The Pennsylvania State University
The Graduate School
Department of Statistics

SEMIPARAMETRIC ESTIMATION FOR FINITE MIXTURE MODELS USING AN EXPONENTIAL TILT

A Dissertation in
Statistics
by
Tracey Ann-Wrobel Hammel

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2010

The dissertation of Tracey Ann-Wrobel Hammel was reviewed and approved* by the following:

Thomas P. Hettmansperger
Professor of Statistics
Dissertation Adviser
Chair of Committee

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

David R. Hunter
Associate Professor of Statistics

Hoben Thomas
Professor of Psychology

*Signatures are on file in the Graduate School.

Abstract

The main purpose of this thesis is to introduce methodology to fit finite mixture models. We propose a method where the component distributions are fitted using an exponential tilt model, in which the log ratio of the density functions of the components is modeled as a quadratic function of the observations. This approach has two key advantages. First, except for the exponential tilt assumption, the marginal distributions of the observations can be completely arbitrary. Second, we have the advantage of having a likelihood. In the multivariate case, a model selection method is introduced for identifying the number of component distributions in the mixture model when theory does not suggest it. In addition, a likelihood ratio test, using the profile likelihood, is used to choose between a model that assumes the coordinates are conditionally independent and one that has *blocks* of conditionally i.i.d. coordinates. Simulations are provided to show the performance of the methods in estimating the component means and standard deviations. The proposed methods are applied to real datasets.

Table of Contents

List of Tables

List of Figures

Acknowledgments

First and foremost, I would like to thank my advisor, Tom. His patience with me during these last couple of years has meant so much to me. His encouragement and input have been so essential. I want to thank him for making me his last student.

I would like to thank my committee members: Bruce Lindsay, David Hunter, and Hoben Thomas. Their comments and suggestions were greatly appreciated.

Most importantly, I want to thank my family. My husband, Ben; my rock. He has been the most patient and understanding person I have ever met. His support and encouragement have been essential. I would have never finished this without him. I love you so much, Ben. I want to thank my Father, Ron, for always having faith and confidence in me. He always gave me encouragement when I needed it the most. I want to thank my sister, Lindsay, for always listening and being there for me. To my sister, Stacey: Thank you for being the best friend I could ever want or could ever wish for. I am where I am and who I am because of her. And last, but certainly not least, I want to thank my Mother, Debbie, for always making me feel that she was proud of me no matter what. I wish she could have been here to see me finally finish this.

I want to thank my amazing friends, Ruth and Tatiana. These two beautiful women gave me a new meaning of family. Without the two of them, I would not be here and I certainly would not have had as much fun in grad school. They were there for me in the best of times and the worst of times. I could not ask for better friends. I will love them always.

Lastly, I want to thank all the friends I have met along the way: Benilton, Eduardo, Lourdes, Elizabeth, Heather, Megan, Chuck, and Christian. Each of them has touched my heart and left an unforgettable imprint.

Chapter 1

Introduction

1.1 Finite Mixture Models

There are many applications where the interest is to classify n observations into m groups based on k measures on each observation. For example, consider a study in which children are given a series of cognitive tasks. Suppose it is suspected that different children employ different cognitive strategies to solve the tasks. Observation does not reveal the strategy a child might employ, and children are not able to articulate how they solve the tasks. Still it may be possible to identify from the data the strategies employed, and which strategies individual children employ. To do so interest focusses on classifying children into homogenous groups, each group hopefully representing a different cognitive strategy. The classification is based only on information in the measurements recorded. Hettmansperger and Thomas (2000) and Cruz-Medina, Hettmansperger, and Thomas (2004) used mixture models to achieve this goal. Mixture models allow one to classify subjects into different groups, probabilistically, based on the task measurements. Using mixture models can provide estimates of the proportion size of each group and they can also be used to estimate the densities of each group, from which we can obtain group characteristic such as means and standard deviations.

Mixture models have been studied extensively. At times, theory suggests the shape(s) or distribution(s) of the group (or component) densities. In these cases, methods are available to perform the mixture analysis (see McLachlan and Peel, 2000). These methods are known as parametric methods. In other cases however, theory does not suggest the form of the component densities. When the densities are not specified, nonparametric methods are available (see Benaglia et al., 2009a). Choosing an incorrect parametric form may lead to incorrect results. Choosing a nonparametric model to analyze the data will be robust but also loses some of the advantages of assuming a

parametric form. In this thesis, we develop a semiparametric method. The methods presented in this thesis have the advantage of being robust but also has some of the advantages of choosing a parametric form, such as having a likelihood. The methods presented here are based on the exponential tilt model. We describe, in more detail, the exponential tilt model in Section 1.4.

The multivariate method proposed in this thesis relaxes the assumptions of previous work (see Section 1.2 for more details). Previous work assumed that the measurements, conditioning on the component (or group) membership, are independent and identically distributed (i.i.d). In practice, however, the measurements may not be identically distributed. Our semiparametric method requires the measurements to be conditionally independent (i.e. conditioning on component membership, the measurements are independent). Relaxing this assumption is advantageous not only for interpretive purposes but the model also allows for measurements that are not similar. For example, a study might record different measurements from the blood of subjects. In this case, the i.i.d. assumption may not be applicable. The independence assumption also may not seem valid but as we discuss in Section 1.2, it is essential for identifiability of the parameters in the model.

Mixture models are composed of a specified number of components, m, and they take the following form:

$$g_{\boldsymbol{\phi}}(X) = \sum_{l=1}^{m} \lambda_l \psi_l(x) \tag{1.1}$$

with $\sum_{l=1}^{m} \lambda_l = 1$ and $l = 1, \ldots, m$, where $\lambda_l \geq 0$ are the mixing proportions, $X \in \mathbb{R}^k$, and $\psi(\cdot)$ belongs to a family of distributions $\mathcal{F} = \{f(\cdot|\nu), \nu \in \mathbb{R}^p\}$ indexed by the p dimensional parameter ν. The parameters to be estimated are $\boldsymbol{\phi}^t = (\lambda_1, \cdots, \lambda_m, \psi_1, \cdots, \psi_m)$. When the component densities are assumed to belong to a parametric family, the parameters, $\boldsymbol{\phi}$, are found using methods described in Titterington et al. (1985), Lindsay (1995) and McLachlan and Peel (2000).

In this thesis, we present two methods for estimating the component densities in the mixture (as well as the component means and standard deviations). We begin by defining and discussing the identifiability in the mixture model and continue with

previous work for both the univariate and multivariate case. Chapter 2 shows in detail the semiparametric method we propose for univariate mixture models. The method is an extension of the density estimation method proposed by Efron and Tibshirani (1996). In that chapter, the first two sections present their idea and the next two sections show how we extended the method to mixtures. We then provide simulation results for various models to show how the method performs. Finally, we use the method to analyze real data collected from the Old Faithful Geyser in Wyoming.

In Chapter 3, we present a semiparametric method for multivariate mixture models based on the exponential tilt. To find the estimates of the parameters we use the EM algorithm to find the maximum likelihood estimates (mle); see Dempster et al. (1977). The details for the algorithm as well as simulations are provided. We present likelihood ratio tests using the profile likelihood. Also, we use a form of the Bayesian Information Criterion (BIC, Schwartz (1978)) to help determine the number of components if theory does not suggest it. We end the chapter with two real data examples; the Reaction Time (RT) data and the Water Level data as discussed in Benaglia et al. (2008).

1.2 Identifiability

In this section, we present identifiability results based on previous work for both univariate and multivariate mixture models. Model (1.1) is not identifiable unless there are assumptions about $\psi_1, \cdots, \psi_m$. The term *identifiable* means that $g_{\boldsymbol{\phi}}$ has a *unique* representation. That is $g_{\boldsymbol{\phi}_1} = g_{\boldsymbol{\phi}_2}$ only if $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$ for all $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$ in the parameters space. In this thesis, we do not consider *label-switching*, the $m!$ permutations of $(\lambda_1, \psi_1), \cdots, (\lambda_m, \psi_m)$, to be a distinct representation.

1.2.1 Univariate case

Suppose the variables $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables from a m component univariate mixture model. The probability density function (pdf) would be of the form:

$$g(x) = \sum_{l=1}^{m} \lambda_l f_l(x) \tag{1.2}$$

where $\sum_{l=1}^{m} \lambda_l = 1$ and $f_l(\cdot)$ is a probability density function. Unless restrictions are imposed on the component densities, f_l, the parameters in the model is not identifiable. It is of interest to determine how strict these restrictions need to be in order to guarantee identifiability. We will present previous work with results on identifiability that are relevant to our model (i.e, the univariate mixture model).

Bordes et al. (2006b) present a nonparametric stochastic method for estimating the densities of a univariate mixture model. The model is not purely nonparametric. If it were, the parameterization of the model would not be identifiable. They find that if the model contained location-shifted symmetric component densities, then the parameters in the model would be identifiable for $m \leq 3$ expect in certain special cases. Also, when $m = 2$, they found the parameters in the model were not identifiable when $\lambda = 1/2$. These are the only restrictions imposed on the model. Hunter et al. (2007) found similar results.

The univariate method we develop in this thesis (see Chapter 2), does not assume the component densities are symmetric. We have not been able to determine the conditions under which the parameters in the univariate model are identifiable. Even though it is not proved, we do show that the estimates produced using this method are close to the true values. The method also produces good estimates when the exponential tilt assumption is not valid, as justified by the simulations in Section 2.7.

1.2.2 Multivariate case

Suppose the variables $X_1, \cdots, X_n$ are independent and identically distributed (i.i.d.) random variables from a m component multivariate mixture model with k coordinates (or repeated measure) and with $X_i' = (x_{i1}, \cdots, x_{ik})$, for $i = 1, \cdots, n$. The probability density function (pdf) would be of the form:

$$g(X) = \sum_{l=1}^{m} \lambda_l f_l(x_1, \cdots, x_k) \tag{1.3}$$

where $\sum_{l=1}^{m} \lambda_l = 1$ and $f_l(\cdot)$ is a probability density function. Unless restrictions are imposed on the component densities, f_l, the parameters in model are not identifiable. It is of interest to determine how strict these restrictions need to be in order to guarantee identifiability. Identifiability is important when estimating the underlying mixture structure. For earlier results on identifiability in nonparametric mixtures see Hall and Zhou (2003); Hall et al. (2005); Elmore et al. (2005). The most important assumption used for identifiability is, conditional on the component membership of the observation, the densities are independent. Under this assumption,

$$g(X) = \sum_{l=1}^{m} \lambda_l \prod_{j=1}^{k} f_j(x_j) \tag{1.4}$$

where m is the number of components and k is the number of repeated measures. When the component densities, $f_{lj}(\cdot)$, do not depend on j for $j = 1, \cdots, k$, the repeated

measure, the densities would be conditionally independent and identically distributed (i.i.d.).

