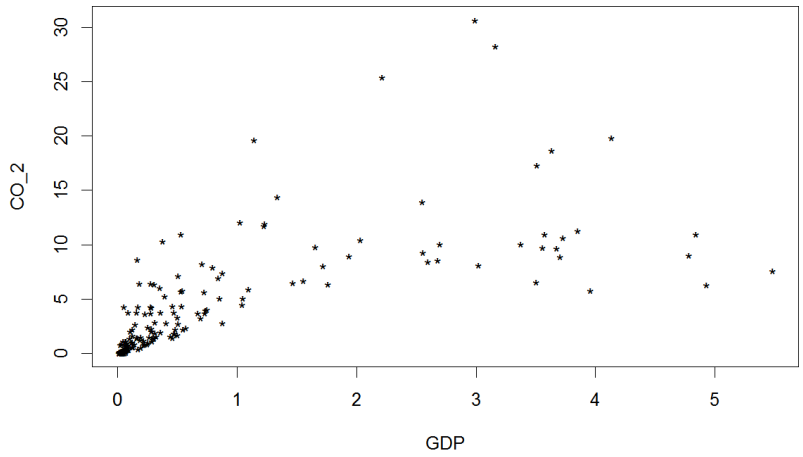


Figure 1.4: 172 Countries' GDP (in 10,000 dollars per capita) vs CO_2 emission (in ton per capita) for the year of 2005.

1.3 Organization of this dissertation

In Chapter 2, we introduce a general mixture model as well as its associated properties and computational algorithms. We continue by introducing a special case called the mixtures of regression model. We also review the Log-concave Maximum Likelihood Estimator (LCMLE) for estimating a single density along with its theoretical property. In Chapter 3, we propose the log-likelihood-type objective function and its maximizer as the LCMLE for the mixtures of log-concave densities. We explore the maximizer's theoretical properties and its associated EM-type algorithm. We show that the LCMLE exists and is consistent under fairly general conditions. Furthermore, we compare the numerical results with the traditional parametric mixture models. We show that our log-concave mixture models are more effective classifiers than parametric mixture models. Moreover, we observe no significant penalties for our log-concave mixture models when the parametric assumptions are satisfied. In Chapter 4, we apply the log-concave shape constraint to the mixtures of regressions models. We develop new EM-type algorithms for log-concave mixture of regression models and detail the process of how the algorithms solve the local maximum while being sensitive to outliers. In Chapter 5, we extend the mixture of regression model to a more general framework. We assume that the mixing proportions are no longer constants and are related via covariates. We developed associated EM-algorithms and compare the results to the EM-algorithms proposed in Chapter 4. We end the dissertation with a short conclusion and discussion in Chapter 6.

Chapter 2

Literature Review

2.1 Mixture Model Estimation

It is natural to estimate the unknown parameters in the mixture models like (1.1) from a maximum likelihood point of view. Lindsay et al. (1983a) and Lindsay et al. (1983b) determined the fundamental properties of MLEs for mixture models. Lindsay (1995) and McLachlan and Peel (2000) further extended the properties.

2.1.1 Maximum Likelihood Estimator (MLE)

Suppose $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ are independent observations from the mixture model (1.1). The log-likelihood for the unknown parameter $\boldsymbol{\psi}$ can be written as:

$$\ell(\boldsymbol{\psi}; \mathbf{X}) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \lambda_j g_j(\mathbf{x}_i|\theta_j). \tag{2.1}$$

The MLE of $\boldsymbol{\psi}$ is the root of the score function:

$$S(\mathbf{x}; \boldsymbol{\psi}) = \frac{\partial \ell(\boldsymbol{\psi}; \mathbf{x})}{\partial \boldsymbol{\psi}} = 0. \tag{2.2}$$

Unfortunately, there's no explicit solution to (2.2), even for the simplest mixture models such as normal mixtures. Consequently, the solution has to be obtained by algorithms. There are two general approaches to solve (2.2): the Newton-Raphson algorithm (McHugh (1956)) and the Expectation-Maximization (EM) algorithm (Dempster et al. (1977)). It is well known that for mixture models, the Newton-Raphson method converges faster than the EM algorithm but cannot guarantee convergence. Thus, the EM algorithm, known as the EM algorithm, is a more popular approach for analyzing mixture models. Readers may refer to McLachlan and Krishnan (2007) for a comprehensive introduction of the EM algorithm and its properties.

2.1.2 EM-algorithm

We assume the observed data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ to be incomplete and define the missing value $\mathbf{Z} = (\mathbf{z}_1^T, ..., \mathbf{z}_n^T)$, where z_i is a K-dimension vector with its j-th element given by:

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to } j\text{-th group,} \\ 0 & \text{otherwise.} \end{cases}$$

The complete log-likelihood is:

$$\log f(\boldsymbol{\theta}, \boldsymbol{\lambda}; \mathbf{X}, \mathbf{Z}) = \log \prod_{i=1}^{n} \prod_{j=1}^{K} [\lambda_j g_j(\mathbf{x}_i; \theta_j)]^{z_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} [\log \lambda_j + \log g_j(\mathbf{x}_i; \theta_j)].$$

Algorithm 2.1. *First we give random initial values to the unknown parameters as* $\boldsymbol{\psi}^{(0)}$ *and* $z_{ij}^{(0)}$*. Then in* t*-th iteration, it consists the following three steps.*

E-step: *Given* $\boldsymbol{\psi}^{(t)}$ *, we calculate*

$$z_{ij}^{(t+1)} = E(Z_{ij}|\boldsymbol{x}, \boldsymbol{\psi}^{(t)}) = \frac{\lambda_j^{(t)} g_j(\boldsymbol{x}_i|\theta_j^{(t)})}{\sum_{h=1}^{K} \lambda_h^{(t)} g_h(\boldsymbol{x}_i|\theta_h^{(t)})}, \quad i = 1, \ldots, n, j = 1, \ldots, K. \tag{2.3}$$

M-step 1: *Update* λ *simply through*

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(t+1)}, \quad j = 1, \ldots, K. \tag{2.4}$$

M-step 2: *Update* θ_j*'s by solving the root of:*

$$\sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij}^{(t+1)} \frac{\partial \log g_j(\boldsymbol{x}_i; \theta_j)}{\partial \theta_j} = 0, \quad j = 1, \ldots, K. \tag{2.5}$$

The algorithm is terminated if either t_{max} *of iterations has been reached, or if* $\ell^{(t+1)} - \ell^{(t)} < \epsilon = 10^{-8}$*, where* $\ell^{(t)} = \sum_{i=1}^{n} \log \sum_{j=1}^{K} \lambda_j^{(t)} g_j(\boldsymbol{x}_i; \theta_j^{(t)})$*.*

There are many different versions of the EM algorithms for mixture models, e.g. McLachlan and Krishnan (2007), Meng and Rubin (1993), and Liu and Rubin (1994). Each extension may focus on solving one particular model, guaranteeing the global convergence, or speeding up the algorithm.

2.1.3 Bayesian estimation

Mixture models can also be estimated via Bayesian approaches through the Markov Chain Monte Carlo estimation method. In this dissertation, we briefly review these approaches.

We denote the likelihood for observed data as $L_0(\boldsymbol{\psi})$ and the complete data likelihood as $L_c(\boldsymbol{\psi})$. Given a proper prior $\pi(\boldsymbol{\psi})$ for $\boldsymbol{\psi}$, and the conditional density for Z given $\boldsymbol{\psi}$ as $\pi(\mathbf{z};\boldsymbol{\psi})$, the posterior is

$$p(\boldsymbol{\psi}|y) = \frac{\pi(\psi)L_0(\boldsymbol{\psi})}{\int \pi(\psi)L_0(\boldsymbol{\psi})d\boldsymbol{\psi}} \tag{2.6}$$

$$= \frac{\sum_z \pi(\mathbf{z};\boldsymbol{\psi})\pi(\boldsymbol{\psi})L_c(\boldsymbol{\psi})}{\int \pi(\mathbf{z};\boldsymbol{\psi})\pi(\boldsymbol{\psi})L_c(\boldsymbol{\psi})d\boldsymbol{\psi}}. \tag{2.7}$$

The estimation of $\boldsymbol{\psi}$ depends on the posterior simulation by MCMC methods. Smith and Roberts (1993) designed the Gibbs sampling methods (Geman and Geman (1984)) to estimate the posterior. Alternatively, we may also apply Metropolis-Hasting sampling (Metropolis et al. (1953)), if there's no close form of distribution for the posterior. For more explicit details, readers may refer to Chapter 4 of McLachlan and Peel (2000).

2.2 Issues in Mixture Model Estimations

2.2.1 Identifiability

One of the biggest problems for the mixture model is the identifiability issue. Let

$$f(\mathbf{x}_i|\boldsymbol{\psi}) = \sum_{j=1}^{K} \lambda_j g_j(\mathbf{x}_i;\theta_j),$$

and

$$f(\mathbf{x}_i|\boldsymbol{\psi}^*) = \sum_{j=1}^{K^*} \lambda_j^* g_j(\mathbf{x}_i;\theta_j^*).$$

A mixture density like (1.1) is *identifiable* when

$$f(\mathbf{x}_i|\boldsymbol{\psi}) = f(\mathbf{x}_i|\boldsymbol{\psi}^*), \tag{2.8}$$

if and only if $K = K^*$, and under permutation of component labels,

$$\lambda_j = \lambda_j^*, \quad \text{and} \quad \theta_j = \theta_j^* \quad (j = 1, \ldots, K). \tag{2.9}$$

Mixture models are not identifiable in general. The lack of identifiability can be avoided by adding constraints such as $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_K$.

2.2.2 Selecting number of Components

For both maximum likelihood estimation and Bayesian approaches, the estimation procedures are under the assumption that the number of components K is known. In practice, K is sometimes unknown. In general, there are several approaches of selecting the number K. One of the methodologies is to adaptively select K in the algorithm. Usually the algorithm starts from a large number of K and will adaptively reduce K until some optimal criteria is reached, see Figueiredo and Jain (2002) as an example. Another way is to pre-determine K before running the algorithm by minimizing the information criteria. Examples include the Akaike information criterion (AIC) by Akaike (1998), the Bayesian information by Schwarz et al. (1978), and the consistent AIC (CAIC)

by Bozdogan (1987):

$$AIC = -2\ell(\hat{\boldsymbol{\psi}}) + 2K, \tag{2.10}$$

$$BIC = -2\ell(\hat{\boldsymbol{\psi}}) + K\log(n), \tag{2.11}$$

$$CAIC = -2\ell(\hat{\boldsymbol{\psi}}) + K(\log(n) + 1), \tag{2.12}$$

where $\hat{\boldsymbol{\psi}}$ is obtained via modeling with component number K. For each criteria, we select the number K that yields the smallest information criteria. It is sometimes known as the "eye-bowl" rule, as we pick the K when the information criteria becomes flat or starts to increase. For example, we fit the K components mixture model for the data of [Example 1.1]. Clearly $K = 2$ is the point which minimizes the information criteria.

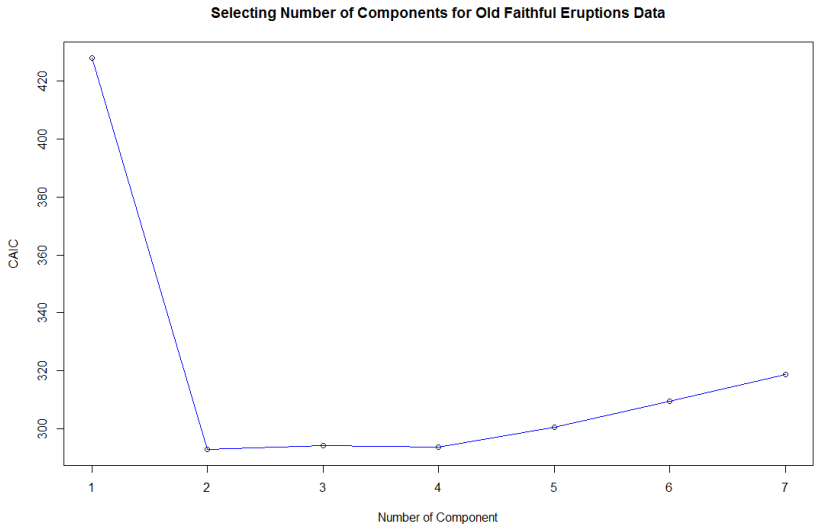


Figure 2.1: Picking the number of components by minimizing CAIC: Old faithful data from Example 1.1. $K = 2$ is the point which is lowest in the $CAIC$ and should be picked by the "eye-bowl" rule.

2.2.3 Label Switching

There is a well-known *label switching* issue for mixture models. There are plenty of articles devoted to this issue, especially for Bayesian mixture models, see e.g. Yao and Lindsay (2009) and Stephens (2000). In this dissertation, we focus on the label switching for frequentist mixture models and its solution in the numeric study. We adopt two labeling methods consistent with Yao (2015). Let $\hat{\boldsymbol{\theta}}^{\omega}$ be the estimator under permutation ω and $\mathbf{z} = \{z_{ij}\}_{i=1,\ldots,n,j=1,\ldots,K}$ be the true label where

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to } j\text{-th group,} \\ 0 & \text{otherwise.} \end{cases}$$

We denote $\hat{z}_{ij}(\boldsymbol{\theta}^{\omega})$ as the estimated label of the EM algorithm under permutation ω. Our goal is to minimize

$$\sum_{i=1}^{n}\sum_{j=1}^{K}\rho(z_{ij}, \hat{z}_{ij}(\boldsymbol{\theta}^{\omega})), \tag{2.13}$$

where ρ is a loss function. There are two approaches in Yao (2015).

Method 1: Complete likelihood based labeling (COMPLH):
We intend to find ω by maximizing

$$L(\hat{\boldsymbol{\theta}}^{\omega}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{K}\{\lambda_j^{\omega} f(\mathbf{x}_i|\boldsymbol{\theta}_j^{\omega})\}^{z_{ij}}, \tag{2.14}$$

through ω, which is equivalent of maximizing

$$\ell_1(\hat{\boldsymbol{\theta}}^{\omega}|\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log\hat{z}_{ij}(\boldsymbol{\theta}^{\omega}). \tag{2.15}$$

Method 2: Distance based labeling (DISTLAT):

We intend to find ω by minimizing

$$\ell_2(\hat{\boldsymbol{\theta}}^{\omega}|\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{K}\{\hat{z}_{ij}(\boldsymbol{\theta}^{\omega}) - z_{ij}\}^2, \tag{2.16}$$

through ω, which is equivalent of maximizing

$$\ell_1(\hat{\boldsymbol{\theta}}^{\omega}|\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\hat{z}_{ij}(\boldsymbol{\theta}^{\omega}). \tag{2.17}$$

2.2.4 Unboundedness of Log-likelihood

The log-likelihood of mixture models may be unbounded for some cases. One classic example is the normal mixture model with unequal variances:

$$L(\boldsymbol{\theta}|x) = \sum_{i=1}^{n}\lambda_1 g(x_i|\mu_1,\sigma_1^2) + \sum_{i=1}^{n}\sum_{j=2}^{K}\lambda_j g(x_i|\mu_j,\sigma_j^2),$$

where $\boldsymbol{\theta} \in \Theta = \{(\lambda_1,\ldots,\lambda_K,\mu_1,\ldots,\mu_K,\sigma_1^2,\sigma_2^2) : \sigma_j^2 > 0, \lambda_j \in (0,1), \sum_{j=1}^{K}\lambda_j = 1\}$ and g is the density function for the standard normal distribution. When $\mu_1 = x_1$ and $\sigma_1^2 \to 0$, $L(\boldsymbol{\theta}|x) \to \infty$ accordingly. One approach to mitigate this issue is to apply the profile likelihood methods. For example, Dempster et al. (1977) ran the EM algorithm over a constrained parameter space:

$$\Theta_\eta = \{\boldsymbol{\theta} \in \Theta : \sigma_h/\sigma_j \geq \eta > 0, 1 \leq h \neq j \leq K\},$$

where $\eta \in (0,1]$. One big challenge is to select the number of η. If η is too small, there might be a chance that some boundary point, which satisfies $\sigma_h/\sigma_j = \eta$, will maximize

the log-likelihood and the solution will depend on the choice of η. If η is too large, it is likely that the solution is a local maximum instead of a global maximum.

In practice, this issue is very rare, mostly due to an inappropriate starting value. Based on our empirical experience, if we start the algorithm from a reasonable initial value, such as the maximum likelihood estimate assuming all components are normal with equal variance, the unboundedness issue is very rare. Consequently, many existing algorithms avoids searching on this constrained subspace and contains a restarting procedure if the log-likelihood goes to infinity after some iterations, e.g. Benaglia et al. (2009).

2.3 Mixture of regressions Models

There is a rich history associated with the FMR models. It was first studied by Quandt (1972) as the *switching regression* in econometric literature. Later it was often used to compare the structural change in a system, e.g. Quandt and Ramsey (1978).

Let $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ be the sample we observed for the mixture of regressions model ($n \gg Kp + K - 1$). In this section, we take a special case of model (1.3) in which g_j is a normal distribution with standard error σ_j. We rewrite the likelihood as:

$$f(y_i, x_i|\boldsymbol{\psi}) = \sum_{j=1}^{K} \lambda_j (2\pi\sigma_j^2)^{-1/2} \exp\{-\frac{1}{2\sigma_j^2}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)^2\}. \tag{2.18}$$

2.3.1 EM-algorithm

The FMR model can also be estimated via maximum likelihood estimation and the EM algorithm. We define the missing value $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T \in \mathbb{R}^{n \times K}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})^T$ $(i = 1, \ldots, n)$ is a K-dimensional indicator vector with its j-th element

given by

$$z_{ij} = \begin{cases} 1 & \text{if } (\mathbf{x}_i, y_i) \text{ belongs to } j\text{-th group;} \\ 0 & \text{otherwise .} \end{cases}$$

Consequently, the complete log-likelihood for equation (2.18) is:

$$\ell_c(\boldsymbol{\psi}|\mathbf{X}, \mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\{\log \lambda_j (2\pi\sigma_j^2)^{-1/2} - (y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j)^2\}. \tag{2.19}$$

Algorithm 2.2. *First we give random initial values to the unknown parameters* $\boldsymbol{\psi}^0$ *and* $z_{ij}^{(0)}$*. Then in t-th iteration, it consists the following three steps.*

E-step: *Given* $\boldsymbol{\psi}^{(t)}$ *, we calculate*

$$z_{ij}^{(t+1)} = E(Z_{ij}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\psi}^{(t)}) = \frac{\lambda_j^{(t)}(2\pi\sigma_j^{2(t)})^{-1/2}\exp\{-\frac{1}{2\sigma_j^{2(t)}}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)})^2\}}{\sum_{h=1}^{K}\lambda_h^{(t)}(2\pi\sigma_h^{2(t)})^{-1/2}\exp\{-\frac{1}{2\sigma_h^{2(t)}}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_h^{(t)})^2\}}, \tag{2.20}$$

for $i = 1, \ldots, n, j = 1, \ldots, K$*.*

M-step 1: *Update* λ *simply through*

$$\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(t+1)}, j = 1, \ldots, K. \tag{2.21}$$

M-step 2: *Update* $\boldsymbol{\beta}_j$*'s through*

$$\boldsymbol{\beta}_j^{(t+1)} = (\boldsymbol{X}^T W_j^{(t+1)}\boldsymbol{X})^{-1}\boldsymbol{X}^T W_j^{(t+1)}\boldsymbol{y}, j = 1, \ldots, K, \tag{2.22}$$

where $W_j^{(t+1)} = diag(z_{1j}^{(t+1)}, \ldots, z_{nj}^{(t+1)})$*.*

M-step 3: *Update σ_j's through*

$$\sigma_j^{(t+1)} = \frac{||W_j^{1/2(t+1)}(\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\beta}_j^{(t+1)})||}{tr(W_j^{1/2(t+1)})}, \tag{2.23}$$

for $j = 1, \ldots, K$.

The algorithm is terminated if either t_{max} of iterations has been reached, or if $\ell^{(t+1)} - \ell^{(t)} < \epsilon = 10^{-8}$.

2.3.2 Robust regression methods

Consider a classic regression model

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i, \tag{2.24}$$

where ϵ_i's are independent and identically distributed as $N(0, \sigma^2)$. Estimating $\boldsymbol{\beta}$ via a maximum likelihood estimator is equivalent to an ordinary least square (OLS) solution:

$$\hat{\boldsymbol{\beta}} = \underset{b\beta}{\operatorname{argmin}} \sum_{i=1}^{n} r_i^2, \tag{2.25}$$

where $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$ corresponds to the i-th residual.

It is well-known that MLEs can be very sensitive to outliers. Sometimes a single outlier could ruin the entire estimation. One approach is to use the least trimmed squares (LTS): (see Rousseeuw (1985) for a detailed description)

$$\hat{\boldsymbol{\beta}}_{LTS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=s+1}^{n} r_{(i)}^2, \tag{2.26}$$

where $r^2_{(i)}$ is the i-th order statistics of r^2_i. The constant s is known as the trimming constant which satisfies $0 < s < \frac{n}{2}$. Through trimming, we sacrifice some efficiency but gain the robustness against outliers. This approach does not work for small sample size, as the results maybe misleading.

2.4 Local Polynomial Regression

Regression models are widely used to analyze the relation between the response y and the covariate x. One typical example is the simple linear regression model, which usually takes the form of:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where $E(\epsilon|X) = 0$. These linear regression models may fail to characterize the form of a curved relation between x and y. To relax the linear assumption, nonparametric methods come to the rescue. There is a rich history of nonparametric local regression, e.g. the Nadaraya-Watson estimator proposed by Nadaraya (1964) and Watson (1964), which is known as the kernel regression estimator.

In this dissertation, we focus our interest on local polynomial regressions (Fan and Gijbels (1996)). We relax the linear assumption to the nonparametric regression form:

$$y = m(x) + \epsilon,$$

where $E(\epsilon|X) = 0$. Suppose we are interested in $m(x)$ at the local point x_0. Then, by applying a Taylor expansion of $m(x)$ in the neighborhood of x_0, we approximate $m(x)$

by:

$$m(x) \approx \mathbf{x}^T \boldsymbol{\beta} = \beta_0 + \sum_{j=1}^{p} \beta_j (x - x_0)^j,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ and $\beta_j = m^{(j)}(x_0)/j!$. The observations, which are closer to x_0, are more trustworthy for the estimation of $m(x_0)$. The observations which are away from x_0 are less trustworthy for the estimation of $m(x_0)$. Intuitively, this suggests a weighted regression model for the local features by minimizing a weighted square error:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j (x_i - x_0)^j)^2 \mathcal{K}_h(x_i - x_0), \tag{2.27}$$

where $\mathcal{K}_h(\cdot) = h^{-1}\mathcal{K}(\cdot/h)$. Such $\mathcal{K}$'s are called kernel functions which satisfy:

$$\mathcal{K}(u) \geq 0, \quad \int \mathcal{K}(u)du = 1, \quad \text{and} \quad \mathcal{K}(u) = \mathcal{K}(-u),$$

Popular choices of kernels include, but are not limited to the following choices:

- Gaussian kernel

$$\mathcal{K}(u) = \frac{1}{\sqrt{2\pi}} \exp\{-u^2/2\}; \tag{2.28}$$

- Cosine kernel

$$\mathcal{K}(u) = \frac{1}{2}(1 + \cos(\pi u)); \tag{2.29}$$

- Symmetric Beta family kernel

$$\mathcal{K}(u) = \frac{1}{Beta(1/2, \gamma + 1)} (1 - u^2)_+^{\gamma}, \tag{2.30}$$

where $\gamma = 0, 1, 2, 3$, which corresponds to uniform, Epanechnikov, biweight and

triweight kernel, respectively.

If we further denote

$$W = diag\{\mathcal{K}_h(x_1 - x_0), \ldots, \mathcal{K}_h(x_n - x_0)\},$$

$$X = \begin{pmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ 1 & x_2 - x_0 & \cdots & (x_2 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{pmatrix}.$$

Then the solution of (2.27) is:

$$\boldsymbol{\beta} = \boldsymbol{\beta}(x_0) = (X^T W X)^{-1} X^T W y. \tag{2.31}$$

Selecting the smoothing bandwidth, h, is challenging. The bandwidth h controls the smoothness of the estimation. If h is too small, then only a few observations are included in the local feature estimation. If h goes to ∞, we are just doing a global estimation. In other words, we are just fitting a linear regression model. To select the best bandwidth, we need to use T-folder cross-validation (CV). We partition the full data set $\mathcal{D}$ into the training data set $\mathcal{R}_t$ and the test data set $\mathcal{T}_t$ for $t = 1, \ldots, T$. We select h by minimizing:

$$CV(h) = \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} (y_i - \hat{y}_i^{(t)})^2, \tag{2.32}$$

where $\hat{y}_i^{(t)}$ is obtained by the estimated coefficients through $\mathcal{R}_t$ for $i \in \mathcal{T}_t$.