The model we present in Chapter 3 requires the conditional independence assumption. Our model has the following form:

$$\begin{aligned} h(X) =& \lambda_1 \prod_{j=1}^{k} f_{1j}(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} f_j(x_j) \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\} \\ =& \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\}\right] \prod_{j=1}^{k} f_j(x_j). \end{aligned} \tag{1.5}$$

Theorem 8 of Allman, Matias, and Rhodes (2009) states that the parameters in the mixture of the form (1.4) are uniquely identifiable up to label switching provided that $k \geq 3$ and, for each $j = 1, \ldots, k$, the m distributions $\{f_{lj}\}_{1 \leq l \leq m}$ are linearly independent. This result makes sense since linear independence precludes expressing any one of the coordinate distributions as a linear combination of the other $m - 1$ distributions. Since, in our case, in an m component mixture, $\sum \lambda_l = 1$ and $\lambda_l > 0$, and for each $j = 1, \ldots, k$

$$\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\} \neq 0 \quad \text{for} \quad -\infty < x_j < \infty$$

identifiability follows for the parameters in the model (1.5) under the linear independence of $\{f_j\}$.

1.3 Previous Work

1.3.1 Previous Work for the Univariate Case

If the component densities of Model (1.2) are completely specified, there are methods to handle these models. Presentation of these methods can be found in Lindsay (1995) and McLachlan and Peel (2000). If the component densities in the mixture model are not specified, more robust methods exist. They, however, need certain distributional assumptions about the component densities as well as assumptions about the mixture.

There are several methods for estimating the attributes of the component densities with the assumption that the component densities are symmetric and differ only by location. Cruz-Medina and Hettmansperger (2004) is an example of such a method. This paper uses tetranomials and sextinomials to estimate the mixing proportions as well as a measure of location and spread for the component densities. At that time, their method was limited to the two component mixture due to computational difficulties when the model has more than two components.

Bordes et al. (2006a) present a semiparametric method for estimating the component densities for a univariate mixture model with a fixed but arbitrary number of components. This method has an advantage over Cruz-Medina and Hettmansperger (2004) because the calculations are easy for a mixture model with more than two components. Hunter et al. (2007) presented a method for estimating the parameters and mixing proportions for two or three component location-shifted mixtures when the component densities are symmetric. Bordes et al. (2006a) use a nonparametric kernel density approach for estimating the densities.

1.3.2 Previous Work for the Multivariate Case

As in the univariate case, some parametric multivariate mixture models have been studied (see, *e.g.*, Titterington, Smith, and Makov, 1985; Lindsay, 1995; McLachlan and Peel, 2000). Early work on nonparametric models assumed the repeated measures were conditionally i.i.d. In the method presented in this thesis the repeated measures are assumed conditionally independent but not necessarily identically distributed. The conditionally i.i.d. assumption may be too strict and relaxing the assumption to conditional independence gives an advantage. Previous work that assume the repeated measures are conditionally i.i.d include Cruz-Medina, Hettmansperger, and Thomas (2004) and Elmore (2003). Their method uses discretization of the data to yield multinomial mixtures. Our method does not require discretization of the data.

Previous work that used the conditional independence assumption include Hettmansperger and Thomas (2000). In the Hettmansperger and Thomas (2000) method, a cut-point is defined on the original measurement scale. Each measurement

is then discretized into a binary outcome, depending on whether the measurement is larger than or smaller than the predefined cut-point. The binary outcomes in an observation are summed and the resulting data is analyzed as a binomial mixture model. This method avoids the need to make strong distributional assumptions for the original data. However, it has a number of shortfalls. First, an arbitrary cut-point has to be defined. Second, the original data has to be discretized, which may lead to a loss of information. Third, the measures must be identically distributed. Cruz-Medina et al. (2004) extended the Hettmansperger and Thomas (2000) method to using multiple cut-points and the resulting data is analyzed as multinomial mixture data. However, the method has the same constraints as the Hettmansperger and Thomas (2000) method.

Benaglia et al. (2009a) present a nonparametric method to estimate the component densities of a multivariate mixture model with the assumption that the repeated measures are conditionally independent. Their method uses kernel densities estimates with the weights being the posterior probabilities. The method has the advantage of being robust, but it does not have the advantage of a likelihood as in the method presented here. They also have to determine bandwidths for the kernel density estimates. Choosing a bandwidth is not always an easy take and different bandwidths may lead to different results. Our method is based on the empirical CDF and does not depend on the bandwidth.

1.4 Exponential tilt model

In this section we will discuss the exponential tilt model. Field and Ronchetti (1990) use an exponent term to sharpen the estimate of sampling distributions, such as the mean. Using the exponent term to obtain a better estimate of the density became know as exponential tilting. They show that using the tilting greatly improved the estimate of the density. Anderson (1979) was the first to use the exponential tilt (with a linear exponent) to estimate the components in a mixture.

The methods proposed in this thesis require the densities to be related by an exponent tilt. Densities functions that are related by an exponential tilt are such that one distribution, $g(x)$, is related to $f(x)$ by $g(x) = f(x)\exp\left(\beta_0 + \beta_1 x + \ldots + \beta_p x^p\right)$.

There are several advantages to the exponential tilt model. Exponential tilt models are flexible. Kay and Little (1986) discuss various versions of the exponential tilt and relate them to common distributions. We restrict the multivariate model in Chapter 3 to have a quadratic exponent, mainly for computations and in Section 1.4.1 we show how normal densities and gamma densities are related by this tilt. Densities related by an exponential tilt can be skewed and Qin et al. (2002) showed satisfactory results for skewed data.

The exponent tilt model can be easily adapted in the mixture setting. Consider a two-component univariate mixture model:

$$g(x) = \lambda f(x) \exp\{\beta_{10} + \beta_{11} x + \ldots + \beta_{1p} x^p\} + (1-\lambda) f(x) \exp\{\beta_{20} + \beta_{21} x + \ldots + \beta_{2p} x^p\} \quad (1.6)$$

The exponential tilt model is a compromise between parametric and nonparametric density estimates (Efron and Tibshirani, 1996). Choosing the parametric family for the component densities in the mixture may be a difficult task. At times, there is theory that suggests the particular distributions but more often this is not the case. Choosing the incorrect component densities may lead to incorrect estimates. Choosing a nonparametric approach will relieve this problem, but also loses the advantage of having a likelihood. The methods based on the tilt estimate the baseline density, $f(x)$, nonparametrically and correct this estimate by the exponential tilt. They combine the benefits of choosing a nonparametric method by producing more robust estimates and the benefit of having a likelihood when choosing a parametric method. Having a likelihood is an advantage because it facilitates model selection.

Efron and Tibshirani (1996) show a moment matching property that results from using an exponential tilt model in density estimation. We discuss the details of the moment matching property in Section 2.2.4. In their method, they use a nonparametric estimator followed by a correction to match predetermined moments. They also claim that the moment matching may reduce the bias of the nonparametric density estimate.

We believe that using an exponential tilt in the mixture setting would be advantageous. Since theory often does not suggest the shape of the component densities, it is better to choose a more robust method. It is also possible that the underlying component

densities are not symmetric. The exponential tilt models are flexible and provide good estimates for skewed data. Although a nonparametric approach is more robust than a semiparametric method, the latter has the advantage of using a likelihood. Likelihoods are important in helping with model selection. Suppose, for example, that theory suggests that subjects could be grouped according to some characteristic, but does not suggest the number of groups. Using the likelihood could be used to help determine the number of components from the data. It might also make sense that some of the measures are similar (i.i.d perhaps) while others are different, but independent. We develop tools in this thesis to help test this using the exponential tilt model. There are no other semiparametric or nonparametric methods available that suggest this capability. Overall, the exponential tilt model seems like a good compromise between a parametric and nonparametric approach.

1.4.1 Common distributions

We assume that the component densities are related by the exponential tilt model of Anderson (1979) (sometimes called the density-ratio model). For example, suppose we have a two component mixture with densities f and g. The exponential tilt model with a linear exponent assumes that f and g are related by $\frac{g(x)}{f(x)} = \exp(\alpha+\beta x)$. Kay and Little (1986) discussed various versions of the density-ratio model for some conventional distributions. For example, if f and g are normal density functions with different means and variances, then a quadratic term is needed in the density-ratio model. To show this relationship more clearly, suppose f is a normal density function with mean μ and variance σ^2. Then,

$$g(x) = f(x)\exp\{\alpha+\beta x+\gamma x^2\} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\exp\{\alpha+\beta x+\gamma x^2\}$$

$$= \frac{\exp(\alpha)}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\mu^2}{2\sigma^2}+\frac{(\mu+\sigma^2\beta)^2}{2\sigma^2(1-2\sigma^2\gamma)^2}\right\}\exp\left\{-\frac{1}{2\left(\frac{\sigma^2}{1-2\sigma^2\gamma}\right)}\left(x-\frac{\mu+\beta\sigma^2}{1-2\sigma^2\gamma}\right)^2\right\}$$

Let $\mu^* = \frac{\mu+\beta\sigma^2}{1-2\sigma^2\gamma}$ and $\sigma^{*2} = \frac{\sigma^2}{1-2\sigma^2\gamma}$. If $\exp(\alpha) = \frac{\sqrt{1-2\sigma^2\gamma}}{\exp\left\{\frac{-\mu^2(1-2\sigma^2\gamma)^2+(\mu+\sigma^2\beta)^2}{2\sigma^2}\right\}}$, then $g(x)$ would be a normal density function with mean μ^* and variance σ^{*2}.

The point of interest here is that we can solve for the parameters in the exponential tilt in terms of the "old mean", μ, the "old variance", σ^2, the "new mean", μ^*, and the "new variance", σ^{*2}. Thus,

$$\gamma = \frac{\sigma^{*2}-\sigma^2}{2\sigma^2\sigma^{*2}} \tag{1.7}$$

$$\beta = \frac{\mu^*-2\sigma^2\gamma\mu-\mu}{\sigma^2} \tag{1.8}$$

$$\alpha = -\frac{1}{2}\log(1-2\sigma^2\gamma) + \frac{\mu^2(1-2\sigma^2\gamma)^2+(\mu+\sigma^2\beta)^2}{2\sigma^2} \tag{1.9}$$

Therefore, if we start with a normal distribution $f(x)$, with mean μ and variance σ^2 and wanted a normal distribution, $g(x)$, with mean μ^* and variance σ^{*2}, we find $\alpha, \beta,$ and γ using the equations above and calculate $g(x) = f(x)\exp\{\alpha+\beta x+\gamma x^2\}$.

The exponential tilt model fits for other distributions as well. Suppose $h(x)$ is a gamma distribution with parameters k and θ. If we multiply $h(x)$ by an exponential tilt containing only the linear term, we obtain:

$$\begin{aligned} h(x)\exp(\alpha+\beta x) &= \left(\frac{1}{\Gamma(k)\theta^k}x^{k-1}e^{-x/\theta}\right)\exp(\alpha+\beta x) = \left(\frac{e^\alpha}{\Gamma(k)\theta^k}\right)x^{k-1}\exp\left(-\frac{x}{\frac{\theta}{1-\theta\beta}}\right) \\ &= \frac{e^\alpha}{\Gamma(k)\theta^k}x^{k-1}\exp\left(-x/\theta^*\right) \end{aligned}$$

Since α is the normalizing term, if we set $\alpha = \log\left(\frac{\theta^k}{(\theta^*)^k}\right)$ we would have a gamma distribution with parameters k and $\theta^* = \frac{\theta}{1-\theta\beta}$. Thus the exponential tilt model is valid for gamma distribution components with a common shape parameter. Similar to the normal example above, if you begin with a gamma distribution, $h(x)$, with parameters k and θ and would like to obtain a gamma distribution, $g(x)$, with parameters k and θ^*, find α and $\beta = \frac{\theta-\theta^*}{\theta\theta^*}$, and calculate $h(x)\exp\{\alpha+\beta x\}$.

The gamma mixture simulations that we will present in this thesis are those with different shape parameters, i.e. do not satisfy the exponential tilt assumption. Even though the assumption is not valid, we show that the models estimate the component means and standard errors fairly well.

Chapter 2

Semiparametric Density Estimation for Univariate Mixtures

2.1 Introduction

In this chapter, we will describe the density estimation method proposed in Efron and Tibshirani (1996). Their method combines both maximum likelihood fitting and nonparametric methods. It has the advantage of likelihood theory and the robustness of nonparametric methods. We extend their method to univariate mixtures hoping to construct a method that will have the same advantages. Section 2.2 describes the univariate, non-mixture case as in Efron and Tibshirani (1996). Next, we will show how to extended this method to the univariate finite mixture model. We include how the EM algorithm is used to find estimates of the model parameters. We will finish the chapter by using simulations to judge performance of the method. Lastly, we use the method to estimate the component densities of the data collected on the Old Faithful Geyser in Yellowstone National Park, Wyoming.

2.2 The non-mixture case

An outline of the Section is as follows: In Section 2.2.1, we present the method for estimating the density for the univariate non-mixture model. In Section 2.2.2, there is a brief summary of the steps for the density estimation. Then, we will show the theory behind the method. Finally, we will end this section with a few examples.

2.2.1 Semiparametric density estimation for univariate data

In this section, we give a brief overview of the method proposed by Efron and Tibshirani (1996). The idea behind this density estimation method stems from those proposed by Lindsey (1974). Suppose that we observe a random sample, $y_1, \cdots, y_n$, and

wish to estimate the probability density $g(y)$ from which the data came. In other words, we assume

$$y_j \overset{\text{i.i.d}}{\sim} g(y), \quad j = 1, \cdots, n, \tag{2.1}$$

where $g(y)$ is unknown. In their paper, Efron and Tibshirani (1996) propose the following model for estimating the density,

$$g_\beta(y) = g_0(y) \exp\left\{\beta_0 + t(y)'\boldsymbol{\beta}_1\right\} \tag{2.2}$$

and call it the exponential family through $g_0(y)$ with sufficient statistics $t(y)$. In (2.2), $g_0(y)$ is the carrier density, $t(y)$ is a $p \times 1$ vector of sufficient statistics, $\boldsymbol{\beta}_1$ is a $p \times 1$ parameter vector, and β_0 is the normalizing parameter. In this model, $\boldsymbol{\beta}_1$ is the unknown vector to be estimated.

Suppose $y_1, \cdots, y_n$ are independent and identically distributed observations from $g_\beta(y)$ and suppose that $\left\{g_\beta(y),\ \beta \in \Theta\right\}$ is a family of densities on $\mathcal{Y}$. Begin by discretizing the data by partitioning the sample space, $\mathcal{Y}$, into k disjoint cells. Let $\mathcal{Y}_i$ be such that

$$\mathcal{Y} = \bigcup_{i=1}^{k} \mathcal{Y}_i.$$

The data are then reduced to cell counts, $s_i = \#\left\{y_j \in \mathcal{Y}_i\right\}$, such that $\sum_{i=1}^{k} s_i = n$, for $i = 1, \ldots, k$ and $j = 1, \ldots, n$. Then the probability of observing y in the i-th cell is

$$\pi_i(\beta) = \int_{\mathcal{Y}_i} g_\beta(y) dy,$$

with $\sum_{i=1}^{k} \pi_i(\beta) = 1$. In this case, $\mathbf{s}' = (s_1, \ldots, s_k)$ can be considered a Multinomial random vector and parameters n and $\boldsymbol{\pi}(\beta)$, where $\boldsymbol{\pi}(\beta)' = \left(\pi_1(\beta), \cdots, \pi_k(\beta)\right)$.

Now, consider the density estimation problem. The family of densities in this case is $\left\{g_{\boldsymbol{\beta}}(y),\ \boldsymbol{\beta} \in \Theta\right\}$. The special exponential family is

$$g_{\boldsymbol{\beta}}(y) = g_0(y) \exp\left\{t'(y)\boldsymbol{\beta}\right\}$$

with $t'(y) = (1, y, y^2, \cdots, y^p)$ and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of parameters, i.e. $\boldsymbol{\beta}' = (\beta_0, \boldsymbol{\beta}_1')$. Efron and Tibshirani (1996) call $g_0(y)$ the *carrier density*. The carrier density can be estimate using a kernel density estimate. Although it is calculated from the data it is treated as fix and known. We want to estimate

$$\pi_i(\boldsymbol{\beta}) = \int_{\mathcal{Y}_i} g_0(y) \exp\left\{t_i'\boldsymbol{\beta}\right\} dy = \int_{\mathcal{Y}_i} g_{\boldsymbol{\beta}}(y) dy.$$

If we estimate the probability at a convenient point in $\mathcal{Y}_i$, say the midpoint, $\tilde{y}_i$, then we can estimate $\pi_i(\boldsymbol{\beta})$ by

$$\begin{aligned}\pi_i(\boldsymbol{\beta}) = \int_{\mathcal{Y}_i} g_0(y) \exp\left\{t'(y)\boldsymbol{\beta}\right\} dy &\approx \int_{\mathcal{Y}_i} g_0(y) \exp\left\{t_i'\boldsymbol{\beta}\right\} dy \\ &= \exp\left\{t_i'\boldsymbol{\beta}\right\} \int_{\mathcal{Y}_i} g_0(y) dy\end{aligned}$$

where $t_i' = \left(1, \tilde{y}_i, \tilde{y}_i^2, \ldots, \tilde{y}_i^p\right)$. Let $\pi_i^* = \int_{\mathcal{Y}_i} g_0(y) dy$ and let the discretized version of π_i^* be $\pi_i^0 = \Delta_i g_0(\tilde{y}_i)$, where Δ_i is the width of the interval $\mathcal{Y}_i$. This implies

$$\hat{\pi}_i(\boldsymbol{\beta}) = \exp\left\{t_i'\boldsymbol{\beta}\right\} \pi_i^0.$$

In this thesis we assume that the widths of the intervals, Δ_i, are the same so we simply write Δ. The method can be adapted for unequal interval widths.

As suggested in Lindsey (1974), instead of considering the distribution of $\mathbf{s}$ as being a Multinomial, we can treat s_i, $i = 1, \cdots, k$, as independent Poisson random

variables with

$$s_i \overset{\text{ind}}{\sim} Pois\left(\mu_i(\gamma,\beta)\right)$$

where $\mu_i(\gamma,\beta) = \gamma\pi_i(\beta)$. In this case, we can consider γ to be n. Therefore, $s_i \sim Pois\left(n\pi_i(\theta)\right)$. And this, in turn, implies that

$$\hat{\mu}_i(\gamma,\widehat{\boldsymbol{\beta}}) = n\Delta g_0(\tilde{y}_i)\exp\left\{t_i'\widehat{\boldsymbol{\beta}}\right\}. \tag{2.3}$$

In summary, Efron and Tibshirani (1996) begin with considering the cell counts to be Multinomial and instead treating the counts as independent Poisson random variables. In the next section, we will show that considering either model will lead to the same estimates of the exponential tilt parameters, $\boldsymbol{\beta}$.

2.2.2 The computation method

We want to estimate the parameters from the following model,

$$g_\beta(y) = g_0(y)\exp\left\{\beta_0 + t(y)'\boldsymbol{\beta}_1\right\}. \tag{2.4}$$

The following are the steps to estimate the density. The outline of this density estimation method were presented in Section 2.2.1 and the theory is in Section 2.2.3.

1. Discretize the data. Suppose that we let $\mathcal{Y}$ be the sample space of the data. Partition $\mathcal{Y}$ into k disjoint cells, $\mathcal{Y}_i$ for $j = 1,\ldots,k$, i.e $\mathcal{Y} = \bigcup_{i=1}^{k}\mathcal{Y}_i$. Let

$$s_i = \#\left\{y_j \in \mathcal{Y}_i\right\}, \text{ for } i = 1,\cdots,k \text{ and } j = 1,\ldots,n$$

   Thus, the data are reduced to cell counts. Denote the midpoints of $\mathcal{Y}_i$ as $\tilde{y}_i$ and the vector of cell counts to be $\mathbf{s}$.

2. Calculate the estimated carrier vector and denote it as $\hat{\boldsymbol{\mu}}^0$. Efron and Tibshirani (1996) suggest calculating the estimate using a smoothing matrix, $\hat{\boldsymbol{\mu}}^0 = \mathbf{M}(h)\mathbf{s}$

and took $\mathbf{M}(h)$ to be a normal kernel smoother with

$$\mathbf{M}_{k,j}(h) = \frac{c_k}{h}\Phi\left(\frac{\tilde{y}_k - \tilde{y}_j}{h}\right) \tag{2.5}$$

The constants c_k were chosen so that $\mathbf{M}_{k+} = 1$, $\Phi(\cdot)$ is the standard normal density function, and h is the bandwidth.

3. Find the discrete analogs of the special exponential family, say $\hat{\boldsymbol{\mu}}' = (\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_k)$, where

$$\hat{\mu}_i = \hat{\mu}_i^0 \exp\left(\hat{\beta}_0 + t_i'\hat{\beta}_1\right)$$

and $t_i' = \left(\tilde{y}_i, \cdots, \tilde{y}_i^p\right)$ where p is the number of moments you would like to match; see Section 2.2.4 for an explanation of the moment matching property. The estimate for $\boldsymbol{\beta}$ is calculated using Poisson regression for counts, s_i, against the vector $\mathbf{t}_i$, using $\log\left(\hat{\mu}^0\right)$ as an offset.

4. Plot the estimated function $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}^0 \exp\left\{\hat{\beta}_0 + t_i'\hat{\beta}_1\right\}$. Note that this is **not** a density. To plot the estimated density, $g_{\widehat{\boldsymbol{\beta}}}(\tilde{y}_i)$, plot the midpoints, $\tilde{y}_i$, against $\frac{\hat{\mu}_i}{n\Delta}$, where Δ is the length of the cells. Recall, we assume that each of the cells, $\mathcal{Y}_i$, has the same width for $i = 1, \ldots, k$.

We have provided a brief overview of the density estimation method. In the next section, we will show the mathematical details of this method. We show how considering the cell counts to be a Multinomial random variable is the same as considering the counts to be from independent Poisson distributions and how they use Poisson regression to calculate estimates of the unknown exponential tilt parameters, $\boldsymbol{\beta}$.

2.2.3 The Theoretical Details

In this section, we will show, in detail, how to estimate the density using the method proposed in Efron and Tibshirani (1996). We will prove that finding the maximum likelihood estimate for the exponential tilt parameters in the Multinomial setting is the same as finding those using Poisson regression.