2.5 Log-concave Shape Constraint

It is well-known that for a linear regression model, when estimating the coefficients and obtaining the inferences, we usually assume that the residuals are normally distributed. However, this assumption may be too restrictive and the parameter estimation may be biased if the parametric family is misspecified. To overcome this problem, nonparametric shape constraints are becoming increasingly popular. In this paper, we make one shape constraint instead of a specific parametric assumption for each component density. We assume each component density g_j to be log-concave. A density $g(x)$ is log-concave if its log-density, $\phi(x) = \log g(x)$, is concave. Examples of log-concave densities include, but are not limited to normal, Laplace, chi-square, logistic, gamma with shape parameter greater than 1, and beta distribution with both parameters greater than 1.

2.5.1 Log-concave Maximum Likelihood Estimator of density estimations and its theoretical properties

Dümbgen et al. (2011) proposed an estimator by maximizing a log-likelihood-type functional:

$$L(\phi, Q) = \int \phi dQ - \int \exp\{\phi(x)\}dx + 1. \tag{2.33}$$

The profiled log-likelihood is:

$$L(Q) = \sup L(\phi, Q). \tag{2.34}$$

The maximizer is called Log-concave Maximum Likelihood Estimator (LCMLE). If ψ

is the LCMLE for a fixed Q such that $L(\phi, Q) = L(Q) \in \mathbb{R}$, then,

$$\int e^{\psi(x)} dx = 1,$$

which guarantees that ψ is a log-density.

Dümbgen et al. (2011) also shows that under fairly general conditions, the maximizer $\psi = \psi(\cdot|Q)$ exists uniquely.

Theorem 2.3. *For any $Q \in \mathcal{Q}$, $L(Q)$ is finite if Q satisfies*

$$(A1) \int ||x|| dQ < \infty \quad \textit{and} \quad (A2)\ \textit{interior}(\textit{csupp}((Q)) \neq \emptyset.$$

Moreover, it exist a unique function

$$\psi = \psi(\cdot|Q) = \underset{\phi \in \Phi}{\operatorname{argmax}}\, L(\phi, Q),$$

such that $\int e^{\psi(x)} dx = 1$.

We further suppose a sequence Q_n (corresponding with the maximizer ψ_n) converges to the true density Q_0 (corresponding to the maximizer ψ_0) in Mallows distance D_1: $D_1(Q, Q') = \inf_{(X,X')} E||X - X'||$, where Q and Q' are two distributions and the infimum is taken over all pairs of (X, X') such that $X \sim Q$ and $X' \sim Q'$. The convergence with respect to Mallows distance, i.e. $\lim_{n\to\infty} D_1(Q_n, Q) = 0$, is equivalent with Q_n weakly converges to Q_0, denoted by $Q_n \to^w Q$ and $\int \|x\| dQ_n(x) \to \int \|x\| dQ(x)$ as $n \to \infty$. Then, the estimator ψ_n is an consistent estimator of ψ_0:

Theorem 2.4. *Let Q_n be a sequence such that $\lim_{n\to\infty} D_1(Q_n, Q_0) = 0$ for some Q_0.*

Then,

$$\lim_{n\to\infty} L(Q_n) = L(Q_0).$$

Let ϕ_n be the maximizer corresponding to profile log-likelihood $L(Q_n)$, $f_n(x) = \exp\{\phi_n(x)\} = f(\cdot|Q_n)$. For $f_0(x) = f(\cdot|Q_0)$, we have:

$$\lim_{n\to\infty, x\to y} f_n(x) = f_0(y) \quad \textit{for all} \quad y \notin \partial\{f_0 \geq 0\}, \tag{2.35}$$

$$\lim_{n\to\infty, x\to y} f_n(x) \leq f_0(y) \quad \textit{for all} \quad y \in \mathbb{R}^d, \tag{2.36}$$

$$\lim_{n\to\infty} \int |f_n(x) - f_0(x)|dx = 0. \tag{2.37}$$

For more details and other properties, reader may refer to Theorem 2.2 and Theorem 2.15 of Dümbgen et al. (2011).

2.5.2 Computational Aspect of LCMLE

The theorem below (Theorem 2.1 of Dümbgen and Rufibach (2009)) shows the existence of $\hat{f}_n$ in the sample version (a special case of Theorem 2.3 with $Q = Q_n$):

Theorem 2.5. *Suppose $n \geq d+1$. Then, $\hat{f}_n$ exists uniquely with probability 1 on the convex hull of the data denoted by $C_n = conv(X_1, \ldots, X_n)$. Moreover, $\log \hat{f}_n$ is a tent function $\bar{h}_y(x)$ for given $y = (y_1, \ldots, y_n)$, i.e. $\bar{h}_y$ is the least concave function that $\bar{h}_y(X_i) \geq y_i$ for all $i = 1, \ldots, n$.*

In the proof of Theorem 2.5, it is shown that the actual maximizer $\log \hat{f}_n$ is a "tent" function. It is shown that:

(i) The maximizer $\log \hat{f}_n$ is also a log-density and is supported on the convex hull of

the data, i.e.

$$\hat{f}_n(x) \begin{cases} > 0 & \text{for } x \in C_n; \\ = 0 & \text{for } x \notin C_n. \end{cases}$$

(ii) $\log \hat{f}_n \in \mathcal{H}$, i.e. the log-function of the maximizer is a "tent" function, and $\log \hat{f}_n$ is closed.

(iii) $\hat{f}_n \in \mathcal{F}_0$, i.e. the maximizer is a log-density.

(iv) There exists $M > 0$ such that if $\max_i |\bar{h}_y(X_i)| \geq M$, then $\ell_n(\exp(\bar{h}_y)) \leq \ell_n(\hat{f}_n)$, i.e. the maximum of the log-density is bounded.

For the detailed proof, readers may refer to Dümbgen and Rufibach (2009). Based on Theorem 2.5, Cule et al. (2010) provided the methodologies of estimating the LCMLEs. It is shown that 2.33 can be rewritten as:

$$\underset{y \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} \tau(y) = -\sum_{i=1}^{n} \omega_i \bar{h}_y(X_i) + \int_{C_n} \exp(\bar{h}_y(x)) dx, \tag{2.38}$$

where $\bar{h}_y(x) = \inf\{h(x) : h \text{ is concave}, h(X_i) \geq y_i \text{ for } i = 1, \ldots, n\}$. The ω_i's are the weights of each observation when estimating the log-concave density, where $\omega_i = 1/n$ in most cases.

However, τ is non-convex and finding it is extremely computationally intensive even for small sample size. Alternatively, Cule et al. (2010) proposed

$$\underset{y \in \mathbb{R}^{n \times d}}{\operatorname{argmin}} \sigma(y) = -\sum_{i=1}^{n} \omega_i y_i + \int_{C_n} \exp(\bar{h}_y(x)) dx. \tag{2.39}$$

Theorem 2.6. *σ is convex and $\sigma \geq \tau$. It has a unique minimum y^* such that $\log \hat{f}_n = \bar{h}_{y^*}$.*

As σ is non-differentiable, there is no gradient based optimization method to find σ. However, sub-gradient methods are still valid. For the univariate case, one may use the R package `logcondens` (Dümbgen et al. (2010)). For multivariate density estimation, one may use `LogConcDEAD` (Cule et al. (2009)), which uses the Shor's r-algorithm (Kappel and Kuntsevich (2000)). For details of sub-gradient optimization techniques, see the Appendix of Cule et al. (2010). In Figure 2.2, we plotted the estimated log-concave density via `LogConcDEAD` for four different log-concave densities. We observe that even for a finite sample size of 400, these LCMLEs approximate the true densities precisely.

2.5.3 Application to Simple Linear Regression

For linear regression with log-concave error density, Dümbgen et al. (2011) proposed an estimator by maximizing:

$$\hat{L}(\phi, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i - \mathbf{x}_i^T\boldsymbol{\beta}) - \int \exp\{\phi(x)\}dx + 1. \tag{2.40}$$

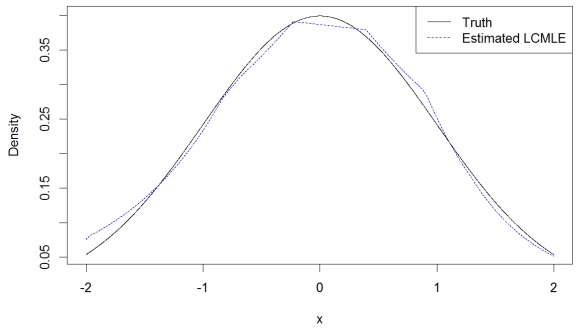
The estimator of (2.40) is consistent under fairly general conditions. For more details, readers may refer to Theorem 3.6 of Dümbgen et al. (2011). To implement (2.40), Dümbgen et al. (2011) proposed several algorithms, one of which is the following algorithm:

Algorithm 2.7. *Initialize $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^0$ which satisfied $\sum(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^0) = 0$ and estimate the log-density $\phi^{(0)}$. In t-th iteration, it contains the following three steps*

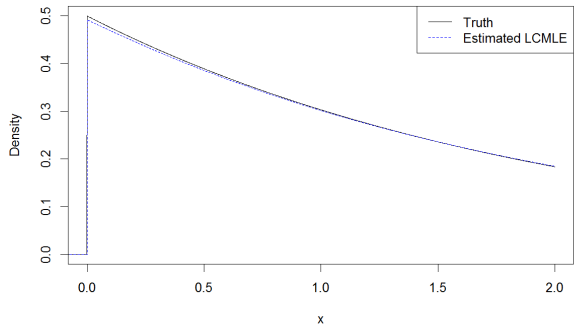
(A) *Determine*

$$\tilde{\boldsymbol{\beta}}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\beta}} \hat{L}(\phi^{(t)}, \boldsymbol{\beta}). \tag{2.41}$$

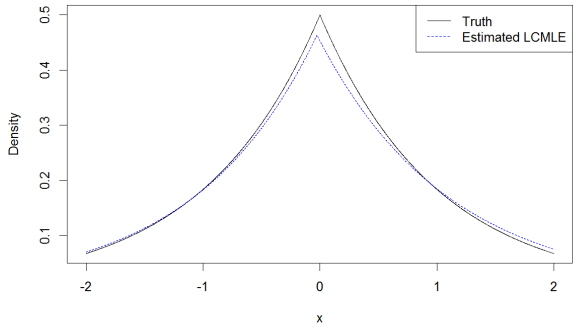
(B) *Shift the intercept of $\tilde{\boldsymbol{\beta}}^{(t+1)}$ so that the residuals have zero mean and obtain*



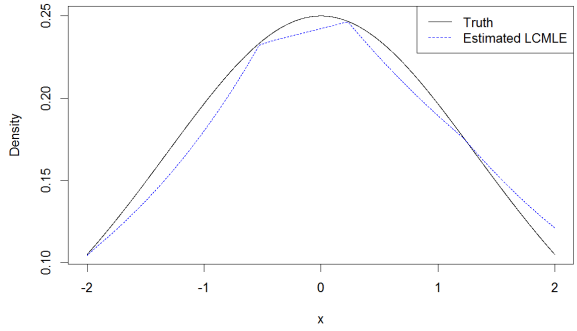
(a) N(0,1)



(b) Gamma(1,2)



(c) Laplace(0,1)



(d) Logistic(0,1)

Figure 2.2: Estimated LCMLEs for four different log-concave densities. Solid line represents the truth and dashed line represents the estimation results. For the finite sample size of 400, these LCMLEs approximates the true density well.

$\hat{\boldsymbol{\beta}}^{(t+1)} = (\hat{\beta}_0^{(t+1)}, \tilde{\beta}_1^{(t+1)} \ldots, \tilde{\beta}_{p-1}^{(t+1)})$, *where*

$$\hat{\beta}_0^{(t+1)} = \tilde{\beta}_0^{(t+1)} + c^{(t+1)} \quad \textit{with} \quad c^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}^{(t+1)}). \tag{2.42}$$

(C) *Determine*

$$\phi^{(t+1)} \leftarrow \arg\max_{\phi} \hat{L}(\phi, \boldsymbol{\beta}^{(t+1)}). \tag{2.43}$$

The algorithm is terminated if either the increasing amount of the log-likelihood value is smaller than a threshold or if the maximal number of iterations has been reached.

Chapter 3

Maximum Likelihood estimation of mixture of log-concave densities

3.1 Introduction

Recall the traditional parametric model takes the form of (1.1)

$$f(x) = \sum_{j=1}^{K} \lambda_j g_j(x; \theta_j) \quad x \in \mathbb{R}^d,$$

where $\lambda_1, \ldots, \lambda_K$ are the mixing proportions and $g_j(x; \theta_j)$'s are component densities. The unknown parameters in the mixture model (1.1) can be estimated by the EM algorithm. One major drawback of the traditional mixture model like (1.1) is the strong parametric assumption about the component density g_j. It is often too restrictive and the density estimation may be inaccurate due to the model misspecification. Another drawback is that each model requires a specific EM algorithm based on the parametric assumption.

To relax the parametric assumption, nonparametric shape constraints are becoming

increasingly popular. In this paper, we make one fairly general shape constraint for our mixture model. We assume that each component density is log-concave. A density g is log-concave if $\log g$ is concave. Examples of log-concave densities include normal, Laplace, logistic, as well as gamma and beta with certain parameter constraints. Log-concave densities have lots of nice properties as described by Balabdaoui et al. (2009). Their nonparametric maximum likelihood estimators were studied by Dümbgen and Rufibach (2009), Cule et al. (2010), Cule and Samworth (2010), Chen and Samworth (2013), Pal et al. (2007) and Dümbgen et al. (2011) (referred as [DSS 2011] thereafter). The convergence rates of these estimators for log-concave densities were studied by Doss and Wellner (2013) and Kim and Samworth (2014). Such estimators provide more generality and flexibility without any tuning parameters.

In our model, we assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent d-dimensional random variables with distribution Q_0 and the mixture density f_0. The mixture density f_0 belongs to a given class

$$\mathcal{F} = \{f : f(x) = \sum_{j=1}^{K} f_j(x) = \sum_{j=1}^{K} \lambda_j \exp\{\phi_j(x)\}, \boldsymbol{\lambda} \in \Lambda, \boldsymbol{\phi} \in \Phi\}, \tag{3.1}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K)$, $\Lambda = \{(\lambda_1, \ldots, \lambda_K) : 0 < \lambda_j < 1, \sum_{j=1}^{K} \lambda_j = 1\}$, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, and $\Phi = \{(\phi_1, \ldots, \phi_K) : \phi_j \text{ is concave}\}$. We assume that each ϕ_j is continuous and is coercive in the sense that $\phi_j(x) \to -\infty$ as $||x|| \to \infty$ $(j = 1, \ldots, K)$.

One issue for mixture models is that the likelihood might be unbounded in some cases. For example, the likelihood function for a normal mixture takes the form of $L(\boldsymbol{\theta}|x) = \sum_{i=1}^{n}(\lambda g(x_i|\mu_1, \sigma_1^2) + (1-\lambda)g(x_i|\mu_2, \sigma_2^2))$, where $\boldsymbol{\theta} = \{(\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) : \sigma_1^2, \sigma_2^2 > 0, \lambda \in (0, 1)\}$ and g is the density function for the standard normal distribution. When $\mu_1 = x_1$ and $\sigma_1^2 \to 0$, $L(\boldsymbol{\theta}|x) \to \infty$ (see Section 3.10 of McLachlan and Peel (2000) for detailed

discussions). Many methods have been proposed to solve the unboundedness issue of mixture likelihood, see for example, Hathaway (1985), Chen et al. (2008), and Yao (2010). Note that, similar to traditional normal mixture models with unequal variances, the likelihood functions for mixtures of log-concave densities are unbounded as well. Thus, similar to Hathaway (1985), we will define LCMLE on a constrained parameter space. Let $M_j(\boldsymbol{\phi}) = \max_{x\in\mathbb{R}^d}\{\phi_j(x)\}$, $M_{(1)}(\boldsymbol{\phi}) = \min_j\{M_j(\boldsymbol{\phi})\}$, and $M_{(K)}(\boldsymbol{\phi}) = \max_j\{M_j(\boldsymbol{\phi})\}$. We further define the ratio $\mathcal{S}(\boldsymbol{\phi}) = M_{(1)}(\boldsymbol{\phi})/M_{(K)}(\boldsymbol{\phi})$. Here, we borrow the idea of Hathaway (1985) by restricting our interest to a constrained subspace $\boldsymbol{\Phi}_\eta$ such that $\boldsymbol{\Phi}_\eta = \{\boldsymbol{\phi} \in \boldsymbol{\Phi} : |\mathcal{S}(\boldsymbol{\phi})| \geq \eta > 0\}$ for some $\eta \in (0, 1]$. This restriction avoids estimating the case that the modes of different components differ a lot. By restricting on $\boldsymbol{\Phi}_\eta$, we focus our interest on $f \in \mathcal{F}_\eta$, where

$$\mathcal{F}_\eta = \{f : f(x) = \sum_{j=1}^{K} f_j(x) = \sum_{j=1}^{K} \lambda_j \exp\{\phi_j(x)\}, \boldsymbol{\lambda} \in \Lambda, \boldsymbol{\phi} \in \Phi_\eta\}. \tag{3.2}$$

Let Q_n be the empirical distribution of $X_1, \ldots, X_n$. The (restricted) log-concave maximum likelihood estimator (LCMLE) is

$$f_n = f(\cdot|Q_n) = \underset{f\in\mathcal{F}_\eta}{\operatorname{argmax}} \int \log(f) dQ_n. \tag{3.3}$$

In practice, similar to Hathaway (1985), picking η can be tricky for some extreme case. If η is too small, there might be a chance that some boundary point $|\mathcal{S}(\boldsymbol{\phi})| = \eta$ maximizes the log-likelihood and the solution will depend on the choice of η. In this paper, we do not focus on the issue of choosing η. The constrained subspace $\boldsymbol{\Phi}_\eta$ is mainly used for theoretical development. Based on our empirical experience, if we start the algorithm from a reasonable initial value, such as the maximum likelihood estimate assuming all

components are normal with equal variance, the unboundedness issue is very rare.

Many methods have been proposed to relax the parametric assumption of (1.1). Hunter et al. (2007), Bordes et al. (2006b), Butucea and Vandekerkhove (2014), and Chee and Wang (2013) considered the extension of (1.1) by assuming all component densities are symmetric but unknown. Bordes et al. (2006a), Bordes and Vandekerkhove (2010), Hohmann and Holzmann (2013), Xiang et al. (2014), and Ma and Yao (2015) considered the extension of (1.1) when $K = 2$ and one of the component densities is symmetric but unknown. Mixtures of log-concave densities have been studied by Chang and Walther (2007), Cule et al. (2010) and Balabdaoui and Doss (2014). Chang and Walther (2007) provided an EM-type algorithm and demonstrated sound numerical results in the simulation study. Cule et al. (2010) applied the log-concave mixture model to the Wisconsin breast cancer data set. Balabdaoui and Doss (2014) considered a special case when all components have the same symmetric log-concave densities but with different location parameters, and proved the $\sqrt{n}$-consistency of their proposed M-estimators for mixing proportion as well as location parameters. Note that these models are special cases from the family of $\mathcal{F}$. Therefore, their estimators and asymptotic results cannot be applied here. For example, the mixture of normal distributions with different component means and variances belongs to $\mathcal{F}$ but does not belong to the model family considered by Balabdaoui and Doss (2014).

In this section, we show that theoretically, the LCMLE (in the restricted subset $\mathcal{F}_\eta$) exists, and is consistent under fairly general conditions. However, we want to point out that the extension of the properties of the log-concave density to mixtures of log-concave densities is not trivial. The log-density $l_n = l(\cdot|Q_n) = \log f_n$ is no longer guaranteed to be a concave function. Consequently, many nice theoretical properties stated in DSS 2011 no longer hold for our mixture model.

3.2 Log-concave maximum likelihood estimator

Let $\mathcal{Q} = \mathcal{Q}(d)$ be the family of all distributions Q on $\mathbb{R}^d$. Our goal is to maximize a log-likelihood-type functional:

$$L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q) = \int \log[\sum_{j=1}^{K} \lambda_j \exp\{\phi_j(x)\}] dQ(x) - \sum_{j=1}^{K} \pi_j (\int \exp\{\phi_j(x)\} dx - 1), \quad (3.4)$$

where π_j's are Lagrange multipliers to incorporate the constraint $\int \exp\{\phi_j(x)\} dx = 1$ $(j = 1, \ldots, K)$. We define a profile log-likelihood:

$$L(Q) = \sup_{\boldsymbol{\phi} \in \Phi_\eta, \boldsymbol{\lambda} \in \Lambda, \boldsymbol{\pi}} L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q). \quad (3.5)$$

If, for fixed Q, $(\boldsymbol{\psi}, \boldsymbol{\lambda}^*, \boldsymbol{\pi}^*)$ maximizes $L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q)$, it will automatically satisfy that:

$$\pi_j^* = E(\pi(j|x)) = \int \frac{\lambda_j^* \exp\{\psi_j(x)\}}{(\sum_{h=1}^{K} \lambda_h^* \exp\{\psi_h(x)\})} dQ(x); \quad (3.6)$$

$$\int \exp\{\psi_j(x)\} dx = 1 \quad (j = 1, 2, \ldots, K). \quad (3.7)$$

Note that differing from the non-mixture setting in DSS 2011, π_j^* is not equal to 1.

To verify this, note that $\boldsymbol{\phi} + \boldsymbol{c} \in \Phi$ for any fixed vector of functions $\boldsymbol{\phi} \in \Phi$ and arbitrary $\boldsymbol{c} = (c_1, \ldots, c_K)^T \in \mathbb{R}^K$, and

$$\frac{\partial L(\boldsymbol{\psi} + \boldsymbol{c}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q)}{\partial c_h}|_{\boldsymbol{c}=0} = (\int \frac{\lambda_h \exp\{\psi_h(x)\}}{\sum_{j=1}^{K} \lambda_j \exp\{\psi_j(x)\}} dQ(x) - \pi_h \int e^{\psi_h(x)} dx) = 0,$$

$$\frac{\partial L(\boldsymbol{\psi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q)}{\partial \pi_h} = 1 - \int \exp\{\psi_h(x)\} dx = 0.$$

The maximizer $(\boldsymbol{\psi}, \boldsymbol{\lambda}^*)$ forms the log-likelihood maximizer $l^*(x) = \log \sum_{j=1}^{K} \lambda_j^* e^{\psi_j(x)}$.

3.3 Theoretical Properties

Before we state the main theories, we first define the convex support of a distribution.

Definition For any distribution Q, let $Q(C)$ be the probability measure of the set C. The convex support of Q is the set such that:

$$csupp(Q) = \bigcap \{C : C \subseteq \mathbb{R}^d \text{ closed and convex}, Q(C) = 1\}.$$

The convex support is itself closed and convex with $Q(csupp(Q)) = 1$.

In the following text, we define:

$\mathcal{Q}^1 = \{Q \in \mathcal{Q} : \int ||x|| dQ < \infty\}$, (we define $||x||$ as Euclidean norm in our paper).

$\mathcal{Q}^0 = \{Q \in \mathcal{Q} : interior(csupp(Q)) \neq \emptyset\}$.

Theorem 3.1. *For any $Q \in \mathcal{Q}^1 \cap \mathcal{Q}^0$, the value of $L(Q)$ is real and there exists a maximizer:*

$$(\boldsymbol{\psi}, \boldsymbol{\lambda}^*, \boldsymbol{\pi}^*) = \underset{\boldsymbol{\phi} \in \Phi_\eta, \boldsymbol{\lambda} \in \Lambda, \boldsymbol{\pi}}{\operatorname{argmax}} L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q) \text{ such that } \int e^{\psi_j(x)} dx = 1 \quad for \quad j = 1, \ldots, K.$$

Proposition 3.2. *The profiled log-likelihood is only real if $Q \in \mathcal{Q}^1 \cap \mathcal{Q}^0$. Moreover it can be proved that:*

$$L(Q) = \begin{cases} -\infty & \text{if } Q \in \mathcal{Q} \setminus \mathcal{Q}^1; \\ \infty & \text{if } Q \in \mathcal{Q}^1 \setminus \mathcal{Q}^0. \end{cases}$$

Proposition 3.3. *Suppose X has distribution $Q \in \mathcal{Q}^1 \cap \mathcal{Q}^0$. For arbitrary $a \in \mathbb{R}^d$ and nonsingular $B \in \mathbb{R}^{d \times d}$, define $Q_{a,B}$ be the distribution of $a + BX$. Then $Q_{a,B} \in \mathcal{Q}^1 \cap \mathcal{Q}^0$,*

$$L(Q_{a,B}) = L(Q) - \log |\det B|,$$

and

$$l^*(x|Q_{a,B}) = l^*(B^{-1}(x-a)|Q) - \log|\det B|.$$

Proposition 3.4. *For $Q \in \mathcal{Q}^1 \cap \mathcal{Q}^0$, the profile log-likelihood is convex. More detailedly, for any $0 < t < 1$,*

$$L((1-t)Q_0 + tQ_1) \leq (1-t)L(Q_0) + tL(Q_1).$$

The equality honds if and only if $l(\cdot|Q_0) = l(\cdot|Q_1)$.