2.2.3.1 The Multinomial

Suppose that we observe a Multinomial random vector $\mathbf{s}' = (s_1, \ldots, s_k)$ such that

$$\mathbf{s} \sim Mult(n, \boldsymbol{\pi})$$

where $\boldsymbol{\pi}' = (\pi_1, \cdots, \pi_k)$ with

$$\pi_i = \Delta g_0(\tilde{y}_i) \exp\left\{\mathbf{X}_i' \boldsymbol{\beta}\right\}, \tag{2.6}$$

$\mathbf{X}_i' = \left(1, t_i'\right)$ with $t_i' = (\tilde{y}_i, \cdots, \tilde{y}_i^p)$ and the $\tilde{y}_i$'s are known, $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$, p is the order of the exponential tilt, and $\sum_{i=1}^k s_i = n$, for $i = 1, \ldots, k$. The likelihood and the log likelihood given the data are

$$L(\boldsymbol{\pi}) = \frac{n!}{s_1! s_2! \cdots s_k!} \prod_{i=1}^k \pi_i^{s_i} \tag{2.7}$$

and

$$\ell(\boldsymbol{\pi}) = \log(n!) - \sum_{i=1}^k \log(s_i!) + \sum_{i=1}^k s_i \log(\pi_i), \tag{2.8}$$

respectively.

Substituting (2.6) in (2.8) and including a Lagrange multiplier for the constraint, $\sum_{i=1}^{k} \pi_i = 1$, gives the following log likelihood

$$\begin{aligned}
\ell(\boldsymbol{\pi}) &\propto \sum_{i=1}^{k} s_i \log\left(\pi_i\right) + \eta\left(\sum_{i=1}^{k} \pi_i - 1\right) \\
&= \sum_{i=1}^{k} s_i \log\left(\Delta g_0(\tilde{y}_i) \exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\}\right) + \eta\left(\sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\} - 1\right) \\
&\propto \sum_{i=1}^{k} s_i \mathbf{X}_i'\boldsymbol{\beta} + \eta \sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\left\{\mathbf{X}_i'\boldsymbol{\beta}\right\} - \eta \qquad (2.9)
\end{aligned}$$

To find the maximum likelihood estimator for $\boldsymbol{\beta}$, we must take the gradient with respect to $\boldsymbol{\beta}$ and η and solve the system of equations. The gradient with respect to $\boldsymbol{\beta}$ yields

$$\bigtriangledown_{\boldsymbol{\beta}} \ell(\pi) = \sum_{i=1}^{k} s_i \mathbf{X}_i + \eta \sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}\mathbf{X}_i \qquad (2.10)$$

and the derivative with respect to η is

$$\frac{\partial \ell}{\partial \eta} = \sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i'\boldsymbol{\beta}\} - 1 \qquad (2.11)$$

The derivative with respect to β_0 is

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{i=1}^{k} s_i + \eta \sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}. \qquad (2.12)$$

Setting (2.10), (2.11), and (2.12) equal to zero yields the following equations,

$$\begin{aligned}
&\sum_{i=1}^{k} s_i \mathbf{X}_i = n \sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i'\boldsymbol{\beta}\}\mathbf{X}_i \\
&\Rightarrow \sum_{i=1}^{k} s_i \mathbf{X}_i = \sum_{i=1}^{k} \exp\left\{\log\left(n\Delta g_0(\tilde{y}_i)\right) + \mathbf{X}_i'\boldsymbol{\beta}\right\}\mathbf{X}_i. \qquad (2.13)
\end{aligned}$$

Therefore, the maximum likelihood estimate for $\boldsymbol{\beta}$ is the solution to equation (2.13), denoted $\widehat{\boldsymbol{\beta}}$. We can *trick* computers into finding this solution easily by using generalized linear model software. Recognize that the solution to (2.13) is the same as the solution found by using Poisson regression with $\log\left(n\Delta g_{0}(\tilde{y}_{i})\right)$ as the offset. This offset is used to adjust for the values of the carrier density for each of the bins.

2.2.3.2 The Poisson Regression

In this section, we are going to prove that the maximum likelihood estimator for $\boldsymbol{\beta}$ is the same whether you consider the Multinomial case or the Poisson case; see Section 2.2.1. Now suppose that $s_{1}, \cdots, s_{k}$ are independent Poisson random variables such that

$$\mathbb{E}(s_{i}) \doteq n\Delta g_{0}(\tilde{y}_{i}) \exp\{\mathbf{X}_{i}'\boldsymbol{\beta}\}. \tag{2.14}$$

Then the likelihood and the log likelihood given the data are:

$$L(\delta) = \frac{1}{\prod_{i=1}^{k} s_{i}!} \exp\left\{-\sum_{i=1}^{n} n\Delta g_{0}(\tilde{y}_{i}) \exp(\mathbf{X}_{i}'\boldsymbol{\beta})\right\} \prod_{i=1}^{k} \left(n\Delta g_{0}(\tilde{y}_{i}) \exp(\mathbf{X}_{i}'\boldsymbol{\beta})\right)^{s_{i}} \tag{2.15}$$

$$\ell(\delta) = -\sum_{i=1}^{k} \log(s_{i}!) - \sum_{i=1}^{k} n\Delta g_{0}(\tilde{y}_{i}) \exp\{\mathbf{X}_{i}'\boldsymbol{\beta}\} + \sum_{i=1}^{k} s_{i} \log\left(n\Delta g_{0}(\tilde{y}_{i})\right) + \sum_{i=1}^{k} s_{i}\mathbf{X}_{i}'\boldsymbol{\beta}. \tag{2.16}$$

To find the maximum likelihood estimator for $\boldsymbol{\beta}$ we take the derivative and find the maximum. The derivative of the loglikelihood with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\sum_{i=1}^{k} n\Delta g_{0}(\tilde{y}_{i}) \exp\left\{\mathbf{X}_{i}'\boldsymbol{\beta}\right\} \mathbf{X}_{i} + \sum_{i=1}^{k} s_{i}\mathbf{X}_{i}. \tag{2.17}$$

Finding the MLE for $\boldsymbol{\beta}$ involves solving the following equation

$$\sum_{i=1}^{k} \exp\left\{\log\left(n\Delta g_{0}(\tilde{y}_{i})\right) + \mathbf{X}_{i}'\boldsymbol{\beta}\right\} \mathbf{X}_{i} = \sum_{i=1}^{k} s_{i}\mathbf{X}_{i}. \tag{2.18}$$

To solve for $\boldsymbol{\beta}$ here, we have Poisson regression with $\log\left(n\Delta g_0(\tilde{y}_i)\right)$ as an offset.

Hence, maximizing the likelihood for the Poisson regression model is the same as maximizing the likelihood for the Multinomial (see (2.13)). The advantage of using poisson regression is the inference on the parameter $\boldsymbol{\beta}$. Efron and Tibshirani (1996) discuss an estimate for the covariance of $\widehat{\boldsymbol{\beta}}$ using the Poisson form to check the significance of the exponential tilt parameters for the different sufficient statistics. For more information on the inference, see Efron and Tibshirani (1996).

2.2.4 Discussion

In this section, we will discuss certain aspects of the density estimation method. We will show the moment matching property found in Efron and Tibshirani (1996). We will show later that the same holds when extending the method to mixtures. Discussion of the choice of carrier, the number of breaks, or cells, and the choice of bandwidth is in Sections 2.4 and 2.6.

In their paper, Efron and Tibshirani (1996) state:

> The parameter values $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1)$ were chosen by maximum likelihood, that is, by maximizing $\prod_{i=1}^{n} g_{\boldsymbol{\beta}}(y_i)$, ignoring the fact that the carrier $\hat{g}_0(y)$ is itself data-dependent. This choice of $\widehat{\beta}$ matches the $t(y)$ moments of $g_{\widehat{\boldsymbol{\beta}}}(y)$ to their empirical averages:
>
> $$\int_{\mathcal{Y}} t(y) g_{\widehat{\boldsymbol{\beta}}}(y) dy = \frac{1}{n} \sum_{i=1}^{n} t(y_i)$$

In their case, $p = 1$ (matching the first moment) and $g_{\widehat{\boldsymbol{\beta}}}(y) = \hat{g}_0(y) \exp\{t'(y)\widehat{\boldsymbol{\beta}}\}$ and $t'(y) = (1, y, y^2)$. We will show this property for the general case with arbitrary p.

For model (2.4) and for observed values $y_1, \ldots, y_n$, the log likelihood given the data is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left(g_0(y_i)\right) + n\beta_0 + \sum_{i=1}^{n} t'_1(y_i)\beta_1 + \ldots + \sum_{i=1}^{n} t'_p(y_i)\beta_p. \tag{2.19}$$

Note that β_0 is a normalizing constant. If we let $t'(y) = (y, \ldots, y^p)$ and $\boldsymbol{\gamma}' = (\beta_1, \ldots, \beta_p)$, then,

$$\exp\{\beta_0\} = \frac{1}{\int g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy} \Rightarrow \exp\{-\beta_0\} = \int g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy \tag{2.20}$$

and the log likelihood becomes

$$\ell(\boldsymbol{\beta}) \propto -n \log \left(\int_{\mathcal{Y}} g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy \right) + \sum_{i=1}^{n} t'_1(y_i)\beta_1 + \ldots + \sum_{i=1}^{n} t'_p(y_i)\beta_p. \quad (2.21)$$

The derivative of (2.21) with respect to β_j for $j = 1, \ldots, p$ is:

$$\begin{aligned} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} =& -n \left(\frac{1}{\int_{\mathcal{Y}} g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy} \right) \int_{\mathcal{Y}} t_j(y) g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy \\ & + \sum_{i=1}^{n} t_j(y_i) \qquad (2.22) \\ =& -n \frac{\int_{\mathcal{Y}} t_j(y) g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy}{\int_{\mathcal{Y}} g_0(y) \exp\{t'(y)\boldsymbol{\gamma}\} dy} + \sum_{i=1}^{n} t_j(y_i) \qquad (2.23) \\ =& -n \int_{\mathcal{Y}} t_j(y) g_0(y) \exp\{t'(y)\boldsymbol{\beta}\} dy + \sum_{i=1}^{n} t_j(y_i). \qquad (2.24) \end{aligned}$$

Setting the derivative equal to 0, we obtain:

$$\frac{1}{n} \sum_{i=1}^{n} t_j(y_i) = \int_{\mathcal{Y}} t_j(y) g_{\widehat{\boldsymbol{\beta}}}(y) dy \quad (2.25)$$

Hence, we have shown that this choice of $\widehat{\boldsymbol{\beta}}$ matches the $t(y)$ moments of $g_{\widehat{\boldsymbol{\beta}}}(y)$ to their empirical moments. The exponential tilt model yields an estimate of the density that adjusts the kernel density estimate to match the designated set of moments. It is a nice compromise between parametric and nonparametric density estimates.