Next, we establish the consistency of the estimated mixture density. In the following, we refer to the concept of convergence of distribution as converging with respect to Mallows distance D_1: $D_1(Q, Q') = \inf_{(X,X')} E||X - X'||$, where Q and Q' are two distributions and the infimum is taken over all pairs of (X, X') such that $X \sim Q$ and $X' \sim Q'$. The convergence with respect to Mallows distance, i.e. $\lim_{n\to\infty} D_1(Q_n, Q) = 0$, is equivalent with Q_n weakly converges to Q_0, denoted by $Q_n \to^w Q$ and $\int \|x\| dQ_n(x) \to \int \|x\| dQ(x)$ as $n \to \infty$.

Theorem 3.5. *Let $Q_n \in \mathcal{Q}^1 \cap \mathcal{Q}^0$ be a sequence such that $\lim_{n\to\infty} D_1(Q_n, Q_0) = 0$ for some $Q_0 \in \mathcal{Q}^1 \cap \mathcal{Q}^0$. Then,*

$$\lim_{n\to\infty} L(Q_n) = L(Q_0).$$

Let ϕ_{nj}'s and λ_{nj}'s be the maximizer corresponding to profile log-likelihood $L(Q_n)$, i.e,

$f_n(x) = \sum \lambda_{nj} \exp\{\phi_{nj}(x)\} = f(\cdot|Q_n) \in \mathcal{F}_\eta$. *For* $f_0(x) = f(\cdot|Q_0) \in \mathcal{F}_\eta$, *we have:*

$$\lim_{n\to\infty, x\to y} f_n(x) = f_0(y) \quad \textit{for all} \quad y \notin \partial\{f_0 \geq 0\}, \tag{3.8}$$

$$\lim_{n\to\infty, x\to y} f_n(x) \leq f_0(y) \quad \textit{for all} \quad y \in \mathbb{R}^d, \tag{3.9}$$

$$\lim_{n\to\infty} \int |f_n(x) - f_0(x)| dx = 0. \tag{3.10}$$

The above theorem showed the consistency of the estimated mixture density. If we further assume that the true mixture density $f_0(x)$ is identifiable, then each estimated component density and mixing proportions are also consistent. We will discuss more about the identifiability issue in Section 3.7.

3.4 EM-type algorithm

The EM algorithm for estimating log-concave mixture densities has already been developed by Chang and Walther (2007). Here we briefly summarize it. First we randomly generate initial values for the normal mixture EM-algorithm and run the normal mixture EM-algorithm until convergence, which will provide a good initial value. Then we use the outcome as the starting values for our EM-type algorithm. We assume the observed data $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times d}$ to be incomplete and define the missing value $\mathbf{Z} = (\mathbf{z}_1, ..., \mathbf{z}_n)^T$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^T \in \mathbb{R}^d$ and $\mathbf{z}_i$ is a K-dimension vector with its j-th element given by:

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to } j\text{th group,} \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, n,\ j = 1, \ldots, K$. So the complete log-likelihood is:

$$\log f(\boldsymbol{\phi}, \boldsymbol{\lambda}; \mathbf{X}, \mathbf{Z}) = \log \prod_{i=1}^{n} \prod_{j=1}^{K} [\lambda_j e^{\phi_j(\mathbf{x}_i)}]^{z_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} [\log \lambda_j + \phi_j(\mathbf{x}_i)],$$

In E-step, we replace z_{ij} by

$$z_{ij}^{(t+1)} = \frac{\lambda_j^{(t)} e^{\widehat{\phi}_j^{(t)}(x_i)}}{\sum_{h=1}^{K} \lambda_h^{(t)} e^{\widehat{\phi}_h^{(t)}(x_i)}}.$$

In M-step, first we update λ by $\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(t+1)}, j = 1, \ldots, K$. Then we update ϕ_j by maximizing $\sum_{i=1}^{n} z_{ij}^{(t+1)} \phi_j(x_i)$ with respect to ϕ_j through the function called `mlelcd` in the R package `LogConcDEAD` (Cule et al. (2009)) and get estimator $\widehat{\phi}_j^{(t+1)}$ for $j = 1, \ldots, K$. The estimation of $\widehat{\phi}_j$ has been studied by Walther (2002) and Rufibach (2007). Given i.i.d. data $X_1, \ldots, X_n$ from distribution f, the Log-concave Maximum Likelihood Estimator (LCMLE) $\widehat{f}_n$ exists uniquely and has support on the convex hull of the data (by Theorem 2 of Cule et al. (2010)). The log-likelihood estimator $\log \widehat{f}_n$ is a piecewise linear function with knots which are a subset of $\{X_1, \ldots, X_n\}$. Walther (2002) and Rufibach (2007) provided algorithms for computing $\widehat{f}_n(X_i), i = 1, \ldots, n$. The entire log-density $\log \widehat{f}_n$ can be computed by linearly interpolating between between $\log \widehat{f}_n(X_{(i)})$ and $\log \widehat{f}_n(X_{(i+1)})$. Walther (2002) and Rufibach (2007) also pointed out that it is natural to apply weights for an EM-type algorithm. The $z_{1j}^{(t+1)}, \ldots, z_{nj}^{(t+1)}$ can be viewed as weights for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ when estimating the log-concave density ϕ_j in our algorithm for $j = 1, \ldots, K$. The algorithm stops once the increasing increment $\ell^{(t+1)} - \ell^{(t)}$ is below 10^{-7}, where $\ell^{(t)} = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j^{(t)} \exp\{\phi_j^{(t)}(x_i)\}$.

To avoid the local maximum, we restart the algorithm 20 times and choose the result with the highest log-likelihood. As we discussed in Section 1, the unboundedness issue of

the log-likelihood does happen infrequently, mostly due to an inappropriate initial. In our algorithm, we borrow the idea of restarting process in many existing EM-algorithms for parametric mixture models, e.g. Benaglia et al. (2009). If the log-likelihood goes to infinity in any iteration, our EM-type algorithm will be forced to restart from the beginning with a new randomly chosen initial.

3.5 Simulation Results

3.5.1 Copula procedure to generate multivariate log-concave mixtures

As we do not have a tuning issue for LCMLE, the most attractive application of LCMLE is the density estimation with dimensionality higher than 1. To generate data from a multivariate log-concave mixture model, we borrow the idea of the copula procedure from Chang and Walther (2007). For a d-dimensional log-concave mixture density, we observe n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^T \in \mathbb{R}^d$. To simplify our simulation, we focus on the model whose univariate marginal distributions are log-concave. We model the dependence structure with a normal copula. Suppose $(N_1, \ldots, N_d)^T$ be multivariate normal with mean $\mathbf{0}_d$ and covariance matrix Σ. Let $F_1, \ldots, F_d$ be the CDFs of the desired univariate log-concave distributions. Then,

$$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^T = (F_1^{-1}(\Phi(N_1)), \ldots, F_d^{-1}(\Phi(N_d)))^T.$$

3.5.2 Significant Improvement when densities are log-concave mixtures

We first generate 500 observations from a univariate log-concave mixture model: $0.3Logistic(0,1) + 0.7Laplace(5,1)$ (referred as Model I). This setup is a more general form of Chang and Walther (2007), as Chang and Walther (2007) only considered the case that one component is a location shift of the other. For the multivariate cases, we generate 500 observations based on the copula procedure as we discussed in Section 3.5.1 for Model II through IV, which are multivariate log-concave mixture models with dimensionality d from 2 to 4. For each model, component 1 (with probability 0.3) is generated as a joint normal distribution $N(\mathbf{0}_d, \mathbf{I}_d)$; component 2 (with probability 0.7) is generated through a normal copula $N(\mathbf{0}_d, 0.5\mathbf{I}_d + 0.5\mathbf{1}_d)$, where $\mathbf{I}_d$ is a $d \times d$ identity matrix and $\mathbf{1}_d$ is a $d \times d$ matrix of ones. The marginal distributions of component 2 are summarized in Table 3.1.

Table 3.1: The simulation setups of Model II - IV.

Model	d	Marginal Distribution of Component 2
II	2	$N(0,1)$, and $Gamma(2,1) + 2$
III	3	$N(0,1)$, $Gamma(2,1) + 1$, and $Beta(4,1)$
IV	4	$N(0,1)$, $Gamma(2,1) + 2$, $Beta(4,1)$, $Laplace(0,1) + 1$

We repeat the simulation 100 times for each model. When evaluating the simulation results for mixture models, there is a well-known label switching issue when sorting the labels for mixture models (Stephens (2000); Yao and Lindsay (2009)). In this paper,we adopt the method of Yao (2015) to find labels by minimizing the distance between the

estimated classification probabilities and the true labels over different permutations. After sorting the labels, we compute the mean square errors obtained by the log-concave EM algorithm (MSE_2) and compare them with the parametric normal EM-algorithm (MSE_1). As mixture models also serve as methods of classification, we compute the average misclassification number (denoted as AMN_2 for the log-concave EM-algorithm and AMN_1 for the normal EM-algorithm) among the 100 replicates. We are also interested in the difference between two classification methods. One of many possible measurements to summarize the similarity between two clusterings is the Adjusted Rand Index (ARI), which ranges from -1 to 1, see Hubert and Arabie (1985) for detailed description of ARI. In this paper, we compute the average Adjusted Rand Index ($AvARI$) among the 100 replicates.

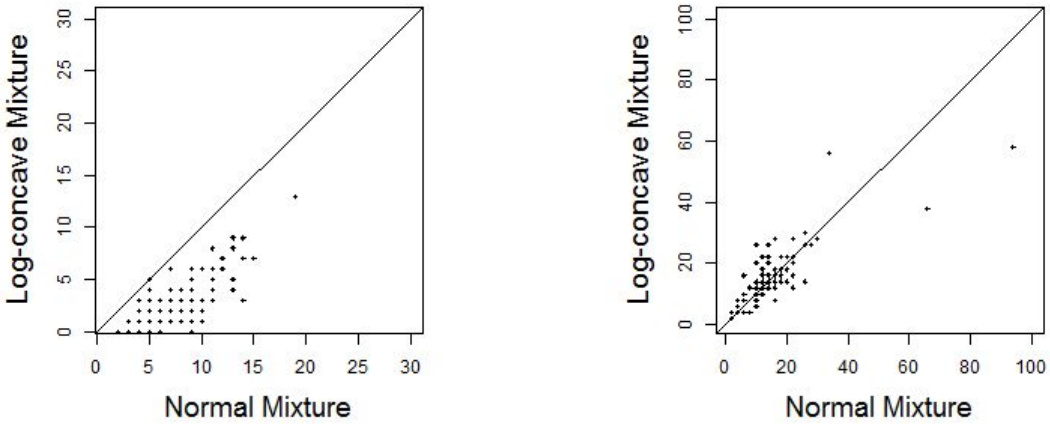
We report results over the 100 replicates in Table 3.2. We observe significant smaller MSEs for the estimated λ obtained by log-concave mixture models. Especially for Model I and II, the mean square errors obtained by log-concave mixture model are less than half of those obtained by normal mixture model. In terms of classification, the average misclassification number among the 500 observations are significantly reduced as well.

Table 3.2: Simulation results of Model I - IV.

Model	d	AMN_1	AMN_2	MSE_1	MSE_2	$AvARI$
I	1	17.56	**10.86**	0.0016	**0.0007**	0.91
II	2	30.35	**13.28**	0.0085	**0.0013**	0.86
III	3	4.76	**2.98**	0.0019	**0.0016**	0.95
IV	4	7.97	**3.21**	0.0006	**0.0004**	0.95

To compare the classification result for each replicate, we take $d = 4$ as an example

and show the clustering results in Figure 3.1. In Figure (3.1), each point represents a single replicate from Model IV's setup. The x-axis represents the number of misclassification by the Normal mixture EM-algorithm. The y-axis represents the number of misclassification by our log-concave mixture EM-algorithm. We observe significant improvement in the misclassification rates, as all the points for our 100 replicates are below the identity line. For the classification plot of Model I-III, please refer to the appendix.



(a) Model IV

(b) Model VIII

Figure 3.1: Four-dimensional clustering result: normal mixture EM-algorithm vs log-concave mixture EM-algorithm by number of misclassifications. The solid lines represent the identity.

To better illustrate the finite sample performance of the LCMLE, we pick one replicate from Model I. To compare the fitted densities with the true densities, in Figure 3.2, we plot the true component densities in the solid lines and the fitted densities in the dashed line. Even for a finite sample size of 500, the LCMLE for the log-concave mixture model

approximates the true component densities well.

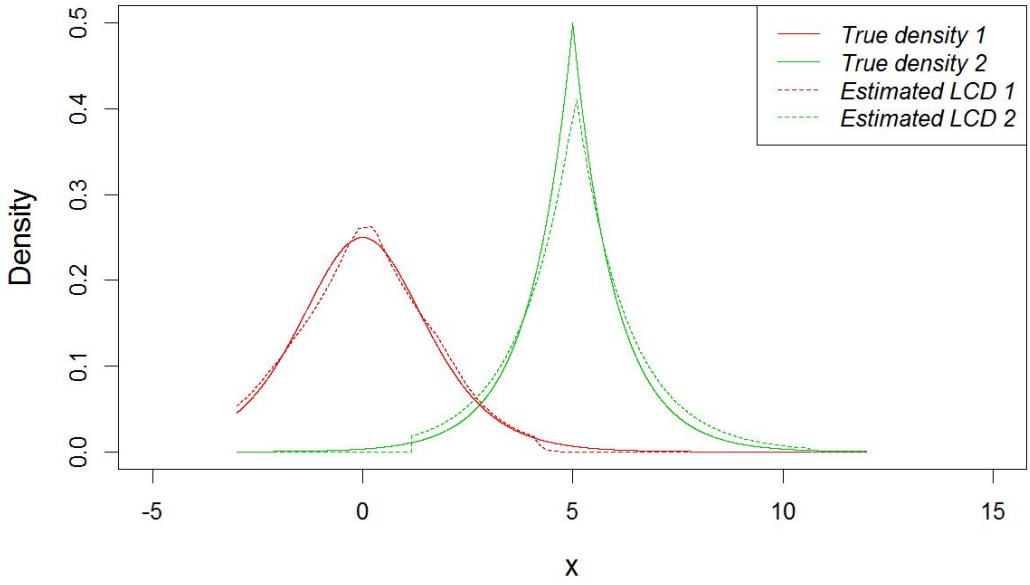


Figure 3.2: EM-type algorithm estimation for log-concave mixtures for a single replicate of Model I. Solid line represents the truth and dashed line represents the estimation results. The fitted $\hat{\lambda} = 0.3076$.

3.5.3 Insignificant penalty when the parametric assumptions are correct

We are also interested in the price that we have to pay for the flexibility while the data actually are from normal mixtures. For Model V - VIII, we generate $n = 500$ observations from a normal/joint normal mixture distribution, in which the first component (with probability 0.4) is a d-dimensional normal distribution with mean $\mathbf{0}_d$ and covariance

matrix $0.5\mathbf{I}_d + 0.5\mathbf{1}_d$, and the second component (with probability 0.6) is a d-dimensional normal distribution with mean $\boldsymbol{\mu}_d$ and the same covariance matrix, where $\mu_1 = 5$, $\boldsymbol{\mu}_2^T = (3, 2)$, $\boldsymbol{\mu}_3^T = (3, 2, 2)$, and $\boldsymbol{\mu}_4^T = (3, 3, 3, 3)$. We also repeat the simulation 100 times and compare the same criteria.

From Table 3.3, we observe no significant penalty for applying log-concave mixture models instead of normal mixture models. The MSEs and average misclassification numbers for log-concave mixture models are either almost the same or only a little bit higher than those for the multivariate normal mixture model. This phenomena is further supported in Figure (3.1), which demonstrate the classification results for Model VIII ($d = 4$). We observe no significant difference in terms of misclassifications, as most points in Figure (3.1) are around the identity line. Consequently, we conclude that the log-concave mixture model is a more flexible methodology without significant penalties. When the data are actually from normal mixtures, the only trade-off is the computational time.

Table 3.3: Simulation results of Model V - VIII.

Model	d	AMN_1	AMN_2	MSE_1	MSE_2	$AvARI$
V	1	3.17	3.51	0.0004	0.0005	0.99
VI	2	33.95	37.12	0.0018	0.0020	0.91
VII	3	21.95	23.57	0.0008	0.0008	0.93
VIII	4	19.85	20.37	0.0016	0.0018	0.94

3.6 Real Data Application

To further illustrate the performance of log-concave mixture models, we apply the log-concave EM algorithm to the crab data set of Campbell and Mahon (1974), which contains two types of crabs in the data set: 100 blue crabs and 100 orange crabs. We focus on these blue crabs, which include $n_1 = 50$ males and $n_2 = 50$ females referred to as groups G1 and G2, respectively. For each crab, there are five measurements. We are only interested in two of them: RW (rear width) and BD (body depth), both in unit of mm. In Figure 3.3, we give the scatter plot of RW and BD.

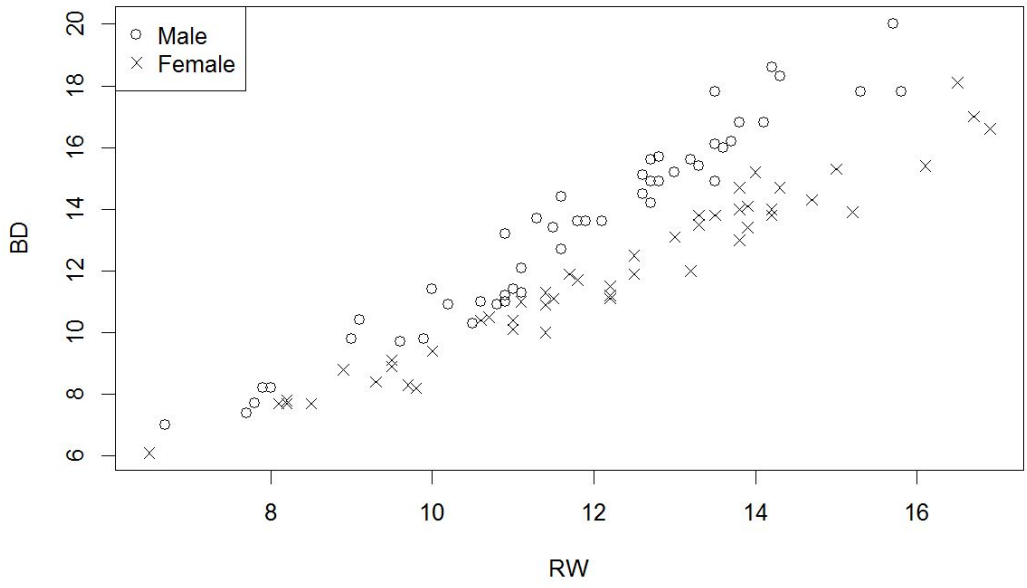


Figure 3.3: Scatter plot of RW (rear width) and BD (body depth) for the Blue Crab data set.

Fitting a 2-dimensional two component log-concave mixture model results in 18 ob-

servations misclassified. Fifteen observations from G_1 are misclassified into G_2 and three observations from G_2 are misclassified into G_1. The normal mixture model results in 20 observations misclassified in total. Two additional observations from G_1 are misclassified into G_2.

3.7 Conclusion

The log-concave maximum likelihood estimator (LCMLE) provides more flexibility to estimate mixture densities, when compared to the traditional parametric mixture models. The estimation of LCMLE for log-concave mixtures can be achieved by an EM-type algorithm. The LCMLE is not sensitive to the model mis-specification and consequently, only one implementation of the EM-type algorithm is necessary. Through simulation studies, we observed significant improvements in the sense of classification and no significant penalties when the parametric assumption is indeed correct.

In this paper, we proved the existence of the LCMLE for log-concave mixture models. The consistency is also proved for the estimated mixture density. If the true mixture density is identifiable, then the estimated component densities are also identifiable. However, it is not an easy task to prove the overall identifiability for the most general family of mixtures of log-concave distributions in (3.1) from a nonparametric point of view. Some restrictive conditions, such as symmetry, are needed to ensure identifiability. Hunter et al. (2007) and Bordes et al. (2006b) proved the identifiability of (1.1) if $K = 2$ and both component densities are symmetric but with different location parameters. Balabdaoui and Doss (2014) have considered a special case of (3.1), when $\phi_j(x; \theta_j) = \phi(x - \theta_j)$ and ϕ is a symmetric concave function about 0, and the identifiability of (3.1) follows from Hunter et al. (2007) and Bordes et al. (2006b) when $K = 2$.

3.8 Appendix A: Lemmas

Lemma 3.6 is taken from Cule and Samworth (2010). Lemma 3.7 to Lemma 3.10 are taken from DSS 2011. Lemma 3.11 is the extension of Lemma 2.13 of DSS 2011.

Lemma 3.6. *For any log-concave distribution Q with density f, there exist finite constants $B_1 = B_1(Q) > 0$ and $B_2 = B_2(Q) > 0$ such that $f(x) \leq B_1 \exp(-B_2||x||)$ for all $x \in \mathbb{R}^d$.*

Lemma 3.7. *The following properties of Q are equivalent:*

(a) csupp(Q) has non-empty interior.

(b) $Q(H) < 1$ for any hyperplane $H \subset \mathbb{R}^d$.

(c) With Leb denoting Lebesgue measure on $\mathbb{R}^d$,

$$\limsup_{\delta \downarrow 0} \{Q(A) : A \subset \mathbb{R}^d \text{ closed and convex,} Leb(A) \leq \delta\} < 1.$$

Lemma 3.8. *Let ϕ be the function such that for any $x, y \in interior(dom(\phi))$ and $t \in (0,1)$, if $tx + (1-t)y \in interior(dom(\phi))$, $\phi(tx + (1-t)y) \geq t\phi(x) + (1-t)\phi(y)$ and for $C \subseteq \mathbb{R}^d$, $\int_C e^{\phi(x)} dx \leq 1$. We define $D_q = \{x \in C : \phi(x) \geq q\}$. For any $r < M \leq \max_{x \in \mathbb{R}^d} \phi(x)$,*

$$Leb(D_r) \leq (M-r)^d e^{-M} / \int_0^{M-r} t^d e^{-t} dt.$$

Lemma 3.9. *Let $\overline{\phi}, \phi_1, \phi_2, \ldots$ be concave functions and $\phi_n \leq \overline{\phi}$. Further we assume the set $H = \{x : \liminf_{n \to \infty} \phi_n(x) > -\infty\}$ is not empty. Then there exist a subsequence*

$(\phi_{n(k)})_k$ of $(\phi_n)_n$ and a function ϕ such that $H \subset dom(\phi) \stackrel{d}{=} \{\phi > -\infty\}$:

$$\lim_{k\to\infty, x\to y} \phi_{n(k)}(x) = \phi(y) \text{ for all } y \in interior(dom(\phi)),$$
$$\lim_{k\to\infty, x\to y} \phi_{n(k)}(x) \leq \phi(y) \text{ for all } y \in \mathbb{R}^d.$$

Lemma 3.10. *Suppose Q_n is a sequence converged to some distribution Q and h be a nonnegative and continuous function, then*

$$\liminf_{n\to\infty} \int h dQ_n \geq \int h dQ.$$

If the stronger statement $\liminf_{n\to\infty} \int h dQ_n = \int h dQ < \infty$ holds, then for any function f such that $|f|/(1+h)$ is bounded,

$$\lim_{n\to\infty} \int f dQ_n = \int f dQ.$$

Lemma 3.11. *A point $x \in \mathbb{R}^d$ is an interior point of C if and only if*

$$h(Q, x) = \sup\{Q(E) : E \subset C, E \text{ closed and convex}, x \notin interior(E)\}/Q(C) < 1.$$

Proof. For $x \notin interior(E)$ and closed and convex E, there exists a unit vector $u_j \in R^d$ such that E is contained in the closed set H_C which is a subset of C:

$$C \supseteq H_C(x) = \{y \in C : u^T y \leq u^T x\} \supseteq E.$$

By the definition of $h(Q, x)$ we conclude $h(Q, x) \leq Q(H_C)/Q(C) \leq 1$. There are two cases: $E \subset H_C$ and $E = H_C(x)$. For the case $E \subset H_C$, by definition $h(Q, x) < 1$ strictly.