2.3 Extending the Method to a Mixture Distribution

The goal of this section is to show how to extend the density estimation method to a density estimation method for mixture distributions. Recall that we have the following model:

$$g_{\boldsymbol{\beta}}(y) = g_0(y) \exp\left\{t'(y)\boldsymbol{\beta}\right\}. \tag{2.26}$$

It follows that the density for the mixture model would have the following form:

$$g(y) = \sum_{l=1}^{m} \lambda_l g_{\boldsymbol{\beta}_l}(y) \tag{2.27}$$

where λ_l is the mixing proportions for $l = 1, \ldots, m$ and $\sum_{l=1}^{m} \lambda_l = 1$. We show in Section 2.3.1 the extension of the method in Section 2.2.1 to the univariate mixture model using the Multinomial model. In Section 2.3.2 we consider the Poisson model. As we will show, although the two models produced the same MLE for the exponential tilt parameters in the non-mixture case, it is not the case for the extension to mixtures. Hence, for estimation we will use the Multinomial in the future. We then develop an EM algorithm (see, Dempster et al., 1977) to find the maximum likelihood estimators for the parameters in the mixture.

2.3.1 The Mixture of Multinomials

Suppose that $y_1, \cdots, y_n$ is a random sample from a finite mixture model with the following form:

$$g_{\boldsymbol{\gamma}}(y_i) = \sum_{l=1}^{m} \lambda_l g_0(y_i) \exp\left(t'(y_i)\boldsymbol{\beta}_l\right) \tag{2.28}$$

where $\sum_{l=1}^{m} \lambda_l = 1$, where $g_0(y_j)$ is the carrier, $t'(y_j) = (1, y_j, y_j^2, \cdots, y_j^p)$, $\boldsymbol{\beta}_l' = (\beta_{l0}, \beta_{l1}, \cdots, \beta_{lp})$, and p is number of moments to be matched. We begin by discretizing the data into bins and using the midpoints of those bins, $\tilde{y}_i$, and the counts in those bins, s_i, for $i = 1, \cdots, k$, similar to Section 2.2.1.

Consider the Multinomial model. We have already stated (see Section 2.2.1 and (2.6)) that the likelihood of the Mult$(n, \boldsymbol{\pi})$ given the data is:

$$L(\boldsymbol{\pi}) = \binom{n}{s_1, \cdots, s_k} \prod_{i=1}^{k} \pi_i^{s_i}. \tag{2.29}$$

For this mixture model, the likelihood is:

$$L(\boldsymbol{\delta}) = \binom{n}{s_1, \ldots, s_k} \prod_{i=1}^{k} \left(\sum_{l=1}^{m} \lambda_l \pi_{li} \right)^{s_i}$$

where $\boldsymbol{\delta}' = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_m, \lambda_1, \ldots, \lambda_m)$.

Let us consider the variable z_{il} for $i = 1, \cdots, k$ and $l = 1, \cdots, m$, such that

$$z_{il} = \begin{cases} 1, & \text{if the observations in } \mathcal{Y}_i \text{ belong to component } l \\ 0, & \text{otherwise} \end{cases} \tag{2.30}$$

where $\sum_{l=1}^{m} \lambda_l = 1$ and $\sum_{l=1}^{m} z_{il} = 1$. The complete likelihood would then be:

$$L(\delta) \propto \prod_{l=1}^{m} \prod_{i=1}^{k} (\lambda_l \pi_{li})^{s_i z_{il}} \tag{2.31}$$

The constraints on $\boldsymbol{\pi}_l$ are still the same as in Section 2.2.1. Therefore, $\sum_{i=1}^{k} \pi_{li} = 1$ for $l = 1, \ldots, m$. The log of the complete likelihood is:

$$\begin{aligned} \ell_c(\boldsymbol{\delta}) = & \sum_{i=1}^{k} s_i \sum_{l=1}^{m} z_{il} \log(\lambda_l) + \sum_{i=1}^{k} \sum_{l=1}^{m} s_i z_{il} \mathbf{X}_i' \boldsymbol{\beta}_l \\ & + \sum_{l=1}^{m} \eta_l \left(\sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} - 1 \right) \end{aligned} \tag{2.32}$$

We will use an EM algorithm to find the estimates of the parameters. Let $\boldsymbol{\gamma}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_m')$ and $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_m)$. Let the current values of $\boldsymbol{\lambda}'$ and $\boldsymbol{\gamma}'$ be $\boldsymbol{\gamma}^{(t)}$ and $\boldsymbol{\lambda}^{(t)}$, respectively. First, find the expectation of the complete likelihood, given the data. The following shows this expectation.

$$\mathbb{E}(\ell_c|\text{data}) = \sum_{i=1}^{k} s_i \sum_{l=1}^{m} w_{il}^{(t)} \log(\lambda_l^{(t)}) + \sum_{i=1}^{k} \sum_{l=1}^{m} s_i w_{il}^{(t)} \mathbf{X}_i' \boldsymbol{\beta}_l^{(t)} \tag{2.33}$$

where

$$w_{il}^{(t)} = \frac{\lambda_l^{(t)} \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l^{(t)}\}}{\sum_{l'=1}^{m} \lambda_{l'}^{(t)} \exp\{\mathbf{X}_i' \boldsymbol{\beta}_{l'}^{(t)}\}} \tag{2.34}$$

We call $w_{il}^{(t)}$ the posterior probability. It is the probability that a particular observation from y_i is in component l given the data. It is worth noting that this probability only depends on the mixing proportion and the parameters in the exponential tilt.

Next, we need to find the maximum likelihood estimates for λ_l and $\boldsymbol{\beta}_l$. When we maximize the complete likelihood with respect to λ_l, we get

$$\widehat{\lambda}_l^{(t+1)} = \frac{\sum_{i=1}^{k} s_i w_{il}^{(t)}}{n} \tag{2.35}$$

for $l = 1, \ldots, m$.

The next step is to maximize the complete likelihood, (2.32), with respect to $\boldsymbol{\beta}_l$, $l = 1, \ldots, m$. This problem entails finding the maximum, with respect to $\boldsymbol{\beta}_l$, of

$$\Phi(\boldsymbol{\beta}_l, \eta_l) = \sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i' \boldsymbol{\beta}_l + \eta_l \left(\sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} - 1 \right). \tag{2.36}$$

The gradient of Φ with respect to $\boldsymbol{\beta}_l$ is

$$\nabla_{\boldsymbol{\beta}_l} \Phi = \sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i + \eta_l \sum_{i=1}^{k} \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \mathbf{X}_i \tag{2.37}$$

and the derivative with respect to η_l gives $\sum_{i=1}^{k} \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} = 1$. If we look at the gradient with respect to $\boldsymbol{\beta}_l$ more carefully, we obtain

$$\frac{\partial \Phi}{\partial \beta_{l0}} = \sum_{i=1}^{k} s_i w_{il}^{(t)} + \eta_l \sum_{i=1}^{k} \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \tag{2.38}$$

$$\frac{\partial \Phi}{\partial \beta_{l1}} = \sum_{i=1}^{k} s_i w_{il}^{(t)} \tilde{y}_i + \eta_l \sum_{i=1}^{k} \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \tilde{y}_i \tag{2.39}$$

$$\vdots \tag{2.40}$$

$$\frac{\partial \Phi}{\partial \beta_{lp}} = \sum_{i=1}^{k} s_i w_{il}^{(t)} \tilde{y}_i^p + \eta_l \sum_{i=1}^{k} \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \tilde{y}_i^p \tag{2.41}$$

Setting (2.38) equal to zero and solving for η_l yields

$$\eta_l = -\sum_{i=1}^{k} s_i w_{il}^{(t)} \tag{2.42}$$

Combining (2.37) and (2.42) yields

$$\nabla_{\boldsymbol{\beta}_l} \Phi = \sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i - \sum_{i=1}^{k} \left(\sum_{i'}^{k} s_{i'} w_{i'l}^{(t)} \right) \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \mathbf{X}_i \tag{2.43}$$

Thus the maximum likelihood estimate for $\boldsymbol{\beta}_l$ is the solution to

$$\sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i = \sum_{i=1}^{k} \left(\sum_{i'}^{k} s_{i'} w_{i'l}^{(t)} \right) \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \mathbf{X}_i \tag{2.44}$$

We can rewrite the (2.44) as

$$\sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i = \sum_{i=1}^{k} \left(\sum_{i'}^{k} s_{i'} w_{i'l}^{(t)} \right) \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \mathbf{X}_i \begin{pmatrix} w_{il}^{(t)} \\ \hline w_{il}^{(t)} \end{pmatrix}$$

$$\Rightarrow \sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i = \sum_{i=1}^{k} \begin{pmatrix} \sum_{i'}^{k} s_{i'} w_{i'l}^{(t)} \\ \hline w_{il}^{(t)} \end{pmatrix} \Delta g_0(\tilde{y}_1) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\} \mathbf{X}_i w_{il}^{(t)}$$

$$\Rightarrow \sum_{i=1}^{k} s_i w_{il}^{(t)} \mathbf{X}_i = \sum_{i=1}^{k} \exp \left\{ \log \left[\frac{\Delta g_0(\tilde{y}_i) \sum_{i'=1}^{k} s_{i'} w_{i'l}^{(t)}}{w_{il}^{(t)}} \right] + \mathbf{X}_i' \boldsymbol{\beta}_l \right\} \mathbf{X}_i w_{il}^{(t)} \qquad (2.45)$$

The solutions to (2.45) are the MLEs for the exponential tilt parameters. As with the non-mixture case, we can find the solutions to (2.45) by utilizing GLM software. These solutions can be computed using a Poisson regression algorithm with $\log \left[\frac{\Delta g_0(\tilde{y}_i) \sum_{i'=1}^{k} s_{i'} w_{i'l}^{(t)}}{w_{il}^{(t)}} \right]$ as the offset. Using the GLM software is only to *trick* the computer into calculating the solutions to (2.45). In the next section, we show that, unlike the non-mixture case, treating the counts as independent variables from a mixture of Poissons does not provide the same MLEs as treating them as coming from a mixture of Multinomials.

2.3.2 The Mixture of Poisson Regressions

Similar to the non-mixture case, we are going to consider what happens when we write the model as a mixture of independent poisson random variables instead of a multinomial random vector. Recall that in the non-mixture case if $\mathbf{s} \sim \text{Mult}_k(n, \boldsymbol{\pi}(\boldsymbol{\beta}))$ and if $\pi_i(\boldsymbol{\beta}) = \Delta g_0(\tilde{y}_i) \exp\{\mathbf{X}_i' \boldsymbol{\beta}\}$ for $i = 1, \ldots, k$, then we can also find the maximum likelihood estimate for $\boldsymbol{\beta}$ by modeling $s_i \overset{\text{ind}}{\sim} \text{Pois}\left(n\pi_i(\boldsymbol{\beta})\right)$. We might hope the mixture case will be similar.

In the Poison setting we have

$$s_i \overset{\text{ind}}{\sim} \text{Pois}\left(n \sum_{l=1}^{m} \lambda_l \pi_{il}\right) \tag{2.46}$$

Let

$$z_{il} = \begin{cases} 1 & \text{if } s_i \text{ comes for component } l \\ 0 & \text{otherwise} \end{cases} \tag{2.47}$$

for $l = 1, \ldots, m$. In the mixture setting, the z_{il} are not observed. If we did observe the complete data, (s_i, z_{il}), the likelihood, given the indicators, would be

$$L(\boldsymbol{\delta}) = \prod_{l=1}^{m} \prod_{i=1}^{k} \left(\frac{\exp\{-\gamma_{il}\} \gamma_{il}^{s_i}}{s_i!} \right)^{z_{il}} \tag{2.48}$$

where $\gamma_{il} = n\Delta g_o(\tilde{y}_i) \exp\{\mathbf{X}_i' \boldsymbol{\beta}_l\}$ for $i = 1, \ldots, k$ and $l = 1, \ldots, m$ and the complete likelihood would be

$$L(\boldsymbol{\delta})_c = \prod_{l=1}^{m} \prod_{i=1}^{k} \left(\frac{\lambda_l^{s_i} \exp\{-\gamma_{il}\} \gamma_{il}^{s_i}}{s_i!} \right)^{z_{il}} \tag{2.49}$$

Then the log of the complete likelihood is

$$\begin{aligned}\ell(\boldsymbol{\delta})_c &\propto \sum_{l=1}^{m}\sum_{i=1}^{k} s_i z_{il}\log(\lambda_l) - \sum_{l=1}^{m}\sum_{i=1}^{k} z_{il}\gamma_{il} + \sum_{i=1}^{k}\sum_{l=1}^{m} s_i z_{il}\log(\gamma_{il}) \\ &\propto \sum_{l=1}^{m}\sum_{i=1}^{k} s_i z_{ik}\log(\lambda_l) - \sum_{l=1}^{m}\sum_{i=1}^{k} z_{il}\Delta n g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\} + \sum_{i=1}^{k}\sum_{l=1}^{m} s_i z_{il}\mathbf{X}_i'\boldsymbol{\beta}_l.\end{aligned} \tag{2.50}$$

For the EM algorithm, the E-step will give

$$\begin{aligned}w_{il}^{(t)} &= \frac{\lambda_l\left(n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\}\right)^{s_i}\exp\left\{-n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\}\right\}}{\sum_{l'=1}^{m}\lambda_{l'}\left(n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_{l'}\}\right)^{s_i}\exp\left\{-n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_{l'}\}\right\}} \\ &= \frac{\lambda_l\left(\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\}\right)^{s_i}\exp\left\{\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\}\right\}}{\sum_{l'=1}^{m}\lambda_{l'}\left(\exp\{\mathbf{X}_i'\boldsymbol{\beta}_{l'}\}\right)^{s_i}\exp\left\{\exp\{\mathbf{X}_i'\boldsymbol{\beta}_{l'}\}\right\}}.\end{aligned} \tag{2.51}$$

For the M-step of the algorithm, we need to maximize the complete likelihood with respect to both λ_l and $\boldsymbol{\beta}_l$. If we want to maximize the log complete likelihood with respect to λ_l we get:

$$\hat{\lambda}_l^{(t+1)} = \frac{\sum s_i w_{il}^{(t)}}{n} \tag{2.52}$$

Next, we want to maximize the log complete likelihood with respect to $\boldsymbol{\beta}_l$. We maximize

$$\Phi(\boldsymbol{\beta}_l) = -\sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\} + \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i'\boldsymbol{\beta}_l. \tag{2.53}$$

Thus, the MLE for $\boldsymbol{\beta}_l$, denote it $\widehat{\boldsymbol{\beta}}_l^{(t+1)}$, is the solution to

$$\sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\}\mathbf{X}_i = \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i \tag{2.54}$$

We can rewrite it as

$$\sum_{i=1}^{k} \exp\left\{\log\left(n\Delta g_0(\tilde{y}_i)\right) + \mathbf{X}_i'\boldsymbol{\beta}_l\right\} w_{il}^{(t)}\mathbf{X}_i = \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i. \tag{2.55}$$

We can find the MLE for $\boldsymbol{\beta}_l$ by using Poisson regression with $w_{il}^{(t)}$ as the weights and $\log\left(n\Delta g_0(\tilde{y}_i)\right)$ as the offset. At this point, what we have here is not the same as the Multinomial case. So rewriting the problem in terms of a mixture of Poisson distributions instead of a mixture of Multinomials is not the same as in the non-mixture case.

Now, consider what happens when we take into account $\sum_{i=1}^{k} \Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\} = 1$. The formula above becomes:

$$\Phi(\boldsymbol{\beta}_l) = -\sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\} + \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i'\boldsymbol{\beta}_l + \eta\left(\sum_{i=1}^{k}\Delta g_0(\tilde{y}_i)\exp\{\mathbf{X}_i'\boldsymbol{\beta}_l\} - 1\right) \tag{2.56}$$

We set the following derivatives equal to zero:

$$\frac{\partial\Phi}{\partial\beta_{10}} = -\sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)e^{\mathbf{X}_i'\boldsymbol{\beta}_l} + \sum_{i=1}^{k} s_i w_{il}^{(t)} + \eta\sum_{i=1}^{k}\Delta g_0(\tilde{y}_i)e^{\mathbf{X}_i'\boldsymbol{\beta}_l} \tag{2.57}$$

$$\Rightarrow \eta = \sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)e^{\mathbf{X}_i'\boldsymbol{\beta}_l} - \sum_{i=1}^{k} s_i w_{il}^{(t)} \tag{2.58}$$

$$\frac{\partial\Phi}{\partial\boldsymbol{\beta}_l} = \sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}_l)\mathbf{X}_i + \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i + \eta\sum_{i=1}^{k}\Delta g_0(\tilde{y}_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}_l)\mathbf{X}_i \tag{2.59}$$

$$\Rightarrow 0 = \sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}_l)\mathbf{X}_i + \sum_{i=1}^{k} s_i w_{il}^{(t)}\mathbf{X}_i \tag{2.60}$$

$$+\left(\sum_{i=1}^{k} w_{il}^{(t)} n\Delta g_0(\tilde{y}_i)e^{\mathbf{X}_i'\boldsymbol{\beta}_l} - \sum_{i=1}^{k} s_i w_{il}^{(t)}\right)\sum_{i=1}^{k}\Delta g_0(\tilde{y}_i)\exp(\mathbf{X}_i'\boldsymbol{\beta}_l)\mathbf{X}_i \tag{2.61}$$

We find that the MLE using the Poisson mixture model is not the same as that using the Multinomial mixture model. Treating the counts as being independent Poisson random variables does not result in the Multinomial mixture estimates of the exponential tilt parameters. In fact, the equations used to find the estimates, (2.45) in the Multinomial case and (2.61) in the Poisson case, are not the same. The equations for the estimates of the posterior probabilities, (2.34) for the Multinomial case and (2.54) for the Poisson case are not the same. The rest of this thesis proceeds using the Multinomial mixture results in Section 2.3.1.

2.4 The Carrier

Let $y_1, \ldots, y_n$ be a random sample from a density, g. Recall the kernel density estimate of g at y is

$$\hat{g}_h(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right)$$

where h is the bandwidth. There are many choices for the kernel, $K(\cdot)$. A common choice is the normal kernel where $K(w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right)$. The density estimation method purposed by Efron-Tibshirani involves discretizing the data. Therefore, they use a discretized version of the normal kernel density estimate. To find this estimate they use a smoothing matrix, $\hat{\boldsymbol{\mu}}^0 = \mathbf{M}(h)\mathbf{s}$. They took $\mathbf{M}(h)$ to be a normal kernel smoother with

$$\mathbf{M}_{i,j}(h) = \frac{c_i}{h} \phi\left(\frac{\tilde{y}_i - \tilde{y}_j}{h}\right)$$

where $i = 1, \cdots, k$ and $j = 1, \ldots, k$ and $\phi(\cdot)$ is the standard normal density function. The constants c_i were chosen so that $\mathbf{M}_{i+} = 1$ and h is the bandwidth; see Section 2.2.2.

Considering this carrier in more detail, we have

$$\begin{aligned}
\hat{\mu}_i^0 =& \mathbf{M}_i(h) \begin{pmatrix} s_1 \\ \vdots \\ s_k \end{pmatrix} = \sum_{j=1}^{k} \mathbf{M}_{i,j} s_j = \sum_{j=1}^{k} c_i \frac{1}{h} \phi\left(\frac{\tilde{y}_i - \tilde{y}_j}{h}\right) s_j \\
&= \frac{c_i}{h} \sum_{j=1}^{k} \phi\left(\frac{\tilde{y}_i - \tilde{y}_k}{h}\right) s_j
\end{aligned}$$

where $c_i = \dfrac{1}{\frac{1}{h}\sum_{j=1}^{k} \phi\left(\frac{\tilde{y}_i - \tilde{y}_j}{h}\right)}$.

$$\hat{\mu}_i^0 = \begin{bmatrix} c_1 \frac{1}{h}\phi\left(\frac{\tilde{y}_1-\tilde{y}_1}{h}\right) & \cdots & c_1 \frac{1}{h}\phi\left(\frac{\tilde{y}_1-\tilde{y}_k}{h}\right) \\ \vdots & & \vdots \\ c_k \frac{1}{h}\phi\left(\frac{\tilde{y}_k-\tilde{y}_1}{h}\right) & \cdots & c_k \frac{1}{h}\phi\left(\frac{\tilde{y}_k-\tilde{y}_k}{h}\right) \end{bmatrix} \begin{bmatrix} s_1 \\ \vdots \\ s_k \end{bmatrix}$$

$$= \begin{bmatrix} c_1 \frac{1}{h} \sum_{j=1}^{k} \phi\left(\frac{\tilde{y}_1-\tilde{y}_j}{h}\right) s_j \\ \vdots \\ c_k \frac{1}{h} \sum_{j=1}^{k} \phi\left(\frac{\tilde{y}_k-\tilde{y}_j}{h}\right) s_k \end{bmatrix} = n \begin{bmatrix} \Delta g_0(\tilde{y}_1) \\ \vdots \\ \Delta g_0(\tilde{y}_k) \end{bmatrix}$$

The choice of the carrier is very important. Although the carrier is data dependent, it is considered fixed. The carrier is also important in helping us determine identifiability. For the mixture model, we use this version of the carrier as well as a symmetrized version of it. The simulation results shown in Section 2.7 are found using both the kernel density estimate shown here (sp.density) and the one in which we force it to be symmetric (sp.density*). We force symmetry by finding the median the bin that contains the median of the observations. Then we finding the maximum value of the carrier density estimate and its reflection around the median. The result is a symmetric carrier. Even by forcing the carrier to be symmetric, we are unable to determine if the parameters in the model are identifiable.

2.5 Moment Matching

In this section, we are going to show the moment matching property of Section 2.2.4 for the mixture model. Consider the m component mixture model

$$h(y) = \sum_{l=1}^{m} \lambda_l g_{\boldsymbol{\beta}_l}(y)$$

where $g_{\boldsymbol{\beta}_l}(y) = g_0(y)\exp\{t'(y)\boldsymbol{\beta}_l\}$ and $\boldsymbol{\beta}'_l = (\beta_{l0}, \ldots, \beta_{lp})$ for $l = 1, \ldots, m$.