For the case $E = H_C(x)$, as we have $x \notin interior(E)$ but $x \in H_C(x)$, we conclude $x \in \partial H_C(x)$. Now if $x \notin interior(C)$, by definition, $h(Q, x) = 1$. On the other hand, if $h(Q, x) = 1$, then $Q(H_C(x)) = Q(C)$, which leads to $C = H_C(x) = E$. Combined with $x \notin interior(H_C(x))$ we can conclude that $x \notin interior(C)$. Consequently, $x \notin interior(C) \iff h(Q, x) = 1$. Thus, $x \in interior(C) \iff h(Q, x) < 1$. □

3.9 Appendix B: Proof of Theorem 3.1

The first thing is to prove the finiteness of the log-likelihood type function.

$L(Q)$ is the supreme of $L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q)$ over all $\boldsymbol{\phi} \in \boldsymbol{\Phi}, \boldsymbol{\lambda} \in \boldsymbol{\Lambda}, \boldsymbol{\lambda} \in \mathbb{R}^K$. If we take a special case that $\phi_j^*(x) = -(\log \lambda_j^*) - ||x||$, $L(\boldsymbol{\phi^*}, \boldsymbol{\lambda}^*, \boldsymbol{\pi}, Q) = \log K - \int ||x|| dQ > -\infty$. Consequently, $L(Q) > -\infty$.

Now we show $L(Q) < \infty$. As discussed at the end of Section 3.2, we do restrict our interest to the $\boldsymbol{\phi}$ such that $\int e^{\phi_j(x)} dx = 1$ for $j = 1, \ldots, K$. Consequently, we define the log-density as $l(x) = \log \sum_{j=1}^K \lambda_j e^{\phi_j(x)}$ and rewrite the log-likelihood-type function as $L(l, Q) = L(\boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{\pi}, Q)$. For the convenience of the proof, we define an envelope function $\overline{\phi}(x) = \max_j \{\phi_j(x)\}$, i.e. $\overline{\phi}(x) \geq l(x)$ for every $x \in \mathbb{R}^d$. This function is continuous but not smooth on $d-1$ dimensional boundaries. These boundaries divide the $csupp(Q)$ into K sets: $C_1, \ldots, C_K$. Each set C_j is defined as $C_j = \{x \in \mathbb{R}^d : \overline{\phi}(x) = \phi_j(x)\}$. The sets $C_1, \ldots, C_K$ are disjoint except on the boundaries and $Leb(C_i \cap C_j) = 0$ for every $i \neq j$. For any $x, y \in C_j$ and $t \in (0, 1)$, $\overline{\phi}(tx + (1-t)y) \geq t\overline{\phi}(x) + (1-t)\overline{\phi}(y)$ and $\int_{C_j} e^{\overline{\phi}(x)} dx \leq 1$. We define $M_j(\boldsymbol{\phi})$ and $\mathcal{S}(\boldsymbol{\phi})$ as stated in Section 3.1. As $L(l, Q) \leq \sum_{j=1}^K Q(C_j) M_j$, $M_j > -\infty$, and the restriction $|\mathcal{S}(\boldsymbol{\phi})| \geq \eta > 0$, we focus our interest on $M_j > 0$ and the only case which we have to worry about is all M_j's increasing to infinity. We define $D_q = \{x \in \mathbb{R}^d : \overline{\phi}(x) \geq q\}$. For any $c > 0$,

$$
\begin{aligned}
L(l,Q) &\le \int \overline{\phi}(x)dQ = \int_{csupp(Q)} \overline{\phi}(x)dQ \\
&= \int_{csupp(Q)\setminus D_{-cM_{(1)}}} \overline{\phi}(x)dQ + \int_{D_{-cM_{(1)}}} \overline{\phi}(x)dQ \\
&\le -cM_{(1)}(1-Q(D_{-cM_{(1)}})) + M_{(K)}Q(D_{-cM_{(1)}}) \\
&= -cM_{(k)}\mathcal{S}(\boldsymbol{\phi})(1-Q(D_{-cM_{(1)}})) + M_{(K)}Q(D_{-cM_{(1)}}) \\
&= \{(1+c\mathcal{S}(\boldsymbol{\phi}))Q(D_{-cM_{(1)}}) + c\mathcal{S}(\boldsymbol{\phi})\}M_{(K)} \\
&\le (1+c\eta)\Big(Q(D_{-cM_{(1)}}) - \frac{c\eta}{c\eta+1}\Big)M_{(K)}.
\end{aligned}
$$

We can always find sufficient large c such that the set $D_{-cM_{(1)}}$ is a closed and convex subset of $\mathbb{R}^d$. We define the set $D_{j,q} = \{x \in C_j : \overline{\phi}(x) \ge q\} \subset C_j$. Obviously $Leb(D_{-cM_{(1)}}) = \sum_{j=1}^{K} Leb(D_{j,-cM_{(1)}})$. For any $c > 0$, applying Lemma 3.8 to set $D_{j,-cM_{(1)}}$ and letting $M = M_{(1)}$ yield $Leb(D_{j,-cM_{(1)}}) \le (1+c)M_{(1)}^d e^{-M_{(1)}}/(d! + o(1)) \to 0$ as $M_{(1)} \to \infty$ for every $j = 1, \ldots, K$. Consequently, $Leb(D_{-cM_{(1)}}) \to 0$ as $M_{(1)} \to \infty$. By our definition, $\eta \in (0, 1]$. Thus, by Lemma 3.7, we can find sufficiently large c and small δ such that

$$
\sup\{Q(D) : D \subset \mathbb{R}^d,\ Leb(D) \le \delta\} < \frac{c\eta}{c\eta+1}.
$$

Thus, $L(l,Q) \to -\infty$ as $M_{(1)} \to \infty$, which indicates that when all modes of log-concave densities increase to infinity, the log-likelihood is poorly characterized. On the other hand, $L(l,Q) \le M_{(K)}$. These considerations show that $L(Q)$ is finite and equals the supremum of $L(l,Q)$ for suitable finite M_j's such that $M_j \in [M_{*j}, M_j^*]$ $(j = 1, \ldots, K)$.

Let $\phi_{m,j}$'s and $\lambda_{m,j}$'s form a sequence $l_m(x) = \log \sum \lambda_{m,j} \exp\{\phi_{m,j}(x)\}$ such that $-\infty < L(l_m, Q) \uparrow L(Q)$ as $m \to \infty$. Next, we will prove that for every $j \in \{1, \ldots, K\}$,

there exists a point, say, $x_{0,j} \in interior(csupp(Q))$, such that $\liminf_{m\to\infty} \phi_{m,j}(x_{0,j}) > -\infty$.

We define $\overline{\phi}_m(x) = \max_j\{\phi_{m,j}(x)\}$, $C_{m,j} = \{x \in \mathbb{R}^d : \overline{\phi}_m(x) = \phi_{m,j}(x)\}$, and $M_{m,j} = \max_{x\in\mathbb{R}^d} \phi_{m,j}(x)$. For any $j^* \in \{1,\ldots,K\}$, by picking any $x_{0,j^*} \in C_{m,j^*}$ such that $\phi_{m,j^*}(x_{0,j^*}) \in [M'_{m,j^*}, M_{m,j^*})$, where $M'_{m,j^*} = \max_{x\in\partial\{C_{m,j^*}\}} \phi_{m,j^*}(x)$, there exists a sufficient small $\epsilon \geq 0$ such that the set $E_{m,j^*} = \{x \in C_{m,j^*} : \phi_{m,j^*}(x) \geq \phi_{m,j^*}(x_{0,j^*}) + \epsilon\}$ is a closed and convex subset of C_{m,j^*} and x_{0,j^*} is not an interior point of E_{m,j^*}. Thus,

$$
\begin{aligned}
L(l_m, Q) = \int l_m dQ &\leq \int \overline{\phi}_m(x) dQ \\
&= \sum\nolimits_{j\neq j^*} \int_{C_{m,j}} \phi_{m,j}(x) dQ + \int_{C_{m,j^*}} \phi_{m,j^*} dQ \\
&= \sum\nolimits_{j\neq j^*} \int_{C_{m,j}} \phi_{m,j}(x) dQ + \int_{C_{m,j^*}\setminus E_{m,j^*}} \phi_{m,j^*} dQ + \int_{E_{m,j^*}} \phi_{m,j^*} dQ \\
&\leq \sum\nolimits_{j\neq j^*} M_{m,j} Q(C_{m,j}) + \phi_{m,j^*}(x_{0,j^*})(Q(C_{m,j^*}) - Q(E_{m,j^*})) + M_{m,j^*}(x_{0,j^*}) Q(E_{m,j^*}) \\
&= \sum\nolimits_{j\neq j^*} M_{m,j} Q(C_{m,j}) + \phi_{m,j^*}(x_{0,j^*}) Q(C_{m,j^*}) + (M_{m,j^*} - \phi_{m,j^*}(x_{0,j^*})) Q(E_{m,j^*}) \\
&\leq \sum\nolimits_{j=1}^{K} \max(M_{m,j}, 0) + \phi_{m,j^*}(x_{0,j^*}) Q(C_{m,j^*})(1 - h_{j^*}(Q, x_{0,j^*})).
\end{aligned}
$$

These inequalities hold for the case of $\phi_{m,j^*}(x_{0,j^*}) = M_{m,j^*}$ as well ($\epsilon = 0$ accordingly). By Lemma 3.11, $h_{j^*}(Q, x_{0,j^*}) < 1$. Due to the fact that M_{m,j^*} is finite, $interior(C_{m,j^*})$ is not empty. Consequently, $\liminf_{m\to\infty} Q(C_{m,j^*}) > 0$, which yields

$$
\begin{aligned}
\phi_{m,j^*}(x_{0,j^*}) &\geq -\frac{\sum_{j=1}^{K} \max(M_{m,j}, 0) - L(l_m, Q)}{Q(C_{m,j^*})(1 - h_{j^*}(Q, x_{0,j^*}))} \\
&> -\frac{\sum_{j=1}^{K} \max(M^*_j, 0) - L(l_1, Q)}{Q(C_{m,j^*})(1 - h_{j^*}(Q, x_{0,j^*}))} > -\infty.
\end{aligned}
$$

Hence, the set $H_j = \{x : \liminf_{m\to\infty} \phi_{m,j}(x) > -\infty\}$ is not empty for every $j \in \{1, \ldots, K\}$. From Lemma 3.6 we conclude that for each ϕ_j, we can find suitable finite positive constants $a_j, b_j > 0$ such that $\phi_j(x) \leq a_j - b_j||x|| \leq a - b||x||$, where $a = \max_j a_j > 0$ and $b = \min_j b_j > 0$. Then by Lemma 3.9, there exist a subsequence $(\phi_{1,m(k_1)})_{k_1}$ of $(\phi_{1,m})_m$ and a concave function ϕ_1 such that:

$$\lim_{k_1\to\infty, x\to y} \phi_{1,m(k_1)}(x) = \phi_1(y) \text{ for all } y \in interior(dom(\phi_1)),$$

$$\lim_{k_1\to\infty, x\to y} \phi_{1,m(k_1)}(x) \leq \phi_1(y) \text{ for all } y \in \mathbb{R}^d.$$

If we define $\phi_1 = -\infty$ on $\mathbb{R}^d \setminus dom(\phi_1)$, then we can rewrite them as:

$$\limsup_{k_1\to\infty} \phi_{1,m(k_1)}(x) \leq \phi_1(x) \quad \text{for all } x \in \partial\{dom(\phi_1)\},$$
$$\lim_{k_1\to\infty} \phi_{1,m(k_1)}(x) = \phi_1(x) \quad \text{for all } x \in \mathbb{R}^d \setminus \partial\{dom(\phi_1)\}.$$

We can find a sub-subsequence in the original subsequence, which has the similar property for $\phi_{2,m(k_2)}$. Keeping doing this sequentially for all $\phi_{m,j}$'s and $\lambda_{m,j}$'s will yield the common subsequence $l_{m(k)}$ and a function $l^*(x) = \log \sum \lambda_j \exp\{\phi_j(x)\}$ such that:

$$\limsup_{k\to\infty} l_{m(k)}(x) \leq l^*(x) \quad \text{for all } x \in \mathcal{P},$$
$$\lim_{k\to\infty} l_{m(k)}(x) = l^*(x) \quad \text{for all } x \in \mathbb{R}^d \setminus \mathcal{P},$$

where $\mathcal{P} = \cup_{j=1}^K \big(\partial\{dom(\phi_j)\}\big)$ and $Leb(\mathcal{P}) = 0$. The next step is to prove that $l^*(x)$ is the maximizer. Applying Fatou's lemma to the subsequence function $l_{m(k)}(x) \leq a - b||x||$

yields

$$\limsup_{k\to\infty} \int l_{m(k)} dQ \leq \int l^* dQ.$$

Hence,

$$L(Q) \geq l(l^*, Q) \geq \limsup_{k\to\infty} L(l_{m(k)}, Q) = L(Q),$$

from which we conclude $L(l^*, Q) = L(Q)$. The first inequality follows by the definition of $L(Q)$. The last equality follows by the definition that $l_{m(k)}$ is a sequence that maximizes $L(l_{m(k)}, Q)$ to $L(Q)$ as $k \to \infty$. Thus, it concludes the existence of the maximizer l^*, which indicates the existence of λ_j^*'s and ϕ_j^*'s.

3.10 Appendix C: Proof of Propositions

Proof of Proposition 3.2:

Suppose $\int ||x|| dQ = \infty$. By Lemma 3.6, $\phi_j \leq a_j - b_j||x||$ for all j's, where $a_j, b_j > 0$. Define $a = \max_j a_j > 0$ and $b = \min_j b_j > 0$, then $\phi_j \leq a - b||x||$ for all j's. Thus, $l \leq a - b||x||$, which entails that $L(Q) = -\infty$.

Now suppose $\int ||x|| dQ < \infty$ but $interior(csupp(Q)) = \emptyset$. By Lemma 3.7, it implies that $Q(H) = 1$ for some hyperplane $H \subset \mathbb{R}^d$. For $c \in \mathbb{R}$, define a function $\phi_{c,j} \in \Phi$ as $\phi_{c,j} = c - \log \lambda_j - ||x||$ when $x \in H$ and $\phi_{c,j} = -\infty$ for $x \notin H$. We further define $l_c = \log \sum \lambda_j \exp(\phi_{c,j})$. Then $L(l_c, Q) = c - \int ||x|| dQ + 1 \to \infty$ as $c \to \infty$.

Proof of Proposition 3.4:

Let $Q_t = (1-t)Q_0 + tQ_1$. Notice that, Q_t is concentrated on the same hyperplane if and only if Q_0 and Q_1 are concentrated on one same hyperplane. Consequently, if $L(Q_t) = \infty$, $L(Q_0), L(Q_1) = \infty$. It is equivalent to the statement of $L(Q_0), L(Q_1) < \infty$

implies that $L(Q_t) < \infty$. Let l_t be the maximizer corresponding to Q_t, then

$$L(Q_t) = L(l_t, Q_t) = (1-t)L(l_t, Q_0) + tL(l_t, Q_1) \leq (1-t)L(Q_0) + tL(Q_1).$$

The equality holds if and only if $l(\cdot|Q_t) = l(\cdot|Q_0) = l(\cdot|Q_1)$.

3.11 Appendix D: Proof of Theorem 3.5

We proof the theorem for a subsequence of Q_n. Let $L(Q_n) \to \Gamma$. As in the proof of Theorem 3.1, $l_n(x) \leq a - b||x||$ and $\inf \phi_{n,j}(x_0) > -\infty$ for some $x_0 \in interior(csupp(Q))$. Therefore, for a subsequence of $(Q_n)_n$, there exists a function l^* such that $l_n(y), l^*(y) \leq a - b||y||$, and

$$\limsup_{k\to\infty} l_{n(k)}(x) \leq l^*(x) \quad \text{for all } x \in \mathcal{P},$$
$$\lim_{k\to\infty} l_{n(k)}(x) = l^*(x) \quad \text{for all } x \in \mathbb{R}^d \setminus \mathcal{P}.$$

By Skorohod's theorem, there exists a probability space with random variables $X_n \sim Q_n$, $X \sim Q$ such that $X_n \to X$ almost surely. We define a random variable $H_n =$

$a - b||X_n|| - l_n(X_n) \geq 0$. Applying Fatou's lemma to H_n yields,

$$\begin{aligned}
\Gamma &= \lim_{n\to\infty} \int l_n dQ_n = \lim_{n\to\infty} \int (a - b||x||) dQ_n - E(H_n) \\
&= a - b\gamma - \liminf_{n\to\infty} E(H_n) \leq a - b\gamma - E\left(\liminf_{n\to\infty}(H_n)\right) \\
&\leq a - b\gamma - E\left(a - b||X|| - l^*(X)\right) \\
&= b\left(\int ||x|| dQ_0 - \gamma\right) + \int l^*(X) dQ_0 \\
&= L(l^*, Q_0) \leq L(Q_0).
\end{aligned}$$

Let $l_0(x) = \log \sum \lambda_j \phi_j(x)$, i.e. λ_j's and ϕ_j's are the results corresponding with l_0. In the following proof we utilize a special approximation scheme.

Let $l^{(\epsilon)}(x) = \log \sum \lambda_j^{(\epsilon)} \phi_j^{(\epsilon)}(x)$, $\lambda_j^{(\epsilon)} = \lambda_j$ and $\phi_j^{(\epsilon)} = \inf_{v,c}(v_j^T x + c_j)$ such that $||v_j|| \leq \epsilon^{-1}$ and $\phi_j(y) \leq v_j^T y + c_j$. DSS 2011 shows that the approximation $\phi_j^{(\epsilon)}$ is real valued and Lipschitz continuous with constant ϵ^{-1}. Consequently, $l^{(\epsilon)}(x)$ is also Lipschitz-continuous with constant ϵ^{-1}. Moreover, $\phi_j^{(\epsilon)} \geq \phi_j$ and $\phi_j^{(\epsilon)} \downarrow \phi_j$ pointwise as $\epsilon \downarrow 0$. Thus, $l^{(\epsilon)} \downarrow l_0$ pointwise as $\epsilon \downarrow 0$ and $l^{(1)} \geq l^{(\epsilon)} \geq l_0$ for $\epsilon \in (0, 1)$. With this approximation, it follows from Lipschitz-continuity, $\int ||x|| dQ_0 = \gamma < \infty$, and the stronger version of Lemma 3.10 that

$$\begin{aligned}
\Gamma &= \lim_{n\to\infty} \int l_n dQ_n \geq \lim_{n\to\infty} L(l^{(\epsilon)}, Q_n) \\
&= \lim_{n\to\infty} \int l^{(\epsilon)} dQ_n - \sum \pi_j \int e^{\phi_j^{(\epsilon)}(x)} dx + 1 \\
&= \int l^{(\epsilon)} dQ_0 - \sum \pi_j \int \exp(\phi_j^{(\epsilon)}(x)) dx + 1.
\end{aligned}$$

Applying monotone convergence theorem to function $l^{(1)} - l^{(\epsilon)}$ and dominated convergence theorem to $\exp\{\phi_j^{(\epsilon)}\}$'s yields, $\lim_{\epsilon\to 0^+} L(l^{(\epsilon)}, Q_0) = L(l_0, Q_0)$. Hence, $\Gamma \geq L(Q_0)$.

Combining with $\Gamma \leq L(l^*, Q_0) \leq L(Q_0)$ yields $\Gamma = L(Q_0) = L(l^*, Q_0)$, which indicates that l^* equals the maximizer $l_0 = l(\cdot|Q_0)$ that corresponds to $L(Q_0)$.

Applying to density $f_n = \exp\{l_n\}$ and $f_0 = \exp\{l_0\}$ yields,

$$\lim_{n\to\infty, x\to y} f_n(x) = f_0(y) \text{ for all } x \in \mathbb{R}^d \setminus \mathcal{P},$$
$$\lim_{n\to\infty, x\to y} f_n(x) \leq f_0(y) \text{ for all } y \in \mathcal{P},$$

where $\mathcal{P} = \cup_{j=1}^{K} \big(\partial\{f_{0j} > 0\}\big)$ and $Leb(\mathcal{P}) = 0$. Consequently, $(f_n)_n \to f_0$ almost everywhere with respect to Lebesgue measure. In addition, $|f_n(x)| \leq e^{a-b||x||}$, and $\int e^{a-b||x||}dx$ is finite. Applying Lebesgue's dominated convergence theorem yields,

$$\lim_{n\to\infty} \int |f_n(x) - f_0(x)|dx = 0.$$

Consequently, we claim Theorem 3.5 to be true for a subsequence of the original sequence $(Q_n)_n$. It remains to show it is true for the entire sequence.

Suppose any assertion about f_n is false, then one could replace the initial sequence $(Q_n)_n$ from the start with a subsequence such that one of the following three conditions is satisfied:

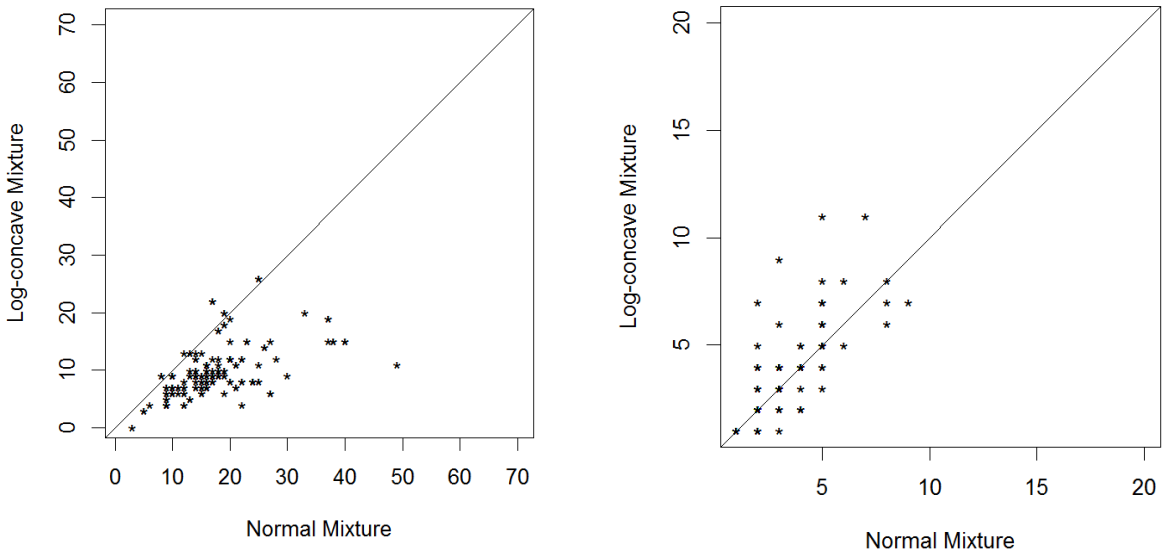
(i)$\lim_{n\to\infty} f_n(x_n) > f_0(y)$ for some sequence $(x_n)_n$ converge to point y;

(ii)$\lim_{n\to\infty} f_n(x_n) < f_0(y)$ for some sequence $(x_n)_n$ converge to point y;

(iii)$\lim_{n\to\infty} \int |f_n(x) - f_0(x)|dx > 0$.

Any of these three properties are transmitted to subsequence of $(Q_n)_n$, which would lead to a contradiction.

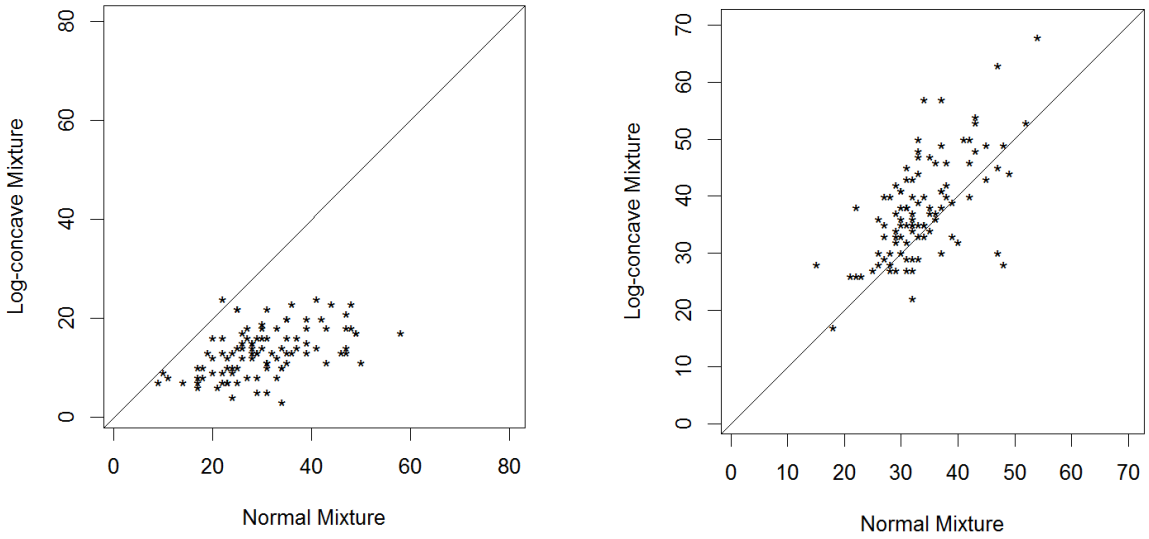
3.12 Appendix E: Classfication Plot of Model I-III and Model V-VII



(a) Model I

(b) Model V

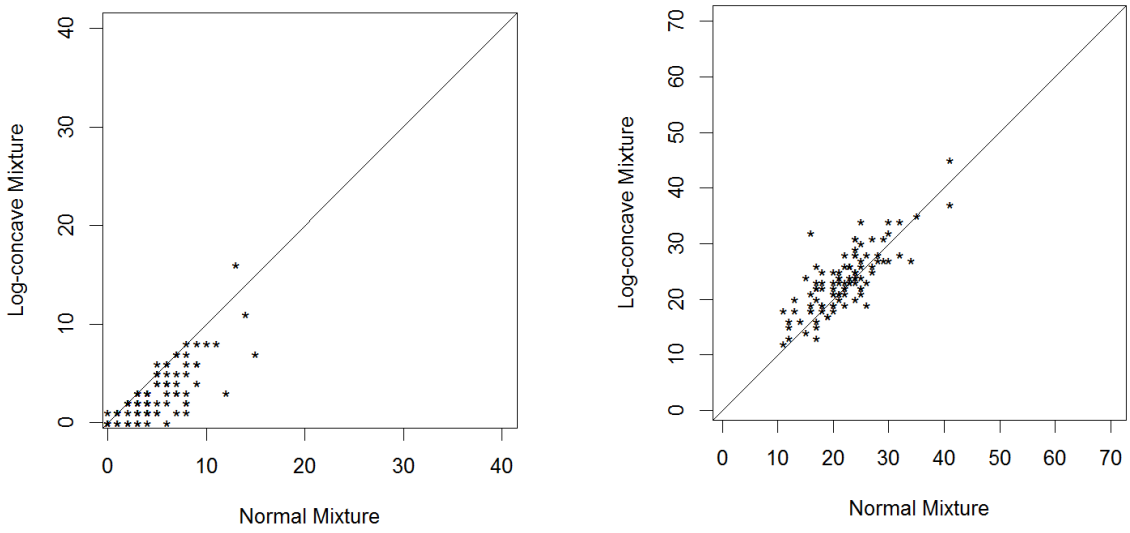
Figure 3.4: One-dimensional clustering result: normal mixture EM-algorithm vs log-concave mixture EM-algorithm by number of misclassifications. The solid lines represent the identity.



(a) Model II

(b) Model VI

Figure 3.5: Two-dimensional clustering result: normal mixture EM-algorithm vs log-concave mixture EM-algorithm by number of misclassifications. The solid lines represent the identity.



(a) Model III

(b) Model VII

Figure 3.6: Three-dimensional clustering result: normal mixture EM-algorithm vs log-concave mixture EM-algorithm by number of misclassifications. The solid lines represent the identity.

Chapter 4

Log-concave Mixtures of Regressions Models

4.1 Introduction

When a random variable has a finite mixture density that depends on certain covariates, we obtain a finite mixture of regression (FMR) model. Suppose we observe univariate response y_i and p-dimensional covariate $\mathbf{x}_i$, the mixture of linear regression can be written as follows:

$$f(y_i, \mathbf{x}_i; \boldsymbol{\psi}) = \sum_{j=1}^{K} \lambda_j g_j(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j), \tag{4.1}$$

where $\boldsymbol{\beta}_j \subseteq \mathbb{R}^p$, $\lambda_j \in (0,1)$, $\sum_{j=1}^{K} \lambda_j = 1$, $\boldsymbol{\psi} = (\lambda_1, \ldots, \lambda_{K-1}, \boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T \in \mathbb{R}^{Kp+K-1}$, and g_j is a parametric distribution function, such as normal, for j-th component, $j = 1, \ldots, K$.

The parametric FMR model (4.1) can be estimated through the maximum likeli-

hood estimators. As there's usually no explicit solutions to the unknown parameters, it is natural to reformulate the likelihood as an incomplete data problem and apply the expectation-maximization (EM) algorithm for the FMR models; see, e.g. Dempster et al. (1977) and McLachlan and Krishnan (2007). Besides estimating the parameters in the FMR models, the EM algorithms also provide the probabilities that an observation belongs to certain classes. Consequently, FMR models can also be considered as unsupervised classification methods, even though clustering might not always be the goal.

One major drawback of model (4.1) is the strong parametric assumption about the component density g_j. It is often too restrictive and the parameter estimation may be biased if the parametric model is misspecified. Another drawback is that each model requires a specific EM algorithm based on the parametric assumption. As a result, it would be valuable to have a universal EM algorithm for all, or at least some classes of the FMR models. Possible solutions include traditional nonparametric methods, e.g. Hunter and Young (2012) and Wu and Yao (2014), to adjust the parametric model misspecification. These traditional nonparametric methods, e.g. kernel methods, bring new difficulties in selecting the tuning parameters. Consequently, we borrow the idea of log-concave shape constraint and combine with the FMR models.

4.2 Mixtures of Regression Models with Log-concave Error Densities

In this chapter, we let Z be a latent variable with $\mathbb{P}(Z = j) = \lambda_j$, where $\lambda_j \in (0, 1)$, and $\sum_{j=1}^{K} \lambda_j = 1$ for $j = 1, \ldots, K$. While given the latent variable $Z = j$, the response

y_i has a linear relationship with $\mathbf{x}_i \in \mathbb{R}^p$:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta}_j + \epsilon_j, \tag{4.2}$$

where $\boldsymbol{\beta}_j = (\beta_{0,j}, \beta_{1,j}, \ldots, \beta_{p-1,j})^T \in \mathbb{R}^p$ and ϵ_j is the error term with the distribution function g_j $(j = 1, \ldots, K)$. We assume that each component's error distribution g_j is an unknown density function with the mean 0 for $j = 1, \ldots, K$. If we do not assume a zero mean for g_j, $\boldsymbol{\beta}_j$ does not contain the intercept term accordingly. To relax the traditional parametric assumption about g_j, we only assume that g_j's are log-concave, i.e. $\log g_j$ is concave for $j = 1, \ldots, K$. We define $\boldsymbol{\theta}_j = (\lambda_j, \boldsymbol{\beta}_j^T)^T$ for $j = 1, \ldots, K$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$. The likelihood function for the mixture of regressions model can be presented as:

$$f(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{g}) = \sum_{j=1}^{K} \lambda_j g_j(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j), \tag{4.3}$$

where $\boldsymbol{\theta} \in \Theta = \{\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T \mid \boldsymbol{\beta}_j \in \mathbb{R}^p, \lambda_j \in (0,1), \sum_{j=1}^K \lambda_j = 1\} \subset \mathbb{R}^{Kp+K-1}$.

Let $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,p-1})^T$ and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T \in \mathbb{R}^{n \times p}$ be the n observations for the mixture of regressions model, where $n \gg Kp + K - 1$. In order to estimate the model (4.3), it is natural to maximize the observed log-likelihood function:

$$\ell(\boldsymbol{\theta}, \mathbf{g}|\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} \log \sum_{j=1}^{k} \lambda_j g_j(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j), \tag{4.4}$$

where $g_j(x) = \exp\{\phi_j(x)\}$ for some unknown concave function $\phi_j(x)$.

4.3 The EM-type Algorithms for Log-concave FMR Models

We define the missing value $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T \in \mathbb{R}^{n \times K}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})^T$ $(i = 1, \ldots, n)$ is a K-dimensional indicator vector with its j-th element given by

$$z_{ij} = \begin{cases} 1 & \text{if } (\mathbf{x}_i, y_i) \text{ belongs to } j\text{-th group;} \\ 0 & \text{otherwise .} \end{cases}$$

Consequently, the complete log-likelihood for equation (4.4) is:

$$\ell_c(\boldsymbol{\theta}, \mathbf{g}|\mathbf{X}, \mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \{\log \lambda_j + \log g_j(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j)\}. \tag{4.5}$$

In the E-step, given the current estimate $\boldsymbol{\theta}^{(t)}$ and $g_j^{(t)}$'s, we need to compute

$$Q(\boldsymbol{\theta}, \mathbf{g}|\boldsymbol{\theta}^{(t)}, \mathbf{g}^{(t)}, \mathbf{X}, \mathbf{y}) = \mathbb{E}\{\ell_c(\boldsymbol{\theta}, \mathbf{g}|\mathbf{X}, \mathbf{y}, \mathbf{Z}) \mid \boldsymbol{\theta}^{(t)}, \mathbf{g}^{(t)}, \mathbf{X}, \mathbf{y}\},$$

which is equivalent to computing

$$\begin{aligned} z_{ij}^{(t+1)} &= E(Z_{ij}|\boldsymbol{\theta}^{(t)}, \mathbf{g}^{(t)}, \mathbf{X}, \mathbf{y}) = Pr(Z_{ij} = 1|\boldsymbol{\theta}^{(t)}, \mathbf{g}^{(t)}, \mathbf{X}, \mathbf{y}) \\ &= \frac{\lambda_j^{(t)} g_j^{(t)}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(t)})}{\sum_{h=1}^{K} \lambda_h^{(t)} g_h^{(t)}(y_i - \mathbf{x}_i^T \boldsymbol{\beta}_h^{(t)})}. \end{aligned} \tag{4.6}$$

In M-step, we need to maximize the following Q function:

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \mathbf{g}|\boldsymbol{\theta}^{(t)}, \mathbf{X}, \mathbf{y}) &= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(t+1)}\{\log\lambda_j + \log g_j(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j)\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(t+1)}\log\lambda_j + \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(t+1)}\log g_j(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j). \qquad (4.7)
\end{aligned}
$$

The first part of (4.7) is maximized by $\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^{(t+1)}, j = 1, \ldots, K$. However, for the second part, there is no explicit solution for $\boldsymbol{\beta}_j$'s and g_j's. Consequently, we propose to alternatively update g_j's and $\boldsymbol{\beta}_j$'s to maximize the second part of (4.7).

It is also well known that MLEs can be sensitive to outliers, see e.g. Yao et al. (2014) and García-Escudero et al. (2009). To overcome this problem, we further propose a robust technique, which adopts the idea of least trimmed squares (LTS), see e.g. Rousseeuw (1985) for a detailed description of LTS. For each algorithm, when updating λ_j's and $\boldsymbol{\beta}_j$'s in the t-th iteration ($j = 1, \ldots, K$), we drop s observations with the least log-likelihood values. In that way, we sacrifice some efficiency to gain the robustness to the outliers. The number s is the trimming tuning parameter, which satisfies $0 < s < n/2$. In this paper, we mainly use this trimmed idea to get a stable estimate of log-concave component error densities while enjoying its robustness when the component error densities are highly skewed or have heavy tails. Our empirical experience suggests that the choice of $s = \lfloor n/40 \rfloor$ works well. Note that a larger value of s would make our algorithms more robust if there are outliers in the data set and the sample size is not too small.

Our methodology is summarized as follows. First, we apply some stochastic search strategy, which will be addressed later, to create the initial value for normal mixtures of regression models (from function `regmixEM` in R package `mixtools`, see Benaglia et al. (2009)), until convergence. We treat the outcome of the normal mixture EM algorithm as

the starting values for our EM-type algorithms, i.e. $\boldsymbol{\psi}^{(0)} = (\hat{\lambda}_1^{(0)}, \ldots, \hat{\lambda}_K^{(0)}, \hat{\boldsymbol{\beta}}_1^{(0)T}, \ldots, \hat{\boldsymbol{\beta}}_K^{(0)T})^T$. The normal mixture of regressions model usually provides good initial values and our proposed EM algorithm will further improve the estimate if the error density is not normally distributed. The initial estimated density g_j can be obtained by the function `mlelcd` in R package `LogConcDEAD` (Cule et al., 2009).

First, we propose the Algorithm 4.1 for the case that all components have the same error density g.

Algorithm 4.1. *The EM-type algorithm when all g_j's are the same, i.e. $g_j \equiv g$.*

Initialize $\boldsymbol{\psi}^{(0)}$ and $z_{ij}^{(0)}$ from normal mixture EM algorithm with equal variances and initialize the trimmed index subset of size $n-s$, denoted by $I^{(0)}$, which has the $n-s$ largest log-likelihoods. Initialize $g^{(0)}$ by the function `mlelcd` through fitted residuals $y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(0)}$ with weights $z_{ij}^{(0)}$ for $i = 1, \ldots, n$, $j = 1, \ldots, K$.

In t-th iteration, it consists of the following steps.

E-step: *Given $\boldsymbol{\psi}^{(t)}$ and $g^{(t)}$, we calculate*

$$z_{ij}^{(t+1)} = E(Z_{ij}|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{\psi}^{(t)}, g^{(t)}) = \frac{\lambda_j^{(t)} g^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)})}{\sum_{h=1}^K \lambda_h^{(t)} g^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_h^{(t)})}, \tag{4.8}$$

for $i = 1, \ldots, n, \quad j = 1, \ldots, K$.

M-step:

(A) *Calculate the log-likelihood value for each observation:*

$$\ell_i^{(t)} = \ell(\boldsymbol{x}_i, y_i|g^{(t)}, \boldsymbol{\psi}^{(t)}) = \log \sum_{j=1}^K \lambda_j^{(t)} g^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)}),$$

from $i = 1, \ldots, n$. Update the trimmed index subset of size $n-s$, denoted by $I^{(t+1)}$, which has the $n-s$ largest log-likelihoods.

(B) *Update λ simply through*

$$\lambda_j^{(t+1)} = \frac{1}{n-s} \sum_{i \in I^{(t+1)}} z_{ij}^{(t+1)}, j = 1, \ldots, K. \tag{4.9}$$

(C) *Update $\boldsymbol{\beta}$:*

$$\tilde{\boldsymbol{\beta}}_j^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\beta}_j} \sum_{i \in I^{(t+1)}} z_{ij}^{(t+1)} \log g^{(t)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j), \quad j = 1, \ldots, K. \tag{4.10}$$

(D) *Shift the intercept of $\tilde{\boldsymbol{\beta}}_j^{(t+1)}$ so that the residuals have a zero mean.*

$$\hat{\boldsymbol{\beta}}_j^{(t+1)} = (\hat{\beta}_{j,0}^{(t+1)}, \tilde{\beta}_{j,1}^{(t+1)} \ldots, \tilde{\beta}_{j,p-1}^{(t+1)}),$$

where

$$\hat{\beta}_{j,0}^{(t+1)} = \tilde{\beta}_{j,0}^{(t+1)} + c_j^{(t+1)} \quad \textit{with} \quad c_j^{(t+1)} = \frac{1}{n-s} \sum_{i \in I^{(t+1)}} z_{ij}^{(t+1)} (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}_j^{(t+1)}),$$

for $j = 1, \ldots, K$.

(E) *Update g by:*

$$g^{(t+1)} \leftarrow \arg\max_{g} \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij}^{(t+1)} \log g(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j^{(t+1)}). \tag{4.11}$$

In (4.10), $\boldsymbol{\beta}_j$ is updated through the function `optim` *in R. The evaluation of $\log \widehat{g}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)})$ is calculated through the function* `dlcd` *in R package* `LogConcDEAD`*. In (4.11), the error density g is updated through the function called* `mlelcd` *in the R package* `LogConcDEAD` *through Kn fitted residuals $y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)}$ with weights $z_{ij}^{(t+1)}$, $i = 1, \ldots, n, j = 1, \ldots, K$. The algorithm is terminated if either t_{max} of iterations has been reached, or if*

$\ell^{(t+1)} - \ell^{(t)} < 10^{-8}$, *where* $\ell^{(t)} = \sum_{i \in I^{(t)}} \ell_{(i)}^{(t)}$ *is the trimmed log-likelihood.*

To avoid the local maximum, we follow the similar stochastic search strategy proposed by Dümbgen et al. (2013). We restart the entire algorithm 20 times. For each restart, we randomly sample $\lfloor \alpha n \rfloor$ ($\alpha \in (0,1)$) observations K times, fit K simple linear regressions, obtain the K groups of coefficients, and treat them as the starting values of $\boldsymbol{\beta}$'s in the normal EM algorithm for K components. In this paper, we set $\alpha = 0.10$, which usually works very well. Additionally, we generate λ_j's from a uniform(0,1) distribution, scale them so that their sum is one, and treat them as the starting values of the mixing proportions in the normal EM algorithm. We then fit a normal FMR model, obtain the estimated coefficients, and use them as the initial values for our algorithm. We repeat this procedure 20 times and select the solution with the highest trimmed likelihood to avoid getting stuck in a local maximum.

The LCMLE $\widehat{g}$ has been studied by Walther (2002) and Rufibach (2007). Here, we briefly summarize the results. Given i.i.d. data $X_1, \ldots, X_n$ which follow a distribution g, the Log-concave Maximum Likelihood Estimator (LCMLE) $\widehat{g}$ exists uniquely and has support on the convex hull of the dataset (by Theorem 2 of Cule et al. (2010)). In addition, $\log \widehat{g}$ is a piecewise linear function whose knots are a subset of $\{X_1, \ldots, X_n\}$. Walther (2002) and Rufibach (2007) provided algorithms for computing $\widehat{g}(X_i), i = 1, \ldots, n$. The entire log-density $\log \widehat{g}$ can be computed by linear interpolation between $\log \widehat{g}(X_{(i)})$ and $\log \widehat{g}(X_{(i+1)})$. Walther (2002) and Rufibach (2007) also pointed out that it is natural to apply weights in the density estimation step of the EM-type algorithms. The $z_{ij}^{(t+1)}$'s can be viewed as weights for the Kn fitted residuals $y_i - \mathbf{x}_i^T \boldsymbol{\beta}_j^{(t+1)}$ ($i = 1, \ldots, n, j = 1, \ldots, K$), while estimating the log-concave density g for M-step 3 in our algorithms.

A more general case is that the components' error terms do not share a common

distribution, i.e. at least one g_j is different. Consequently, we propose the following Algorithm 4.2. The main difference is that, in Algorithm 4.2, each component density g_j is estimated by the iterative residuals only from the according component class, instead of being estimated by the entire residuals from all components in Algorithm 4.1.

Algorithm 4.2. *The EM-type algorithm when g_j's are different.*

Initialize $\boldsymbol{\psi}^{(0)}$ and $z_{ij}^{(0)}$ from normal mixture EM algorithm with unequal variances and initialize the trimmed subset of size $n-s$, denoted by $I^{(0)}$, which has the $n-s$ largest log-likelihoods. For $j \in \{1,\ldots,K\}$, initialize $g_j^{(0)}$ by the function `mlelcd` *through fitted residuals $y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(0)}$ with weights $z_{ij}^{(0)}$ for $i = 1,\ldots,n$.*

In t-th iteration, it consists of the following steps.

E-step: *Given $\boldsymbol{\psi}^{(t)}$ and $g^{(t)}$, we calculate*

$$z_{ij}^{(t+1)} = E(Z_{ij}|\boldsymbol{X},\boldsymbol{y},\boldsymbol{\psi}^{(t)},g^{(t)}) = \frac{\lambda_j^{(t)} g_j^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)})}{\sum_{h=1}^{K}\lambda_h^{(t)} g_h^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_h^{(t)})}, \tag{4.12}$$

for $i = 1,\ldots,n, \quad j = 1,\ldots,K$.

M-step:

(A) *Calculate the log-likelihood value for each observation:*

$$\ell_i^{(t)} = \ell(\boldsymbol{x}_i, y_i|\boldsymbol{g}^{(t)},\boldsymbol{\psi}^{(t)}) = \log\sum_{j=1}^{K}\lambda_j^{(t)} g_j^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)}),$$

for $i = 1,\ldots,n$. Update the trimmed subset of size $n-s$, denoted by $I^{(t+1)}$, which has the $n-s$ largest log-likelihoods.

(B) *Update λ simply through*

$$\lambda_j^{(t+1)} = \frac{1}{n-s}\sum_{i\in I^{(t+1)}} z_{ij}^{(t+1)}, j = 1,\ldots,K. \tag{4.13}$$

(C) *Update* $\boldsymbol{\beta}$*:*

$$\tilde{\boldsymbol{\beta}}_j^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\beta}_j} \sum_{i \in I^{(t+1)}} z_{ij}^{(t+1)} \log g_j^{(t)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j), \quad j = 1, \ldots, k. \tag{4.14}$$

(D) *Shift the intercept of* $\tilde{\boldsymbol{\beta}}_j^{(t+1)}$ *so that the residuals have a zero mean.*

$$\hat{\boldsymbol{\beta}}_j^{(t+1)} = (\hat{\beta}_{j,0}^{(t+1)}, \tilde{\beta}_{j,1}^{(t+1)} \ldots, \tilde{\beta}_{j,p-1}^{(t+1)}),$$

where

$$\hat{\beta}_{j,0}^{(t+1)} = \tilde{\beta}_{j,0}^{(t+1)} + c_j^{(t+1)} \quad \textit{with} \quad c_j^{(t+1)} = \frac{1}{n-s} \sum_{i \in I^{(t+1)}} z_{ij}^{(t+1)} (y_i - \boldsymbol{x}_i^T \tilde{\boldsymbol{\beta}}_j^{(t+1)}),$$

for $j = 1, \ldots, K$*.*

(E) *Update* g_j *by:*

$$g_j^{(t+1)} \leftarrow \arg\max_{g_j} \sum_{i=1}^{n} z_{ij}^{(t+1)} \log g_j(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_j^{(t+1)}). \tag{4.15}$$

for $j = 1, \ldots, K$*.*

In (4.15), the j*-th component density* g_j *is updated through the function called* `mlelcd` *in the R package* `LogConcDEAD` *through* n *fitted residuals* $y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)}$ *with weights* $z_{ij}^{(t+1)}$*,* $i = 1, \ldots, n$ *for* $j \in \{1, \ldots, K\}$*. The algorithm is terminated if either the maximum number of iterations* t_{max} *has been reached, or if* $\ell^{(t+1)} - \ell^{(t)} < 10^{-8}$*, where* $\ell^{(t)} = \sum_{i \in I^{(t)}} \ell_i^{(t)}$ *is the trimmed log-likelihood for* t*-th iteration.*

4.4 Numerical Experiments

In this section, we study the performance of our EM-type algorithms and compare them with the according EM algorithms for the normal FMR models. For the convenience purposes, in the following text and tables, we denote Algorithm 4.1 as "LCD-EM1" and compare it with the normal EM algorithm with equal variance and similar trimming techniques, denoted as "Normal-EM1". We also denote Algorithm 4.2 as "LCD-EM2" and compare it with the normal EM algorithm with unequal variance and similar trimming techniques, denoted as "Normal-EM2". For all trimming techniques, we set the trimming constant $s = \lfloor n/40 \rfloor$ for all algorithms, which usually works well based on our empirical experience.

To evaluate the performance of Algorithm 4.1, we generate data from 2-component log-concave FMR models, which share the same error density family among components:

$$y_i = \begin{cases} \beta_{0,1} + \beta_{1,1} x_i + e_i & \text{with probability} \quad \lambda; \\ \beta_{0,2} + \beta_{1,2} x_i + e_i & \text{with probability} \quad 1 - \lambda. \end{cases} \tag{4.16}$$

where x_i's are independently generated from Uniform$(-1, 3)$. We set $\lambda = 0.3$, $\boldsymbol{\beta}_1 = (\beta_{0,1}, \beta_{1,1})^T = (0, 2)^T$, and $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2})^T = (-2, 5)^T$. For Model I through Model VI, e_i's are independently and identically generated based on the parametric form from Table 4.1. For all six models, we generate data for a finite sample size of $n = 400$.