Suppose we observe $y_1, \ldots, y_n$. Let z_{il} be an unobserved Bernoulli random variable such that

$$z_{il} = \begin{cases} 1 & \text{if } y_i \text{ belongs to component } l \\ 0 & \text{otherwise} \end{cases} \tag{2.62}$$

for $i = 1, \ldots, n$ and $l = 1, \ldots, m$. If we did observe (y_i, z_{il}) then the complete likelihood is:

$$L_c(\boldsymbol{\gamma}) = \prod_{i=1}^{n}\prod_{l=1}^{m}\left(\lambda_l g_0(y_i)\exp\{t'(y_i)\boldsymbol{\beta}_l\}\right)^{z_{il}} \tag{2.63}$$

where $\boldsymbol{\gamma}' = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m, \lambda_1, \ldots, \lambda_m)$. The log of the complete likelihood is

$$\ell_c(\boldsymbol{\gamma}) = \sum_{i=1}^{n}\sum_{l=1}^{m} z_{il}\log\left(\lambda_l g_0(y_i)\right) + \sum_{i=1}^{n}\sum_{l=1}^{m} z_{il}\beta_{l0} + \sum_{i=1}^{n}\sum_{l=1}^{m} z_{il}t'_i(y)\boldsymbol{\alpha}_l \tag{2.64}$$

where $t'_i(y) = (y_i, y_i^2, \ldots, y_i^p)$ and $\boldsymbol{\alpha}'_l = (\beta_{l1}, \ldots, \beta_{lp})$.

Similarly to Section 2.2.4,

$$\exp\{-\beta_{l0}\} = \int_{\mathcal{Y}} g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy \tag{2.65}$$

for $l = 1, \ldots, m$. Substituting (2.65) in (2.64) gives

$$\ell_c(\boldsymbol{\gamma}) \propto -\sum_{i=1}^{n}\sum_{l=1}^{m} z_{il}\left(\int_{\mathcal{Y}} g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy\right) + \sum_{i=1}^{n}\sum_{l=1}^{m} z_{il}t'_i(y)\boldsymbol{\alpha}_l \tag{2.66}$$

Taking the derivative with respect to β_{lj} for $j = 1, \ldots, p$ results in

$$\begin{aligned}
\frac{\partial \ell_c}{\partial \beta_{lj}} &= -\left(\sum_{i=1}^{n} z_{il}\right)\left(\frac{1}{\int_{\mathcal{Y}} g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy}\right)\int_{\mathcal{Y}} t_j(y) g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy \\
&\quad + \sum_{i=1}^{n} z_{il} t_j(y_i) \\
&= -\left(\sum_{i=1}^{n} z_{il}\right)\left(\frac{\int_{\mathcal{Y}} t_j(y) g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy}{\int_{\mathcal{Y}} g_0(y)\exp\{t'(y)\boldsymbol{\alpha}_l\}dy}\right) + \sum_{i=1}^{n} z_{il} t_j(y_i) \\
&= -\left(\sum_{i=1}^{n} z_{il}\right)\int_{\mathcal{Y}} t_j(y) g_0(y)\exp\{t'(y)\boldsymbol{\beta}_l\}dy + \sum_{i=1}^{n} z_{il} t_j(y) \qquad (2.67)
\end{aligned}$$

By setting the derivative equal to 0 we obtain

$$\frac{\sum_{i=1}^{n} z_{il} t_j(y_i)}{\sum_{i=1}^{n} z_{il}} = \int_{\mathcal{Y}} t_j(y) g_{\widehat{\boldsymbol{\beta}}_l}(y)dy \qquad (2.68)$$

The choice of $\widehat{\boldsymbol{\beta}}_l,\ l = 1, \ldots, m$ matches the $t(y)$ moments of $g_{\hat{\beta}_l}(y)$ to their empirical averages. For example, $g_{\widehat{\boldsymbol{\beta}}_1}(y) = g_0(y)\exp(\hat{\beta}_0 + t'(y)\widehat{\boldsymbol{\alpha}}_1)$ matches the first p moments of the data from component 1.

2.6 Bandwidths and Breaks

There are two aspects of the model that still require discussion: the choice of bandwidth and the number of breaks. As with kernel density estimates, the choice of the bandwidth is important. If too large a bandwidth is chosen, the density estimate will over-smooth the true density. Too small a bandwidth will cause the estimate to under-smooth the density. Over-smoothing may mask features of the underlying density and under-smoothing may show features that are not of the underlying density. Thus the appropriate bandwidth choice is important when estimating the true density. Efron and Tibshirani (1996) suggest that since the moments are estimated unbiasedly, this allows the carrier to use a larger bandwidth.

The number of breaks also deserves some explanation. It seems logical that the number of breaks chosen makes a significant difference in how well the method estimates the density. Efron and Tibshirani (1996) suggest that the number of breaks makes very little difference in the estimates. We also found that the number of breaks makes little difference. We have not tried examples where the number of breaks is small, i.e. $k < 30$.

We provide simulations below that show the effects of the choice of bandwidth and number of breaks. Each example had a sample size of $n = 150$. The first example is from a two-component location mixture of normal scores. The component means and standard deviations are $\mu = (0, 3)$ and $\sigma = (1, 1)$, respectively, with a mixing proportion of $\lambda = 0.5$. The normal scores were obtained from the `qqnorm` function in `R`. We chose normal scores because they produce a perfect sample of normals in order to better show the effects of the choice of bandwidth and breaks. Table 2.1 shows the results for the normal scores using different breaks (or bins) and different bandwidths. As seen in the table, changing the number of breaks changed the values very little. The bandwidth seems to change the values significantly, however. When the bandwidth increases, the values tend to be very close to the true values. This is consistent with the claim that there is little danger in over-smoothing the density estimates.

For the next example, we have a location/scale mixture of normal scores with two components and mixing proportion of $\lambda = 0.5$, mean $\mu = (0, 4)$, and standard deviation $\sigma = (1, 2)$. The results for the component means and standard deviations are shown in

	μ_1	σ_1	μ_2	σ_2	λ
True Values	0	1	3	1	0.50
Bandwidth=.5					
Breaks=30	0.1196	1.1533	2.8828	1.1519	0.4996
Breaks=100	0.1251	1.1598	2.8774	1.1583	0.5005
Bandwidth=1					
Breaks=30	0.0271	1.0333	2.9732	1.0331	0.5001
Breaks=100	0.0279	1.0344	2.9724	1.0342	0.5000
Bandwidth=2					
Breaks=30	0.0079	1.0056	2.9919	1.0057	0.5000
Breaks=100	0.0083	1.0060	2.9918	1.0059	0.5000
Bandwidth=3					
Breaks=30	0.0089	1.0069	2.9913	1.0068	0.5000
Breaks=100	0.0090	1.0071	2.9908	1.0072	0.5000

Table 2.1. The semiparametric estimates for the mixing proportion and the component means and standard deviations using different bandwidths and breaks for normal scores.

Table 2.3. The estimates are poor for this example when the bandwidth is small. As in the other examples, the number of breaks does not seem to make much difference in the estimates. The estimated component densities are shown in Figure 2.1 when the bandwidth is 0.5 and Figure 2.2 shows the estimates when the bandwidth is 2. The figures show that the method has difficulty finding the components when the bandwidth is small. But as we increase the bandwidth, the estimates are much closer to the true values and the estimated densities are much closer as well.

For the next example, we constructed gamma scores using a Q-Q plot with the gamma distribution. We used a two component gamma scores mixture with mean of $\mu = (2, 4.5)$ and standard deviation $\sigma = (2, 1.5)$ (or $(\alpha_1, \beta_1) = (1, 2)$ and $(\alpha_2, \beta_2) = (9, 0.5)$ where α is the shape parameter and β is the scale parameter). Table 2.4 shows the estimates for the mixing proportion and the component means and standard deviations. The estimates are fairly close to the true estimates. The number of breaks and the bandwidths do not seem to make much difference as the bandwidth increases. When the bandwidth is smaller however, the fit seems better. The estimated component densities for $bw = 1$ and $bw = 7$ for 30 breaks are shown in Figures 2.3 and 2.4. Although the

	μ_1	σ_1	μ_2	σ_2	λ
True Values	0	1	2	1	0.50
Bandwidth=.5					
Breaks=30	0.1178	1.1015	1.8822	1.1015	0.5000
Breaks=100	0.1344	0.1146	1.8656	1.1146	0.5000
Bandwidth=1					
Breaks=30	0.0098	1.0055	1.9902	1.0055	0.5000
Breaks=100	0.0138	1.0094	1.9862	1.0094	0.5000
Bandwidth=2					
Breaks=30	0.0122	1.0078	1.9878	1.0078	0.5000
Breaks=100	0.0163	1.0119	1.9837	1.0119	0.5000
Bandwidth=3					
Breaks=30	0.0176	1.0131	1.9824	1.0131	0.5000
Breaks=100	0.0223	1.0176	1.9777	1.0176	0.5000

Table 2.2. The semiparametric estimates for the mixing proportion and the component means and standard deviations using different bandwidths and breaks for normal scores.

mixing proportions are a little off from the true values, the shapes of the densities are very close.

In conclusion, it does not seem like there is any disadvantage to over-smoothing the density estimates by choosing a larger bandwidth. We found larger bandwidths work for mixture models perhaps because of the moment matching property; see Section 2.5 and Efron and Tibshirani (1996). The number of breaks does not seem to make much difference. It seems that the estimates for the gamma scores mixture are biased downward. It is worth noting that the gamma model used in this chapter is an example where the component densities do not satisfy the exponential tilt assumption.

	μ_1	σ_1	μ_2	σ_2	λ
True Values	0	1	4	2	0.50
Bandwidth=.5					
Breaks=30	-0.1668	0.9249	2.5217	2.5358	0.1940
Breaks=100	-0.1741	0.8429	2.3837	2.5525	0.1500
Bandwidth=1					
Breaks=30	-0.1071	0.9473	3.2935	2.3466	0.3804
Breaks=100	-0.1283	0.9052	3.1196	2.4124	0.3447
Bandwidth=2					
Breaks=30	-0.0032	0.9991	3.9598	2.0183	0.4945
Breaks=100	-0.0061	0.9958	3.9501	2.0253	0.4929
Bandwidth=3					
Breaks=30	0.0063	1.0027	3.9895	2.0040	0.4995
Breaks=100	0.0036	0.9997	3.9834	2.0073	0.4984

Table 2.3. The semiparametric estimates for the mixing proportion and the component means and standard deviations using different bandwidths and breaks for normal scores.

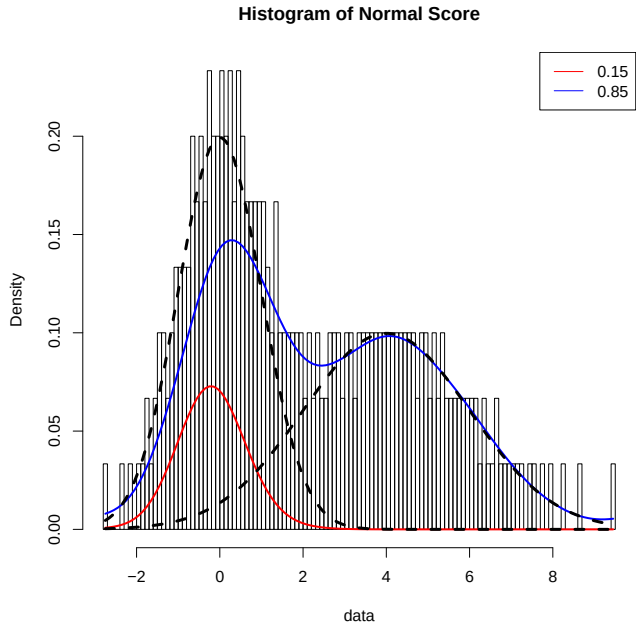


Fig. 2.1. Semiparametric estimates of the component densities for the normal scores when $bw = 0.5$ and 100 breaks. The dashed lines are the true densities. The values in the legend are the estimated mixing proportions.