We take Model III as an example. Figure 4.1(a) shows the scatter plot of the generated data for the log-concave FMR model based on the setup of Model III for one single replicate. We fit this generated data with Algorithm 4.1 and show the property of monotone increasing log-likelihood in Figure 4.1(b). To further illustrate the performance of LCMLE for a single replicate, Figure 4.1(c) shows that for this replicate, the fitted

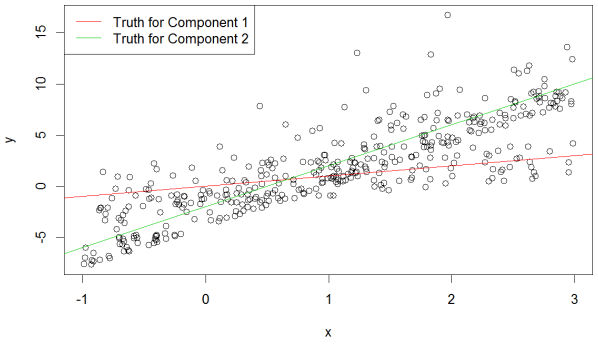
Table 4.1: The error densities for Model I to Model VI and the summary of the according features.

	e_i's distribution	log-concave	symmetric
Model I	Standard Normal: $N(0,1)$	Yes	Yes
Model II	Centered Beta: $3(Beta(1,2)-1/3)$	Yes	No
Model III	Centered Exponential: $Exp(2)-2$	Yes	No
Model IV	Standard Laplace: $Laplace(0,1)$	Yes	Yes
Model V	Centered Beta: $4(Beta(0.25,0.75)-1/4)$	No	No
Model VI	Centered t: t_4	No	Yes

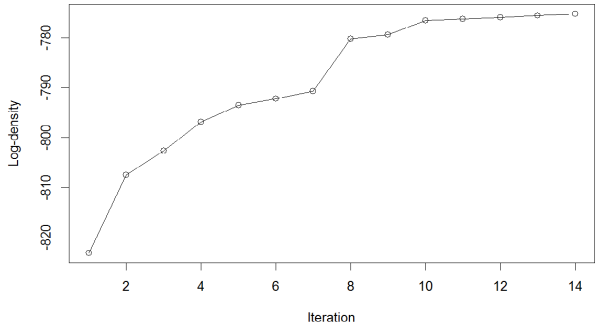
error density (green dashed line) approximates the true density (red solid line) well, even under a finite sample size of 400.

There is a well-known *label switching* issue when sorting the labels for mixture models (Stephens, 2000; Yao and Lindsay, 2009). In this paper, we adopt the method of Yao (2015) to find the labels by minimizing the distance between the estimated classification probabilities and the true labels over different permutations. After sorting the labels, we compute the MSE of all parameters over the N replicates, i.e. $MSE = N^{-1}\sum_{h=1}^{N}(\hat{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_0)^2$, where $\hat{\boldsymbol{\theta}}_h = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T, \hat{\lambda})_h^T$ is the vector of parameter estimates of the h-th replicate and $\boldsymbol{\theta}_0$ is the true value for the vector of the parameters. As the mixtures of regression models serve as a methodology for classification, we compute the average of misclassification numbers (AMN) as well.

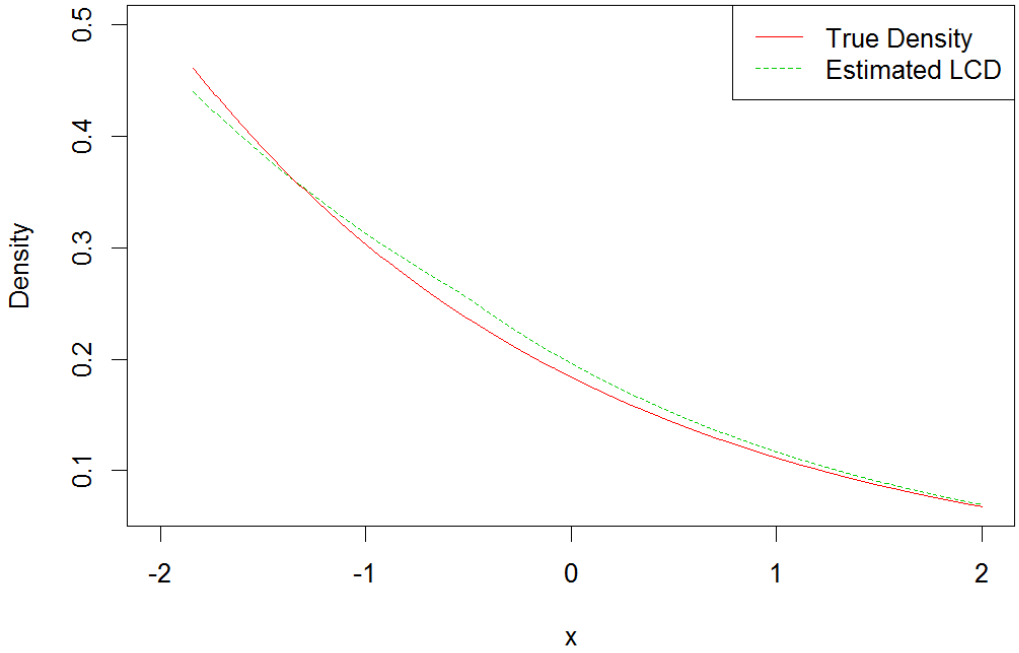
Table 4.2 displays the MSEs of parameter estimates and the average of misclassification numbers over $N = 200$ simulations for Algorithm 4.1. For the density which is not normally distributed, even is not log-concave (Model V and VI), Algorithm 4.1 demonstrates significant improvement over the traditional normal mixture EM algorithm in terms of much smaller MSE. Especially for Model II, III and V, many MSEs from LCD-EM1 are 30% less than those from the Normal-EM1. When the error density truly comes



(a)



(b)



(c)

Figure 4.1: FMR Model III. Plot (a) is the scatter plot of the data generated from Model III's setup for a single replicate. Plot (b) shows that the log-likelihood is monotone increasing through iterations. Plot (c) shows that the estimated density $\hat{g}$ (green dashed line) approximates the true density (centered Exponential, red solid line) well for the finite sample size of 400.

Table 4.2: Simulation results for Model I-VI.

Model	Method	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{0,2}$	$\beta_{1,2}$	λ	AMN
I	LCD-EM1	0.03096	0.01028	0.00874	0.00380	0.00081	41.55
	Normal-EM1	0.02671	0.00992	0.00819	0.00348	0.00072	41.43
II	LCD-EM1	**0.00847**	**0.00273**	**0.00273**	**0.00051**	**0.00101**	**26.95**
	Normal-EM1	0.01310	0.00341	0.00416	0.00134	0.00671	29.68
III	LCD-EM1	**0.10955**	**0.02746**	**0.02039**	**0.01676**	**0.00304**	**47.49**
	Normal-EM1	0.14997	0.04237	0.03357	0.03090	0.00402	62.17
IV	LCD-EM1	**0.03794**	0.01526	**0.01304**	**0.00475**	0.00152	54.25
	Normal-EM1	0.05371	0.01558	0.01538	0.00686	0.00146	55.86
V	LCD-EM1	**0.01639**	**0.00317**	**0.00695**	**0.00031**	0.00113	**33.13**
	Normal-EM1	0.05458	0.01504	0.02268	0.00480	0.00121	51.66
VI	LCD-EM1	**0.10205**	**0.27827**	**0.01654**	**0.01300**	0.00245	53.45
	Normal-EM1	0.13040	0.32327	0.02011	0.01813	0.00236	54.91

from normal distribution (Model I), the new algorithm still has comparable performance and the MSEs of LCD-EM1 are only a little bit higher than those from the traditional normal mixture EM algorithm.

To evaluate the performance of Algorithm 4.2, we generate data from 2-component log-concave FMR models, where $e_{i,1}$ and $e_{i,2}$ are from different families of distributions:

$$y_i = \begin{cases} \beta_{0,1} + \beta_{1,1}x_i + e_{i,1} & \text{with probability} \quad \lambda; \\ \beta_{0,2} + \beta_{1,2}x_i + e_{i,2} & \text{with probability} \quad 1-\lambda. \end{cases} \tag{4.17}$$

where x_i's are independently generated from Uniform$(-1, 3)$. We set $n = 400, \lambda = 0.4$, $\boldsymbol{\beta}_1 = (\beta_{0,1}, \beta_{1,1})^T = (0, 1)^T$, and $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2})^T = (-3, 4)^T$. The component error densities are generated based on the parametric form of Model VII to Model IX in Table 4.3.

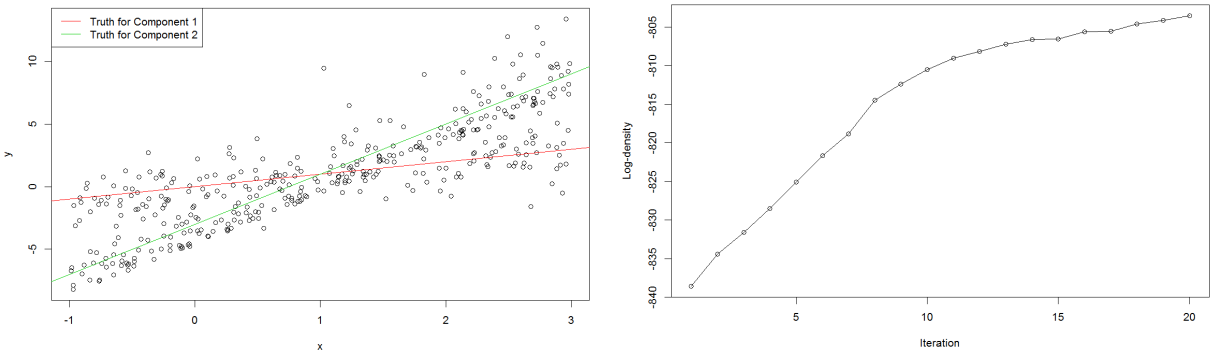
Table 4.3: The error densities for Model VII to Model IX.

	$e_{i,1}$'s distribution	$e_{i,2}$'s distribution
Model VII	$N(0,1)$	$N(0,0.25)$
Model VIII	$3(Beta(1,2)-1/3)$	$N(0,0.25)$
Model IX	$\frac{2}{3}Laplace(0,1)$	$\frac{2}{3}(Exp(2)-2)$

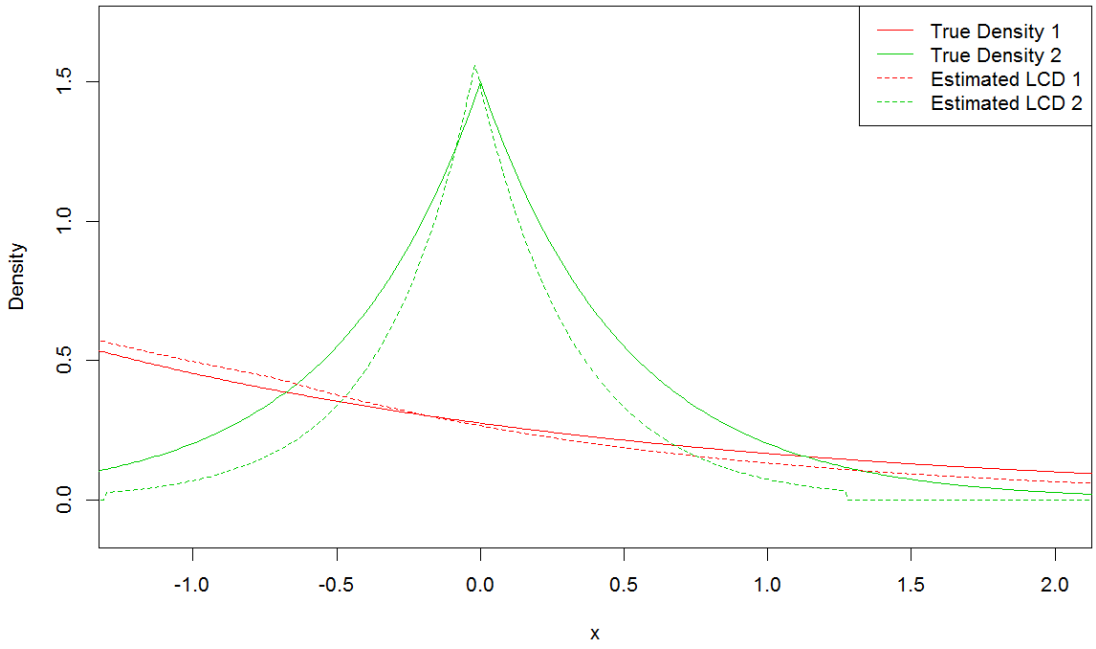
Similar to what we did for Model III, we plot the generated data based on the setup of Model IX for a single replicate in Figure 4.2(a). We fit this generated data with Algorithm 4.2 and demonstrate the property of monotone increasing log-likelihood values for Algorithm 4.2 in Figure 4.2(b). Figure 4.2(c) shows that for this replicate, the fitted log-concave error densities for the two components (red and green dashed lines) approximate the true densities (red and green solid line) for both two components well under a finite sample size of 400.

We report the result over 200 replicates and compare the same criteria as we did for (4.16). Similar phenomena (shown in Table 4.4) are observed for Algorithm 4.2. For the component that truly comes from normal distribution (Model VII and Component 2 of Model VIII), our proposed algorithm has comparable performance to the normal EM algorithm with unequal variances and a similar trimming technique. For the components which are misspecified (Model IX and component 1 of Model VIII), potential improvements are gained if we apply LCD-EM2 instead of the Normal-EM2.

One important feature of the FMR model is that it serves as a tool of unsupervised learning. Consequently, we compare the average number of misclassifications (AMN's) among the 200 replicates in Table 3.1 and Table 3.3. For Model I and Model VII, the average misclassification numbers for our EM-type algorithms are almost the same or only a little bit higher than the normal EM algorithm with similar trimming techniques. When the models are misspecified, the average misclassification numbers obtained from



(a) (b)



(c)

Figure 4.2: FMR Model IX. Plot (a) is the scatter plot of the data generated from Model IX's setup for a single replicate. Plot (b) shows that the log-likelihood is increasing through iterations. Plot (c) shows that the estimated densities $\hat{g}_1$ and $\hat{g}_2$ (red and green dashed lines) approximate the true densities (red and green solid lines) well for the finite sample size of 400.

Table 4.4: Simulation results for Model VII-IX.

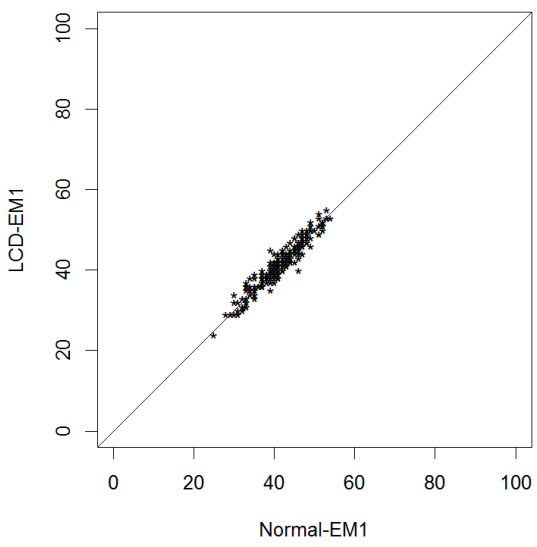
Model	Method	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{0,2}$	$\beta_{1,2}$	λ	AMN
VII	LCD-EM2	0.01473	0.00700	0.00187	0.00088	0.00117	30.21
	Normal-EM2	0.01390	0.00551	0.00175	0.00081	0.00087	29.88
VIII	LCD-EM2	**0.00543**	**0.00122**	0.00199	0.00080	0.00084	26.97
	Normal-EM2	0.00633	0.00266	0.00183	0.00073	0.00075	27.89
IX	LCD-EM2	**0.00658**	**0.00436**	**0.03836**	0.00004	**0.00022**	**49.20**
	Normal-EM2	0.01943	0.01671	0.08099	0.00004	0.00185	61.28

log-concave FMRs are smaller than those from the normal mixture EM algorithm with similar trimming techniques.

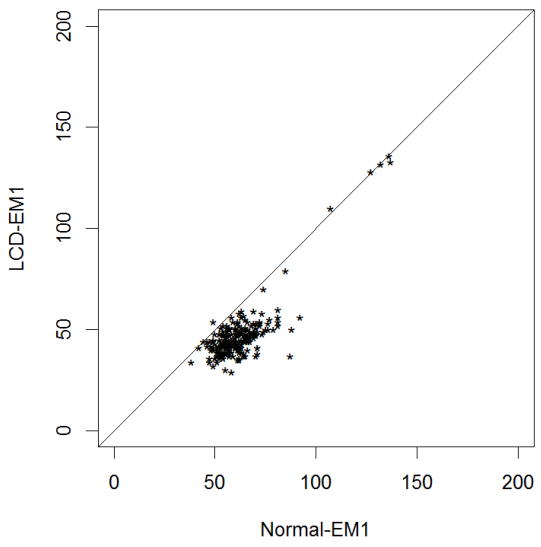
To further illustrate the classification result, we show the classification results for every replicate in Model I, III, VII and IX. In Figure 4.3, each point represents a single replicate. The x-axis represents the number of misclassifications by normal mixture EM algorithm. The y-axis represents the number of misclassifications by our log-concave mixture EM algorithm. We observe significant improvement in the sense of misclassification rates when the models are misspecified (in Figure 4.3(a) and (c), the majority of points are under the identical line). When the component error densities are indeed normal, we observe no significant penalties if we apply the log-concave EM algorithm (Figure 4.3(b) and (d)). The classification plots of the rest Models are presented in the Appendix at the end of this chapter.

4.5 Data Analysis

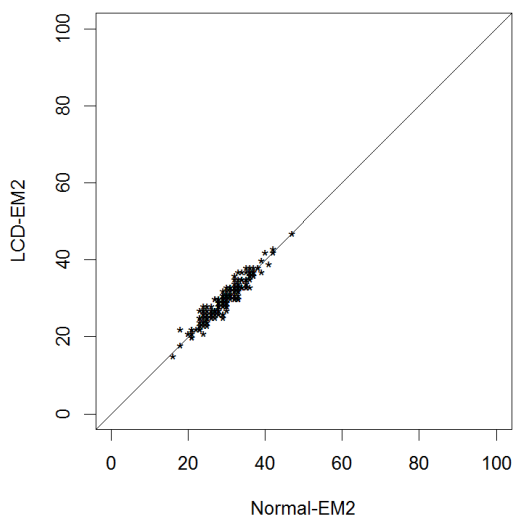
The tone dataset (from package `mixtools`) contains 150 trials from the same musician; see Cohen (1980) for a detailed description. In each trial, a fundamental tone, which was



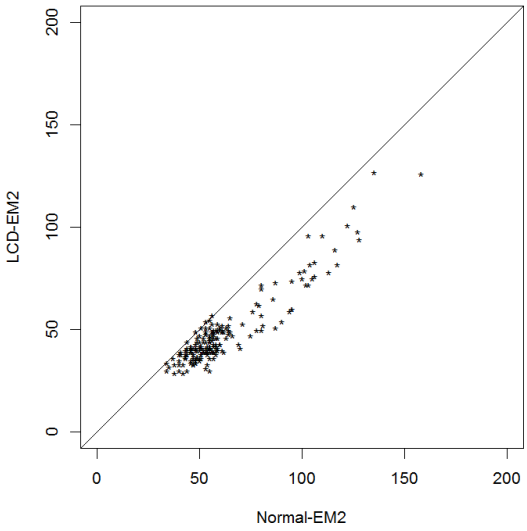
(a) Model I



(b) Model III



(c) Model VII



(d) Model IX

Figure 4.3: Numbers of misclassifications for Model I, III, VII and IX: normal mixture EM algorithm vs log-concave mixture EM algorithm for mixtures of regression models. The solid lines represents the identity. For most replicates, the log-concave FMR significantly improves the classification results as an unsupervised learning method.

purely determined by a stretching ratio, was first provided to the musician. Then the musician tuned the tone one octave above. The tuning ratio, which was measured as the adjusted tone divided by the fundamental tone, was recorded. The purpose of this experiment was to demonstrate the "two musical perception theory". We give the scatter plot of the data in Figure 4.4.

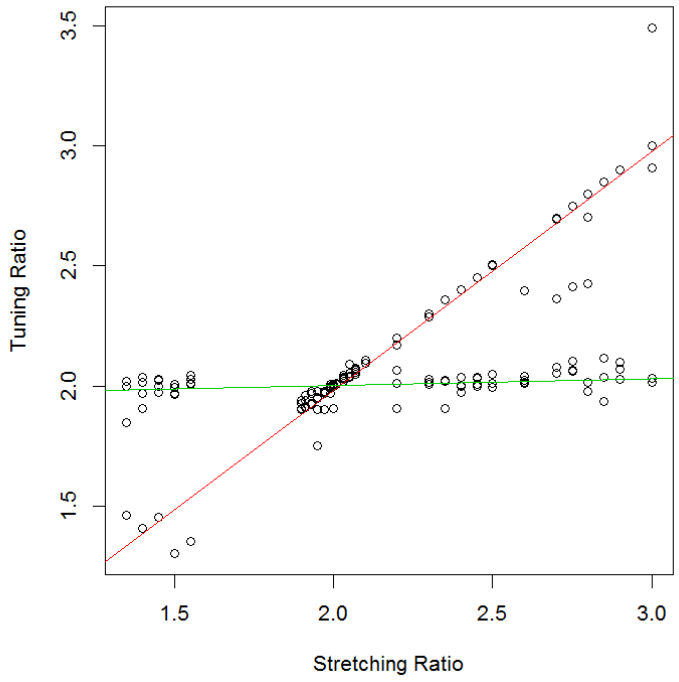


Figure 4.4: Tone data from the tone perception study of Cohen (1980) and the coefficients fitted by LCD-EM1.

For the entire dataset, by applying Algorithm LCD-EM1 with $K = 2$, we obtain the fitted coefficients (shown as the solid lines in Figure 4.4) and the fitted log-likelihood value. We refit the data with Algorithm Normal-EM1, and report the same criteria as

we did for Algorithm 4.1.

To further demonstrate the prediction power of Algorithms 4.1, we apply a 10-folder cross validation to the data set. Denote the full data set as $\mathcal{D}$. We randomly partition $\mathcal{D}$ into a training set $\mathcal{R}_h$ and a testing set $\mathcal{T}_h$ with the property $\mathcal{D} = \mathcal{R}_h + \mathcal{T}_h$ for $h = 1, \ldots, H$, where $H = 10$. For each folder $h \in 1, \ldots, H$, we estimated the parameters $\hat{\lambda}_j^h$'s and $\hat{\boldsymbol{\beta}}_j^h$'s, as well as the estimated log-concave density g^h through the training set $\mathcal{R}_h$. We then calculate the following two types of mean square errors:

- $E_1 = H^{-1} \sum_{h=1}^{H} \sum_{i \in \mathcal{T}_h} \sum_{j=1}^{K} \hat{p}_{ij}^h \{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^h)^2\}$;
- $E_2 = H^{-1} \sum_{h=1}^{H} \sum_{i \in \mathcal{T}_h} \min_j \{(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^h)^2\}$;

where $\hat{p}_{ij}^h$ is the estimated probability that (x_i, y_i) is from j-th component for folder h:

$$\hat{p}_{ij}^h = \frac{\hat{\lambda}_j^h g^h(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^h)}{\sum_{m=1}^{K} \hat{\lambda}_m^h g^h(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_m^h)},$$

for $i \in \mathcal{T}_h$ and $j \in \{1, \ldots, K\}$.

We report the same criteria based on the coefficients obtained by the Normal-EM1 algorithm. The results of fitting the log-concave FMR model and the normal FMR model are summarized in Table 4.5. The fitted result obtained by LCD-EM1 has a much larger log-likelihood. Additionally, Algorithm 4.1 provides much smaller mean square errors for both E_1 and E_2, which indicates that our proposed algorithm predicts the response more precisely than the traditional normal EM algorithm.

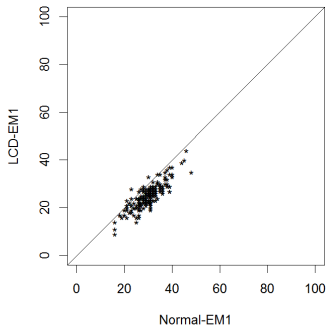
Table 4.5: Estimated parameters and other characteristics of LCD-EM algorithm and Normal-EM algorithm for the tone dataset.

	LCD-EM1		Normal-EM1	
	Comp 1	Comp 2	Comp 1	Comp 2
β_0	-0.0143	1.9488	-0.0388	1.8924
β_1	0.9968	0.0263	0.9989	0.0559
λ	0.4253	0.5747	0.3256	0.6744
ℓ	**170.91**		158.54	
E_1	**0.0039**		0.0105	
E_2	**0.0033**		0.0041	

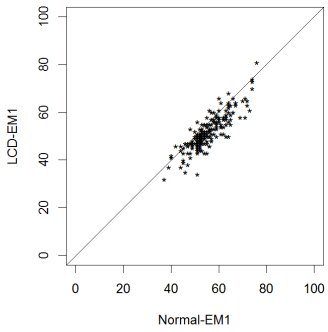
4.6 Conclusion

This paper proposed two robust EM-type algorithms for the log-concave mixtures of regression models. These algorithms provide more flexibility, which allows a large family of error densities in the mixtures of regression models. By estimating the log-concave error density in every M-step of our algorithms, the log-concave maximum likelihood estimator corrects the model misspecification, e.g. adjusting skewness and heavy tails when the error distribution is not normal. Through numerical studies, significant improvements for the two proposed algorithms are observed while comparing them with the normal mixture EM algorithm. Future work includes, but is not limited to the theoretical investigation of the log-concave mixtures of regression models, as an extension of Section 3 of Dümbgen et al. (2011).

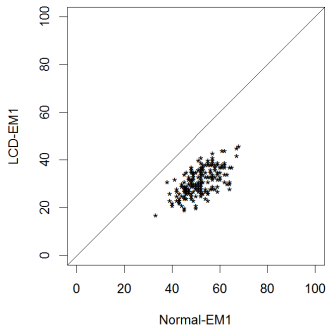
4.7 Appendix



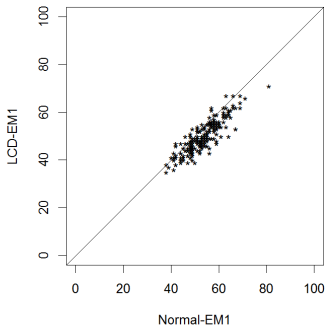
(a) Model II



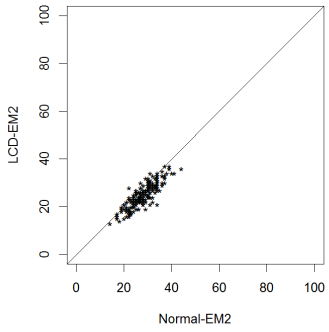
(b) Model IV



(c) Model V



(d) Model VI



(e) Model VIII

Figure 4.5: Numbers of misclassifications for Model II, IV, V, VI, and VIII: normal mixture EM algorithm vs log-concave mixture EM algorithm for mixtures of regression models. The solid lines represents the identity. For all these models, log-concave FMR significantly improve the classification, even for the model which is not actually log-concave distributed (Model V and VI).

Chapter 5

Log-concave Mixtures of Regressions Models with covariates-dependent Mixing Proportions

5.1 Introduction

Consider the $CO_2 - GDP$ Data in Example 1.3. In Figure 1.4, clearly, there exists two or more different patterns between the CO_2 emission and GDP. One group of countries is more energy consuming (referred as Group A, which is expected to have a larger slope for the regression line). The other group is more environmental friendly (referred as Group B, which is expected to have a lower slope for the regression model). However, when the GPD per capita is higher, the country will most likely fall into Group B. If we are given a new data point with GPD per capita of 35,000-50,000, we may conclude it's more likely to fall into Group B, which is the wealthier and more environmental friendly group.

Fitting a model like (4.1) would not capture such dependent structure between the

covariates and the mixing proportions. Thus, we extend the model to a more general form of:

$$f(\mathbf{x}|G,\lambda,\beta)=\sum_{i=1}^{n}\log\sum_{j=1}^{K}\lambda_j(x_i)g_j(y_i-\mathbf{x}_i^T\boldsymbol{\beta}_j),\ \mathbf{x}_i\in\mathbb{R}^p, \tag{5.1}$$

where g_j's are log-concave, i.e. $\log g_j$'s are concave. For $\lambda_j(x_i)\equiv\lambda_j$, it reduces to the log-likelihood of the log-concave FMR model (4.2) which is found in Chapter 4.

Considering the flexible structure of (5.1) instead of (4.2) will not only potentially improve the parameter estimation, but also improve the estimation when given new covariates. Suppose that we've fitted a log-concave mixture of regressions model like (4.2) with n observations $\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$. Now we are given a new observation with the covariate $\mathbf{x}_{n+1}$ and we want to predict the response of the new observation. With K regression equations in the model, we end up with K different predictive values with weights of λ_j's. If we allow λ_j to depend on certain covariates, we can better adjust the weights of the K predictive values.

There are two ways to solve covariate-dependent mixing proportions in the EM algorithm while characterizing the covariate-dependent FMR models with normal error densities. The first way is to fit a ME (*Mixtures of Expert*, see Jacobs et al. (1991)), which assumes a logistic-type structure between covariates and mixing proportions. However, certain parametric forms of $\lambda_j(\mathbf{x}_i)$ may still be too restrictive and cannot characterize certain structures of $\lambda_j(\mathbf{x}_i)$, since the ME structure is strictly monotone. Alternatively, we may model $\lambda_j(\mathbf{x}_i)$ non-parametrically through a kernel density estimation, which makes the dependence structure more flexible, but brings additional difficulty in terms of tuning, see Derek S (2007) and Huang (2009) for different tuning procedures. In this chapter, we adopt these two approaches and combine them with the log-concave FMR model with covariate-dependent mixing proportions.

5.2 Mixture of Expert

The Mixtures of Expert is a supervised learning concept, see Jordan and Jacobs (1994) and Hastie et al. (2005). It contains a logistic type of structure, i.e.

$$\lambda_j(\mathbf{x}_i;\boldsymbol{\theta}_j) = \frac{\exp\{\mathbf{x}_i^T\boldsymbol{\theta}_j\}}{\sum_{j=1}^K \exp\{\mathbf{x}_i^T\boldsymbol{\theta}_j\}},$$

where $j = 1,\ldots,K$ and $\boldsymbol{\theta}_j$ is the unknown parameter vector for component j. When there are multiple layers, it is called the Hierarchical Mixtures of Expert (HME). For more details about ME and HME, readers may refer to Jacobs et al. (1991), Jordan and Xu (1995) and Hastie et al. (2005).

Assuming the observed data y is incomplete and define the missing value $\mathbf{Z} = (\mathbf{z}_1^T, ..., \mathbf{z}_1^T)$ where $\mathbf{z}_i$ is a K-dimension vector with the j-th element given by:

$$z_{ij} = \begin{cases} 1 & \text{if } (\mathbf{x}_i, y_i) \text{ belongs to } j\text{-th group,} \\ 0 & \text{otherwise .} \end{cases}$$

The complete log-likelihood is:

$$\log f(\psi, \mathbf{G}; \mathbf{y}, \mathbf{z}, \mathbf{X}) = log \prod_{i=1}^{n}\prod_{j=1}^{K}[\lambda_j(\mathbf{x}_i)g(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j)]^{Z_{ij}} = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}\log\lambda_j(\mathbf{x}_i)g(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_j),$$

where $\mathbf{y} = (y_1,\ldots,y_n)^T, \mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$.

For now we assume that $g_1 = g_2 = \ldots = g_K = g$.

Algorithm 5.1. *EM algorithm that estimating $\lambda_j(\boldsymbol{x}_i)$'s via ME.*

1. *Initial values $\boldsymbol{\psi}^{(0)}$ and $z_{ij}^{(0)}$ from package* `Mixtools`.

2. *E-step: Calculate*

$$z_{ij}^{(t+1)} = E(Z_{ij}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\psi}^{(t)}, g^{(t)}) = Pr(Z_{ij} = 1|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\psi}^{(t)}, g^{(t)})$$
$$= \frac{\lambda_j^{(t)}(\boldsymbol{x}_i)\widehat{g}^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t)})}{\sum_{h=1}^{K}\lambda_h^{(t)}(\boldsymbol{x}_i)\widehat{g}^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_h^{(t)})}.$$

3. *Update $\boldsymbol{\theta}$ via finding the solution of*

$$\frac{\partial b(\boldsymbol{\theta}_j)}{\partial \boldsymbol{\theta}_j} = \boldsymbol{0}, \tag{5.2}$$

where $b(\boldsymbol{\theta}_j) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(t)} \log \lambda_j(\boldsymbol{x}_i; \boldsymbol{\theta}_j)$. The equation (5.2) can be solved by an iteratively re-weighted least square (IRLS).

4. *M-step: The M-step on the t-th iteration maximizes $Q(\boldsymbol{\psi}, g; \boldsymbol{\psi}^t, g^{(t)})$ with respect to $\boldsymbol{\psi}$ and g.*

   (a) *Update λ simply through*

$$\lambda_j^{(t+1)}(\boldsymbol{x}_i) = \frac{\exp\{\boldsymbol{x}_i^T\boldsymbol{\theta}_j^{(t+1)}\}}{\sum_{j=1}^{K}\exp\{\boldsymbol{x}_i^T\boldsymbol{\theta}_j^{(t+1)}\}},$$

   for $j = 1, \ldots, K$.

   (b) *Update $\boldsymbol{\beta}$ through "`optim`"*

$$\boldsymbol{\beta}_j^{(t+1)} = \arg\max_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} z_{ij}^{(t+1)} \log \widehat{g}^{(t)}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_j^{(t+1)}),$$

   for $j = 1, \ldots, K$.

(c) Estimate g by maximizing

$$Q(\boldsymbol{\psi}, g; \boldsymbol{\psi}^t, g^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{K} z_{ij}^{(t+1)} \log g(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)}),$$

with respect to g through the function called `logConcDEAD` *and get estimator $\widehat{g}^{(t+1)}$. Similar as what we discussed in Section 4.3, the $z_{ij}^{(t+1)}$'s can be viewed as weights of the Kn residuals when estimating the log-concave density g.*

5.3 Nonparametric Covariate-dependent mixing proportions

A major drawback of the ME structure is that the logistic type structure can only characterize a monotone relationship. Alternatively, we propose a nonparametric structure for the covariate-dependent mixing proportions. Similar techniques were applied by Young and Hunter (2010), Hunter and Young (2012) and Huang and Yao (2012). In this chapter, we replace $\lambda_j(\mathbf{x}_i)$ with

$$\lambda_j(x_i) = \frac{\sum_{l=1}^{n} z_{l,j} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_l)}{\sum_{l=1}^{n} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_l)}, \tag{5.3}$$

where

$$\mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_l) = \frac{1}{h_1 \cdots h_{p-1}} \mathcal{K}(\frac{x_{i,1} - x_{l,1}}{h_1}, \cdots, \frac{x_{i,p-1} - x_{l,p-1}}{h_{p-1}}). \tag{5.4}$$

The nonparametric approach is more flexible than the ME methods. However, it adds difficulty in terms of tuning. To tune the bandwidth, we use cross-validation. Suppose we have full data set $\mathcal{D}$. For each of the T folders, we partition the data into training set

$\mathcal{R}_t$ and test set $\mathcal{T}_t$, $t \in \{1, \ldots, T\}$. In practice, we set $T = 10$, which usually works very well.

For the t-th folder, we build the learner using the training set $\mathcal{R}_j$. We are trying to minimize:

$$CV(h) = \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} \{y_i - \hat{y}_i\}^2, \tag{5.5}$$

where $\hat{y}_i = \sum_{j=1}^{K} \hat{z}_{ij} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^t$. The estimated $\lambda_j^t(\cdot)$ and $\hat{\boldsymbol{\beta}}_j^t$ are obtained through the training set $\mathcal{R}_t$. The posterior $\hat{z}_{ij}$ is obtained by:

$$\hat{z}_{ij} = \frac{\lambda_j^t(\mathbf{x}_i)\hat{g}(y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_j^t)}{\sum_{q=1}^{K} \lambda_q^t(\mathbf{x}_i)\hat{g}(y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}}_q^t)}.$$

The $\lambda_j^t(\cdot)$'s are obtained by

$$\lambda_j^t(\mathbf{x}_i) = \frac{\sum_{l \in \mathcal{R}_t} z_{l,j} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_l)}{\sum_{l \in \mathcal{R}_t} \mathcal{K}_h(\mathbf{x}_i - \mathbf{x}_l)}, \tag{5.6}$$

for $i \in \mathcal{T}_t$ and $j \in \{1, \ldots, K\}$.

Algorithm 5.2. *EM algorithm that estimating λ's non-parametrically.*

1. *Initial values $\boldsymbol{\psi}^{(0)}$ and $z_{ij}^{(0)}$ from package* `Mixtools`.

2. *E-step: Same as Algorithm 5.1*

3. *M-step: The M-step on the t-th iteration maximizes $Q(\boldsymbol{\psi}, g; \boldsymbol{\psi}^t, g^{(t)})$ with respect to $\boldsymbol{\psi}$ and g.*

   (a) *Update λ simply through*

$$\lambda_j^{(t+1)}(x_i) = \frac{\sum_{l=1}^{n} z_{l,j}^{(t+1)} \mathcal{K}_h(\boldsymbol{x}_i - \boldsymbol{x}_l)}{\sum_{l=1}^{n} \mathcal{K}_h(\boldsymbol{x}_i - \boldsymbol{x}_l)},$$

for $j = 1, \ldots, K$.

(b) Update $\boldsymbol{\beta}$ *through "*`optim`*"*

$$\boldsymbol{\beta}_j^{(t+1)} = \arg\max_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} z_{ij}^{(t+1)} \log \widehat{g}^{(t)}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)}),$$

for $j = 1, \ldots, K$.

(c) Estimate g *by maximizing*

$$Q(\boldsymbol{\psi}, g; \boldsymbol{\psi}^t, g^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij}^{(t+1)} \log g(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}_j^{(t+1)}),$$

with respect to g *through the function called* `logConcDEAD` *and get estimator* $\widehat{g}^{(t+1)}$. *Similar as what we discussed in Section 4.3, the* $z_{ij}^{(t+1)}$*'s can be viewed as weights of the* Kn *residuals when estimating the log-concave density* g.

5.4 Simulations

5.4.1 Example 1: Monotone increasing $\lambda_j(\mathbf{x}_i)$ Structure

We generate $n = \{200, 400, 600\}$ data points from a 2-component mixture regression model:

$$y_i = \begin{cases} \beta_{0,1} + \beta_{1,1} x_i + e_i & \text{with probability} \quad \lambda(x_i), \\ \beta_{0,2} + \beta_{1,2} x_i + e_i & \text{with probability} \quad 1 - \lambda(x_i), \end{cases} \tag{5.7}$$

where x_i is from Uniform(0,1), $\boldsymbol{\beta}_1 = (\beta_{0,1}, \beta_{1,1})^T = (-1, 3)^T$, and $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2})^T = (2, -1)^T, e_i \sim 0.1(Gamma(2, 1) - 2)$. Moreover, we let

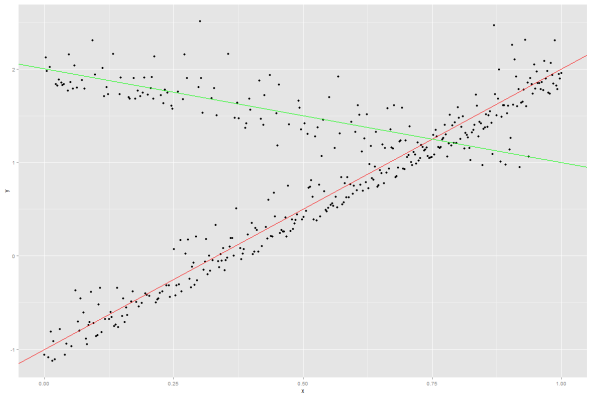
$$\lambda_1(x_i) = \frac{\exp(2x)}{1 + \exp(2x)},$$

where the ME structure will correctly specify the covariate-dependent structure.

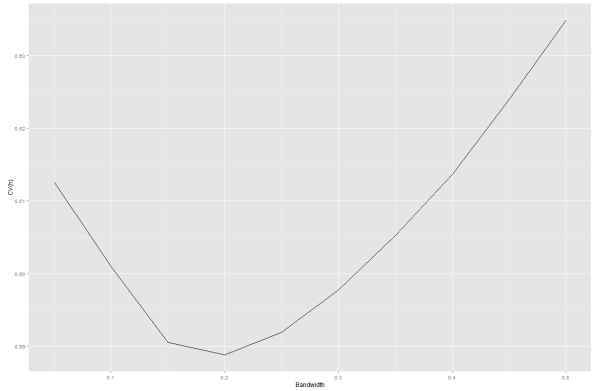
We estimate the model through the following four methods: Algorithm 5.2 (NP-LCD), Algorithm 5.1 (ME-LCD), Algorithm 4.1 (LCD-EM1, which is the log-concave FMR model with fixing mixing proportions and $s = 0$), and normal mixture EM algorithm for equal variance (Normal-EM1, $s = 0$), over $N = 200$ replicates. To estimate $\lambda_1(x_i)$'s by NP-LCD, we use a Gaussian Kernel. For each n, we use the tuning criteria $CV(h)$ from (5.5). To save computational time and avoid selecting the bandwidth for every replicate in the simulation, we repeat the tuning process for the first 10 replicates and average their optimal bandwidths to $\hat{h}$. As suggested by Huang and Yao (2012), we consider three different bandwidths in our simulation: the under-smoothing bandwidth $n^{-2/15}\hat{h}$, the optimal bandwidth $\hat{h}$, and the over-smoothing bandwidth $2\hat{h}$.

We simulate and plot the data for a single replicate with a sample size of 400 in Figure 5.1(a). We show the cross-validation criteria $CV(h)$ vs the bandwidth h to select the optimal bandwidth for this replicate in Figure 5.1(b). To further demonstrate the estimated $\lambda(\mathbf{x}_i)$'s, we plot the estimated mixing proportions (dotted lines) and the theoretical values (solid lines) in Figure 5.1(c) and (d). For both ME-LCD and NP-LCD with optimal bandwidth, the estimated mixing proportions approximate theoretical values well.

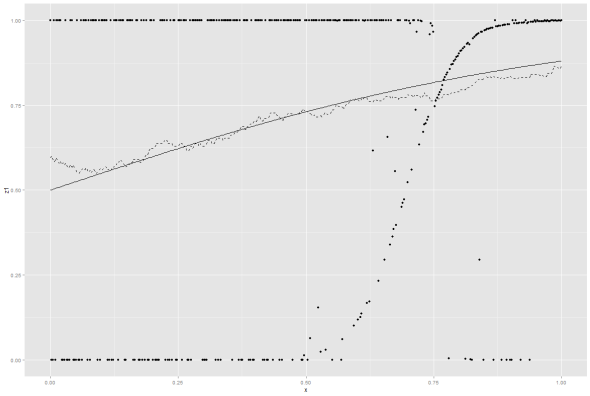
We repeat the simulation $N = 200$ times and compute the mean square errors (MSE)



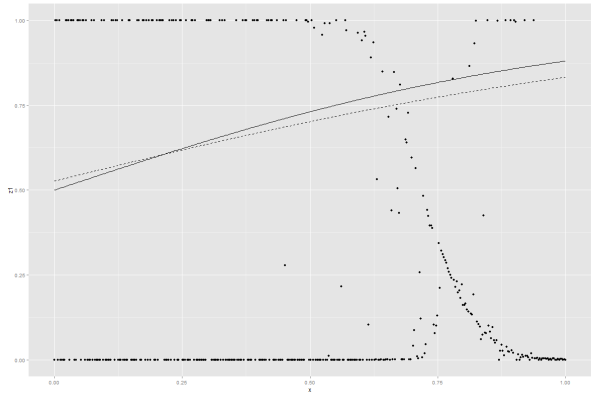
(a) 400 generated values



(b) Selecting the optimal bandwidth through cross-validation



(c) λ vs $\hat{\lambda}$ obtained by Algorithm 5.2



(d) λ vs $\hat{\lambda}$ obtained by Algorithm 5.1

Figure 5.1: Simulation setup of Example 1. (a) 400 observations generated by the setup of (5.7). (b) Selecting the optimal bandwidth through cross-validation. Plots of estimated posterior membership probabilities vs the predictors from (c) Algorithm 5.2 ($h = 0.20$), and (d) Algorithm 5.1. The solid line represents the theoretical value. The dotted line represents the estimated mixing proportion.

for the $\boldsymbol{\beta}$'s and the average square errors (ASE) of the mixing proportions:

$$ASE = (Nn)^{-1} \sum_{q=1}^{N} \sum_{i=1}^{n} (\hat{\lambda}(x_i^{(q)}) - \lambda_0(x_i^{(q)}))^2,$$

where $x_1^{(q)}, \ldots, x_n^{(q)}$ are the observations for replicate $q \in \{1, \ldots, N\}$. We compute these values for optimal bandwidth $\hat{h}$, under-smoothing bandwidth $n^{-2/15}\hat{h}$, and over-smoothing bandwidth $2\hat{h}$ for NP-LCD. We also compute the same criteria for ME-LCD, LCD-EM1, and Normal-EM1. We average these criteria over the 200 replicates and present them in Table 5.1.

We observe significant improvement for both NP-LCD and ME-LCD estimation, especially in the ASEs of λ's. We observe about 70%-90% decrease in ASEs for λ's when comparing ME-LCD with LCD-EM. NP-LCD's performances are slightly worse than ME-LCD, but still improve a lot when comparing with both Normal-EM1 and LCD-EM1.

5.4.2 Example 2: Bell-shape $\lambda_j(\mathbf{x}_i)$ Structure

We generate $n = \{200, 400, 600\}$ data points from a 2-component mixture regression model:

$$y_i = \begin{cases} \beta_{0,1} + \beta_{1,1} x_i + e_i & \text{with probability} \quad \lambda(x_i), \\ \beta_{0,2} + \beta_{1,2} x_i + e_i & \text{with probability} \quad 1 - \lambda(x_i), \end{cases} \tag{5.8}$$

where x_i is from Uniform(0,1), $\boldsymbol{\beta}_1 = (\beta_{0,1}, \beta_{1,1})^T = (-1, 5)^T$, and $\boldsymbol{\beta}_2 = (\beta_{0,2}, \beta_{1,2})^T = (0, 1)^T, e_i \sim Laplace(0, 0.1)$. Moreover,

$$\lambda_1(x_i) = 0.7 sin(\pi x) + 0.1, \tag{5.9}$$

Table 5.1: MSEs for $\boldsymbol{\beta}$'s and ASEs for $\lambda_j(x)$'s for Example 1.

	Model(bandwidth)	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{0,2}$	$\beta_{1,2}$	λ
n=200	NP-LCD(h=0.17)	0.02221	0.02820	0.02858	0.06882	0.00697
	NP-LCD(h=0.35)	0.02475	0.03233	0.02869	0.06647	0.00417
	NP-LCD(h=0.70)	0.02347	**0.03026**	0.02482	0.05882	0.00418
	ME-LCD	**0.02312**	0.03222	**0.02465**	**0.05610**	**0.00270**
	LCD-EM1	0.03077	0.05031	0.03843	0.08974	0.01464
	Normal-EM1	0.03063	0.04742	0.06781	0.19052	0.01502
n=400	NP-LCD(h=0.09)	0.01563	0.02062	0.02347	0.05999	0.00665
	NP-LCD(h=0.20)	**0.01508**	**0.02004**	0.02295	0.06181	0.00344
	NP-LCD(h=0.32)	0.01556	0.02057	0.02339	0.06025	0.00246
	ME-LCD	0.01554	0.02248	**0.02282**	**0.05980**	**0.00176**
	LCD-EM1	0.02198	0.03584	0.03127	0.08557	0.01408
	Normal-EM1	0.02351	0.03657	0.06809	0.19904	0.01449
n=600	NP-LCD(h=0.06)	0.01085	0.01683	0.01974	0.04236	0.00562
	NP-LCD(h=0.15)	**0.01029**	**0.01472**	0.02279	0.04560	0.00289
	NP-LCD(h=0.30)	0.01058	0.01938	0.02004	0.04033	0.00203
	ME-LCD	0.01085	0.01828	**0.02003**	**0.04010**	**0.00123**
	LCD-EM1	0.01557	0.03045	0.02398	0.05834	0.01378
	Normal-EM1	0.01625	0.03050	0.05281	0.15293	0.01368

where the ME structure cannot characterize this relationship.

We use a similar tuning process, select the optimal bandwidth, and estimate the model through NP-LCD of three different bandwidths, ME-LCD, LCD-EM1($s = 0$), and Normal-EM1($s = 0$).

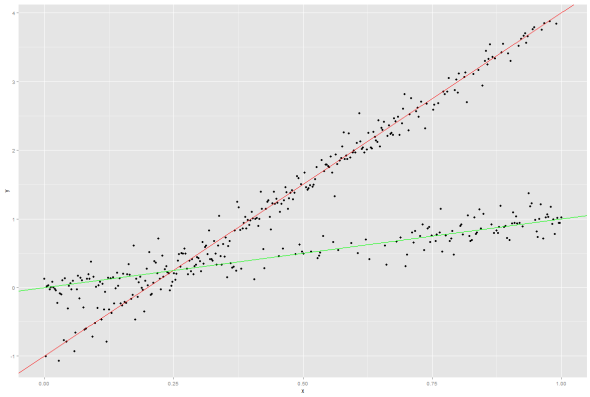
To further illustrate this performance of the algorithms, we simulate and plot the data for a single replicate with a finite sample size of 400 in Figure 5.2(a). We show the cross-validation criteria $CV(h)$ vs the bandwidth h to select the optimal bandwidth for this replicate in Figure 5.2(b). Figure 5.2(c) shows that the NP-LCD with optimal

bandwidth(dotted lines) approximate theoretical values (solid lines) well. However, Figure 5.2(d) shows that the estimated mixing proportion from ME-LCD (dotted lines) is flat and cannot characterize the theoretical values (solid lines) at all.

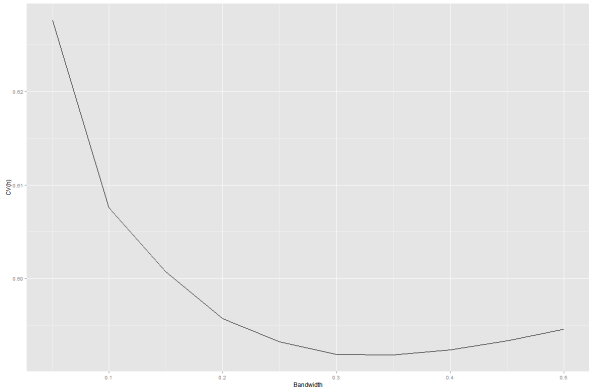
Table 5.2: Mean of MSE for $\boldsymbol{\beta}$'s and for $\lambda_j(x)$'s for Example 2.

	Model(bandwidth)	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{0,2}$	$\beta_{1,2}$	λ
n=200	NP-LCD(h=0.20)	**0.02958**	**0.04917**	**0.02177**	**0.03223**	**0.00698**
	NP-LCD(h=0.40)	0.02974	0.05006	0.02193	0.03249	0.00976
	NP-LCD(h=0.80)	0.03062	0.05164	0.02215	0.03280	0.03539
	ME-LCD	0.03128	0.05220	0.02217	0.03207	0.04942
	LCD-EM1	0.03314	0.05504	0.02284	0.03378	0.04855
	Normal-EM1	0.03229	0.05415	0.02322	0.03454	0.04856
n=400	NP-LCD(h=0.16)	0.02063	0.03408	**0.01430**	**0.02030**	**0.00452**
	NP-LCD(h=0.35)	**0.02048**	**0.03301**	0.01461	0.02060	0.00709
	NP-LCD(h=0.70)	0.02085	0.03386	0.01571	0.02126	0.02702
	ME-LCD	0.02091	0.03464	0.01508	0.02041	0.04802
	LCD-EM1	0.02299	0.03812	0.01673	0.02431	0.04749
	Normal-EM1	0.02332	0.03857	0.01682	0.02424	0.04752
n=600	NP-LCD(h=0.13)	0.01657	0.02625	0.01229	0.01999	**0.00298**
	NP-LCD(h=0.30)	**0.01653**	**0.02578**	0.01223	0.02064	0.00436
	NP-LCD(h=0.60)	0.01686	0.02735	**0.01191**	**0.01981**	0.01836
	ME-LCD	0.01665	0.02604	0.01246	0.02038	0.04738
	LCD-EM1	0.01797	0.02924	0.01241	0.01987	0.04708
	Normal-EM1	0.01782	0.02934	0.01258	0.01991	0.04706

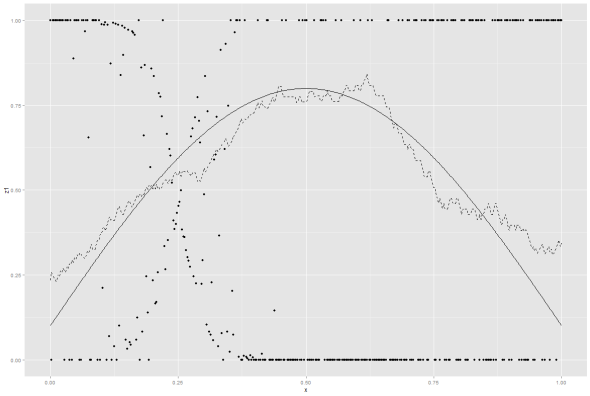
We repeat the simulation 200 times and report the same criteria. As Table 5.2 shows, NP-LCD significantly decreases the MSEs for all $\boldsymbol{\beta}_j$'s. Moreover, NP-LCD decreases $\lambda_j(\mathbf{x}_i)$'s significantly (more than 90%). However, in this example, the ME structure fails to characterize $\lambda(x_i)$ correctly. From Table 5.2, ME-LCD's performance is almost the



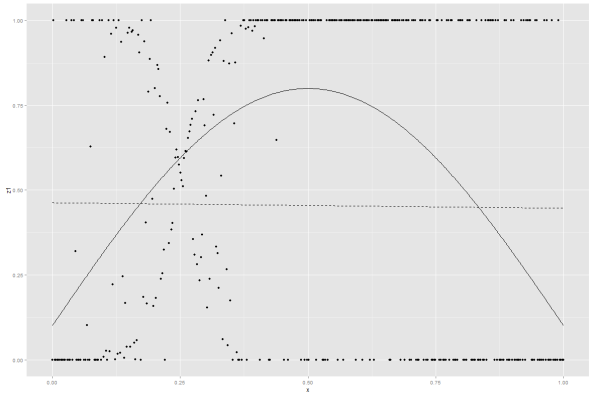
(a) 400 generated values



(b) Selecting the optimal bandwidth through cross-validation



(c) λ vs $\hat{\lambda}$ obtained by Algorithm 5.2



(d) λ vs $\hat{\lambda}$ obtained by Algorithm 5.1

Figure 5.2: Simulation setup of Example 2. (a) 400 observations generated by the setup of (5.8) (b) Selecting the optimal bandwidth through cross-validation. Plots of estimated posterior membership probabilities vs the predictors from (c) Algorithm 5.2 ($h = 0.35$), and (d) Algorithm 5.1. The solid line represents the theoretical value. The dotted line represents the estimated mixing proportion.

same with LCD-EM1's and Normal-EM1's performances and is much worse than the NP-LCD's with all different bandwidths.

5.5 Real data analysis

We apply the two proposed EM algorithms to the CO_2-GDP dataset as we described in Chapter 1 and compare with LCD-EM1 and Normal-EM1, which do not consider the covariate-dependent mixing proportions. By applying the cross-validation criteria of (5.5), we select the optimal bandwidth $\hat{h} \approx 5.85$. In Table 5.3, we present the estimation coefficients of $\boldsymbol{\beta}$'s and λ (only applicable for Normal-EM1 and LCD-EM1), as well as the log-likelihood ℓ for all different algorithms. The estimated coefficients for both NP-LCDs with different bandwidths and ME-LCD (the coefficients estimated by NP-LCD($\hat{h} = 5.85$) is shown in Figure 5.3) are significantly different after considering varying mixing proportion. Without any surprise, all the fitted log-likelihood values are higher than those of FMR model with fixing mixing proportions.

To further illustrate the prediction power of our proposed algorithms, we apply T-folder cross-validation, where $T = 10$. For each of the T folders, we partition the full data set $\mathcal{D}$ into training set $\mathcal{R}_t$ and test set $\mathcal{T}_t$, where $t \in \{1, \ldots, T\}$. For data in each test set $x_i \in \mathcal{T}_t$, we build the learner using the training set and obtain the prediction $\lambda_j^{(t)}(x_i)$ as well as the coefficients $\hat{\boldsymbol{\beta}}_j^t$'s from $\mathcal{R}_t$. We calculate the predicted error sum of squares $PRESS$ by:

$$PRESS = \sum_{t=1}^{T} \sum_{i \in \mathcal{T}_t} \{y_i - \hat{y}_i\}^2,$$

where $\hat{y}_i = \sum_{j=1}^{K} \lambda_j^{(t)}(x_i) \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_j^t$. In Table 5.3, the predicted error sum of squares are significantly decreased if we consider varying mixing proportions for the log-concave

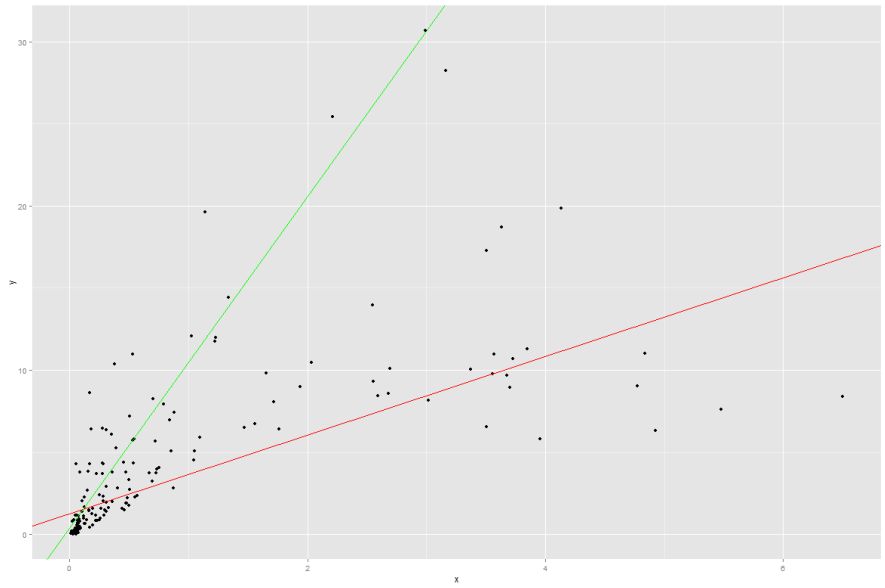


Figure 5.3: The $GDP\text{-}CO_2$ dataset from Example 1.3 and the coefficients fitted by Algorithm 5.2.

FMR models.

Table 5.3: Estimated coefficients and predicted error sum of squares via cross-validation for the $GDP\text{-}CO_2$ dataset.

	Normal-EM1	LCD-EM1	ME-LCD	NP-LCD (h=2.95)	NP-LCD (h=5.85)	NP-LCD (h=11.70)
$\beta_{0,1}$	0.1047	0.1047	0.4020	0.4029	0.4042	0.4020
$\beta_{1,1}$	4.8483	4.8483	10.087	10.087	10.087	10.088
$\beta_{0,2}$	5.3622	5.3622	1.6644	1.6640	1.1274	1.2708
$\beta_{1,2}$	1.5437	1.5437	2.2894	2.2896	2.4340	2.3944
λ_1	0.5807	0.5853	NA	NA	NA	NA
ℓ	-357.31	-324.39	-329.73	-324.82	**-314.55**	-315.03
$PRESS$	19.36	19.13	16.84	4.90	**3.74**	4.67

5.6 Conclusion

We propose two methods, ME-LCD and NP-LCD, for capturing covariate-dependent features that mixing proportions depend on when the mixture regression residuals are not normal. ME-LCD estimates the mixing proportion structure using a logistic-type function. NP-LCD estimates the mixing proportion using kernel methods. While NP-LCD is more flexible than ME-LCD, it introduces in tuning-based complications. Simulation studies and real data studies demonstrates the superior performance of Algorithm 5.1 and Algorithm 5.2, which incorporates the two methods.

Chapter 6

Conclusion

This dissertation combines the log-concave shape constraints with finite mixture models and finite mixtures of regressions models. Rather than assuming a certain parametric family, the log-concave shape constraint surely increases the model's flexibility. This extension from log-concave densities to mixture of log-concave densities considerably broadens the scope of shape-constrained estimation procedures and increases the ability to approximate a large amount of multi-modal distributions. Furthermore, the parameter estimation and classification ability are significantly improved.

6.1 Discussion

In Chapter 3, we extend the LCMLE for a log-concave density to the mixtures of log-concave densities. The main results show the existence of a restricted maximum likelihood estimator (the ratio of the maxima of the components are constrained to be smaller than some pre-specified upper bound), and the consistency of this estimator, under fairly general conditions. These conditions are met in most cases of the real data study. To

guarantee the existence of LCMLE, we only need two conditions: (i) the distribution Q does not collapse on a single point; and (ii) the the tail is not heavy enough to have infinite expectation (such as Cauchy distribution). To guarantee the consistency, in addition to (i) and (ii), we only require that the empirical sequence converges to the truth in Mallows Distance (slightly stronger than weak convergence).

We've also developed an EM-type algorithm in Chapter 3. In every M-step, we update the component densities. Consequently, we call it "EM-type" as the ascending property of traditional parametric EM algorithm no longer holds here. Through numerical studies, we demonstrate superior performance in density estimation and classification.

In Chapter 4, we extend the EM-algorithm for the log-concave mixture model to the mixture of log-concave regressions model. We developed two EM-type algorithms. There are several troublesome issues with these algorithms. The first one is the local maximum. To overcome this problem, we use a stochastic search strategy to select initial values 20 times. We run EM algorithms with those 20 initial values and select the one with the highest log-likelihood value. The second problem is the sensitivity to the outliers. We apply the Least trimmed square method and discard s observations with the least log-likelihood values when updating the parameters in M-steps.

In Chapter 5, we further extend the log-concave mixture of regressions model. We assumes that the mixing proportions are no longer constant, but rather a function depending on the covariates. We use a Mixtures of Expert approach and a Kernel based nonparametric approach to model the flexibility of the $\lambda_j(\mathbf{x}_i)$'s. Mixture of Expert method does not have a tuning issue, but can only handle monotone patterns of the mixing proportions. The kernel-based method does not have restrictive patterns but requires costly bandwidth tuning. For kernel based EM-algorithms, we use cross-validation to select the optimal bandwidth. Numeric simulation demonstrates the great improvement when

considering the covariate-dependent mixing proportions.

6.2 Direction of Future Research

One of major issues of designing mixture models is the identifiability. For now, some restrictive conditions, such as symmetry, are needed to ensure identifiability. Hunter et al. (2007) and Bordes et al. (2006b) proved the identifiability of (1.1) if $K = 2$ and both component densities are symmetric with different location parameters. Balabdaoui and Doss (2014) considered the special case of (3.1), when $\phi_j(x;\theta_j) = \phi(x - \theta_j)$ and ϕ is a symmetric concave function about 0, and the identifiability of (3.1) follows from Hunter et al. (2007) and Bordes et al. (2006b) when $K = 2$. Consequently, it is worth trying to prove the overall identifiability of K-component mixture model from a nonparametric point of view, under a minimum number of conditions. Another issue worth studying is how to select the number of components K, as in some cases, K is unknown. Another issue for the future study is the theoretical properties of the LCMLE (described in Chapter 4) for the log-concave mixture of regressions models, which can be viewed as the extension of Theorem 3.6 of Dümbgen et al. (2011).

There are several possible extensions for the FMR models in Chapter 4. Besides the one we've discussed in Chapter 5, one possible extension is a log-concave FMR model with a nonlinear mean part/deterministic part m_j:

$$y_i = m_j(\mathbf{x}_i) + e_{ij} \text{ with probability } \lambda_j(\mathbf{x}_i), \tag{6.1}$$

where the e_{ij}'s have a log-concave density g_j.

The mean part m_j could be estimated locally, either through spline based methods,

or local regression methods. Such models would provide more flexibility to characterize nonlinear relationships, such as the simulated data in Figure 6.1, in the mixture of regression models.

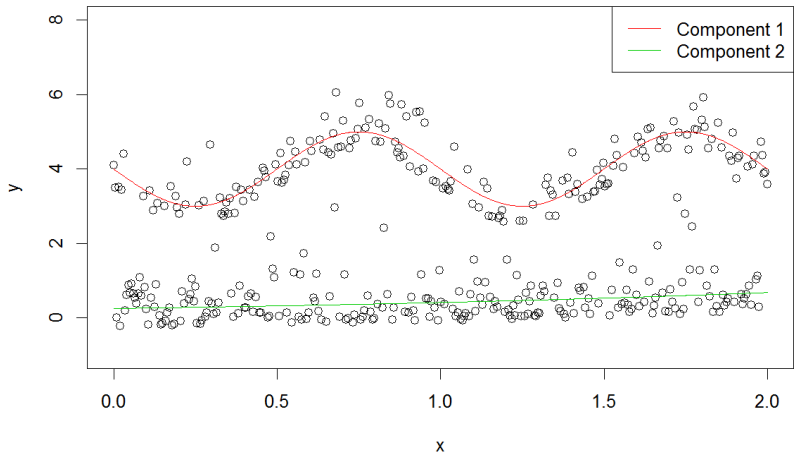


Figure 6.1: Simulated Data for log-concave FMR model with nonlinear mean parts.

Another possible extension is the log-concave FMR model with change points:

$$y_i = \mathbf{x}_i(\boldsymbol{\gamma}_j)^T\boldsymbol{\beta}_j + e_{ij} \text{ with probability } \lambda_j(\mathbf{x}_i), \tag{6.2}$$

where e_{ij}'s have a log-concave density function g_j, $\mathbf{x}_i(\boldsymbol{\gamma}_j) = (1, x_i, (x_i - \gamma_{1,j})I\{x_i > \gamma_{1,j}\}, \ldots, (x_i - \gamma_{c_j,j})I\{x_i > \gamma_{c_j,j}\})^T$, and $\gamma_{1,j}, \ldots, \gamma_{c_j,j}$ are the change points for component j.

Such models are useful when the data has some pattern change, such as the simulated data shown in Figure 6.2. A typical example would be a mixture of time series changes when some economic event occurs.

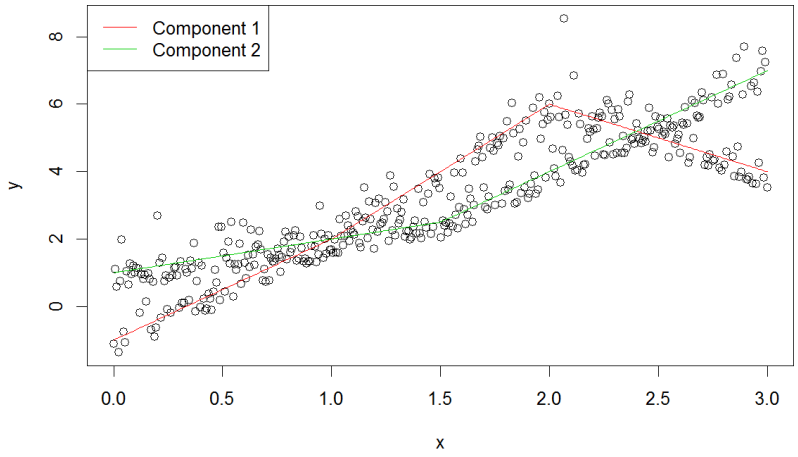


Figure 6.2: Simulated Data for log-concave FMR model with change points.

REFERENCES

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

Balabdaoui, F. and Doss, C. R. (2014). Inference for a mixture of symmetric distributions under log-concavity. *arXiv preprint arXiv:1411.4708*.

Balabdaoui, F., Rufibach, K., and Wellner, J. A. (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of statistics*, 37(3):1299.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

Bordes, L., Delmas, C., and Vandekerkhove, P. (2006a). Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian journal of statistics*, 33(4):733–752.

Bordes, L., Mottelet, S., Vandekerkhove, P., et al. (2006b). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics*, 34(3):1204–1232.

Bordes, L. and Vandekerkhove, P. (2010). Semiparametric two-component mixture model with a known component: an asymptotically normal estimator. *Mathematical Methods of Statistics*, 19(1):22–41.

Bozdogan, H. (1987). Model selection and akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Butucea, C. and Vandekerkhove, P. (2014). Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics*, 41(1):227–239.

Campbell, N. and Mahon, R. (1974). A multivariate study of variation in two species of rock crab of the genus leptograpsus. *Australian Journal of Zoology*, 22(3):417–425.

Chang, G. T. and Walther, G. (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis*, 51(12):6242–6251.

Chee, C.-S. and Wang, Y. (2013). Estimation of finite mixtures with symmetric components. *Statistics and Computing*, 23(2):233–249.

Chen, J., Tan, X., and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18(2):443.

Chen, Y. and Samworth, R. J. (2013). Smoothed log-concave maximum likelihood estimation with applications. *Statist. Sinica*, 23:1373–1398.

Cohen, E. (1980). Inharmonic tone perception. *Unpublished Ph. D. Dissertation, Stanford University*.

Cule, M., Gramacy, R., Samworth, R., et al. (2009). Logconcdead: An R package for maximum likelihood estimation of a multivariate log-concave density. *Journal of Statistical Software*, 29(2):1–20.

Cule, M. and Samworth, R. (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254–270.

Cule, M., Samworth, R., and Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(5):545–607.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Derek S, Y. (2007). A study of mixtures of regressions. *Unpublished Ph. D. Dissertation, Pennsylvania State University*.

Doss, C. and Wellner, J. A. (2013). Global rates of convergence of the mles of log-concave and s-concave densities. *arXiv preprint arXiv:1306.1438*.

Dümbgen, L. and Rufibach, K. (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40–68.

Dümbgen, L., Rufibach, K., et al. (2010). logcondens: Computations related to univariate log-concave density estimation. *Journal of Statistical Software, to appear*.

Dümbgen, L., Samworth, R., and Schuhmacher, D. (2011). Approximation by log-concave distributions, with applications to regression. *The Annals of Statistics*, 39(2):702–730.

Dümbgen, L., Samworth, R. J., Schuhmacher, D., et al. (2013). Stochastic search for semiparametric linear regression models. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 78–90. Institute of Mathematical Statistics.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volume 66. CRC Press.

Figueiredo, M. A. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96(453):194–209.

García-Escudero, L. A., Gordaliza, A., San Martin, R., Van Aelst, S., and Zamar, R. (2009). Robust linear clustering. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):301–318.

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

Grün, B. and Hornik, K. (2012). Modelling human immunodeficiency virus ribonucleic acid levels with finite mixtures for censored longitudinal data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(2):201–218.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, pages 795–800.

Hohmann, D. and Holzmann, H. (2013). Semiparametric location mixtures with distinct components. *Statistics*, 47(2):348–362.

Huang, M. (2009). Nonparametric techniques in finite mixture of regression models. *Unpublished Ph. D. Dissertation, Pennsylvania State University.*

Huang, M. and Yao, W. (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association*, 107(498):711–724.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hunter, D. R., Wang, S., and Hettmansperger, T. P. (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics*, pages 224–251.

Hunter, D. R. and Young, D. S. (2012). Semiparametric mixtures of regressions. *Journal of Nonparametric Statistics*, 24(1):19–38.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Jordan, M. I. and Xu, L. (1995). Convergence results for the em approach to mixtures of experts architectures. *Neural networks*, 8(9):1409–1431.

Kappel, F. and Kuntsevich, A. V. (2000). An implementation of shor's r-algorithm. *Computational Optimization and Applications*, 15(2):193–205.

Kim, A. K. and Samworth, R. J. (2014). Global rates of convergence in log-concave density estimation. *arXiv preprint arXiv:1404.2298*.

Kostantinos, N. (2000). Gaussian mixtures and their applications to signal processing. *Advanced Signal Processing Handbook: Theory and Implementation for Radar, Sonar, and Medical Imaging Real Time Systems*.

Liang, F. (2008). Clustering gene expression profiles using mixture model ensemble averaging approach. *JP J Biostat*, 2:57–80.

Lindsay, B. G. (1995). Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR.

Lindsay, B. G. et al. (1983a). The geometry of mixture likelihoods: a general theory. *The annals of statistics*, 11(1):86–94.

Lindsay, B. G. et al. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, 11(3):783–792.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4):633–648.

Ma, Y. and Yao, W. (2015). Flexible estimation of a semiparametric two-component mixture model with one parametric component. *Electronic Journal of Statistics*, 9:444–474.

McHugh, R. B. (1956). Efficient estimation and local identification in latent class analysis. *Psychometrika*, 21(4):331–347.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

Mcnicholas, P. D. and Murphy, T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Pal, J. K., Woodroofe, M., and Meyer, M. (2007). Estimating a polya frequency function. *Lecture Notes-Monograph Series*, pages 239–249.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *Journal of the American statistical association*, 67(338):306–310.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American statistical Association*, 73(364):730–738.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297.

Rufibach, K. (2007). Computing maximum likelihood estimators of a log-concave density function. *Journal of Statistical Computation and Simulation*, 77(7):561–574.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Smith, A. F. and Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Walther, G. (2002). Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97(458):508–513.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Wu, Q. and Yao, W. (2014). Mixtures of quantile regressions. *Computational Statistics & Data Analysis*.

Xiang, S., Yao, W., and Wu, J. (2014). Minimum profile hellinger distance estimation for a semiparametric mixture model. *Canadian Journal of Statistics*, 42(2):246–267.

Yao, W. (2010). A profile likelihood method for normal mixture with unequal variance. *Journal of Statistical Planning and Inference*, 140(7):2089–2098.

Yao, W. (2015). Label switching and its solutions for frequentist mixture models. *Journal of Statistical Computation and Simulation*, 85(5):1000–1012.

Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of the American Statistical Association*, 104(486).

Yao, W., Wei, Y., and Yu, C. (2014). Robust mixture regression using the t-distribution. *Computational Statistics & Data Analysis*, 71:116–127.

Young, D. S. and Hunter, D. R. (2010). Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266.