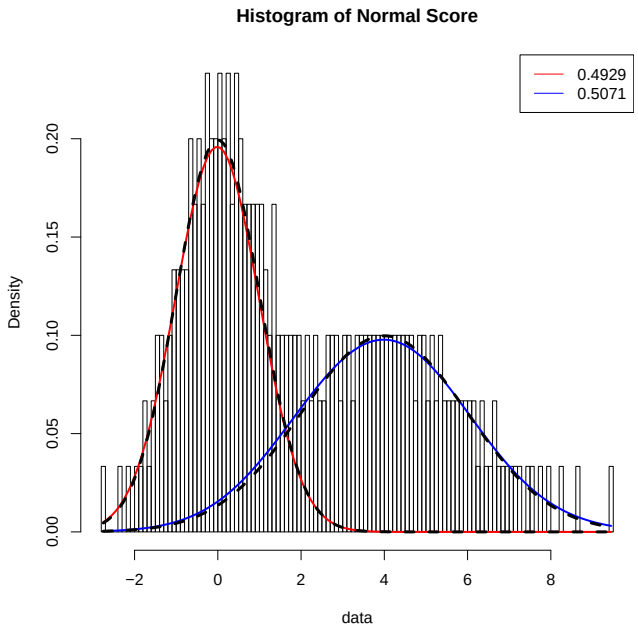


Fig. 2.2. Semiparametric estimates of the component densities for the normal scores when $bw = 2$ and 100 breaks. The dashed lines are the true densities.

	μ_1	σ_1	μ_2	σ_2	λ
True Values	2	2	4.5	1.5	0.50
Bandwidth=1					
Breaks=30	2.3716	2.3094	3.9192	1.7423	0.4343
Breaks=100	2.0217	2.2934	3.7791	1.8465	0.3028
Bandwidth=2					
Breaks=30	1.8495	2.2088	3.9495	1.7371	0.3344
Breaks=100	1.8259	2.2072	3.9483	1.7349	0.3304
Bandwidth=5					
Breaks=30	1.7697	2.1593	4.0019	1.7076	0.3382
Breaks=100	1.7534	2.1558	4.0063	1.7025	0.3370
Bandwidth=7					
Breaks=30	1.7699	2.1596	4.0067	1.7071	0.3383
Breaks=100	1.7534	2.1560	4.0067	1.7019	0.3372

Table 2.4. The semiparametric estimates for the mixing proportion and the component means and standard deviations using different bandwidths and breaks for gamma scores.

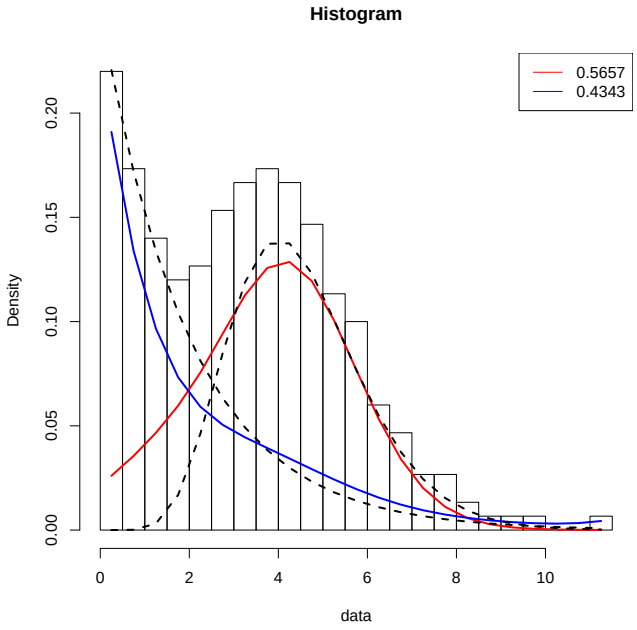


Fig. 2.3. Semiparametric estimates of the component densities for the gamma scores when $bw = 1$ and 30 breaks. The dashed lines are the true densities.

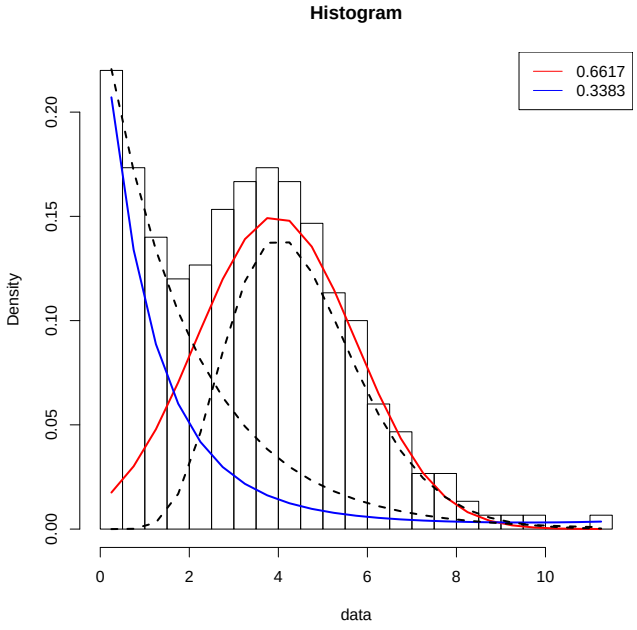


Fig. 2.4. Semiparametric estimates of the component densities for the gamma scores when $bw = 7$ and 30 breaks. The dashed lines are the true densities.

2.7 Monte Carlo Simulations

In this section, we present simulations to determine how well the density estimation method proposed in the chapter works. We compare our results using those from a univariate normal mixture and also using the method proposed in Bordes et al. (2007). The functions for their method can be found in the `mixtools` package in R under `spEMsymloc` (see Benaglia et al. (2009b)). For the tilted method and the normal method, we started the simulations with a random $n \times m$ matrix of posterior probabilities, where n is the number of observations and m is the number of components. The number of breaks for the tilted method was set at 30 and the bandwidths for these simulations were chosen using twice the bandwidth found by

$$h = 0.9 \min \left\{ SD, \frac{IQR}{1.34} \right\} n^{-1/5} \tag{2.69}$$

where SD is the standard deviation of the data, IQR is the interquartile range, and n is the number of observations; see Silverman (1986). The nonparametric method finds starting values using `kmeans` and we used the non-stochastic option (see Benaglia et al. (2009a) for more details on these functions).

For the first set of simulations, we simulated 200 datasets of sample size $n = 200$ from the following mixture

$$g(x) = \lambda N(x; -1, 1) + (1 - \lambda) N(x; 2, 1) \tag{2.70}$$

where $\lambda = 0.5$ and $N\left(\cdot; \mu, \sigma^2\right)$ is the normal density function with mean μ and variance σ^2. The results for the semiparametric tilted method (sp.density), the semiparametric tilted method with a symmetric carrier (sp.density*), the univariate normal mixture (normal), and the method proposed by Benaglia et al. (2009a) (np) are shown in Table 2.5. The plot of one randomly selected simulated dataset is shown in Figure 2.5. We expect the results from the exponential tilt method to be similar to those from the normal since the normal mixture does satisfy the exponential tilt assumption. We can see from the table that this is the case. The results from both of those methods are

similar and are close to the true values. The estimates from the tilted method and from the normal method provide closer results than those from the nonparametric method. We also generated 200 simulations from Model (2.70) with $\lambda = 0.3$ and $n = 200$. The results are shown in Table A.2 in Appendix A and they are similar to the results for $\lambda = 0.3$.

	True	sp.density	sp.density*	normal	np
λ	0.5	0.4860(0.0624)	0.4837(0.0700)	0.4902(0.0516)	0.5013(0.0430)
μ_1	-1	-1.0304(0.1724)	-1.0342(0.1856)	-1.0499(0.1440)	-0.7526(0.4222)
μ_2	2	1.9310(0.1896)	1.9239(0.2166)	1.9732(0.1572)	1.7394(0.4522)
σ_1	1	0.9995(0.1258)	0.9953(0.1416)	0.9647(0.1018)	1.2239(0.2025)
σ_2	1	1.0603(0.1223)	1.0631(0.1395)	1.0107(0.1002)	1.2289(0.1995)

Table 2.5. The semiparametric estimates of the component means (standard errors) and standard deviations for Model (2.70).

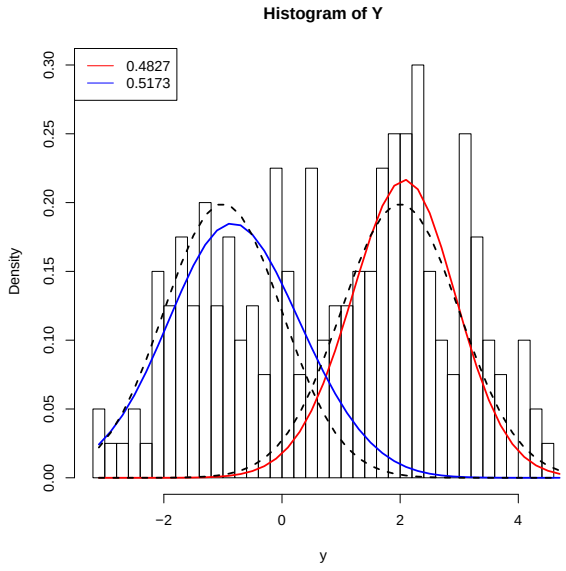


Fig. 2.5. The semiparametric component density estimates for a randomly selected simulated dataset from Model (2.70).

The next example is a location/scale mixture. The method proposed by Benaglia et al. (2009a) is for a location mixture with symmetric distributions. Since the function is not designed to take into account mixtures with different variances, we compare the results of the tilted method with those from a normal mixture method. We generated 200 simulated datasets of sample size $n = 200$ from the following model:

$$h(x) = \lambda N(x; 0, 1) + (1 - \lambda) N(x; 5, 4) \tag{2.71}$$

with $\lambda = 0.3, 0.5$. We report the results for $\lambda = 0.5$ here and the results for $\lambda = 0.3$ are in Appendix A. Table 2.6 displays the estimated component means and standard deviations and Figure 2.6 shows the estimated component densities for a randomly selected simulated dataset. From the table, we can see that the tilted method performs similarly to the normal mixture method.

Parameter	True	sp.density	sp.density*	normal
λ	0.5	0.4904(0.0332)	0.4872(0.0449)	0.5133(0.0292)
μ_1	0	-0.0067(0.1112)	-0.0063(0.1090)	0.0266(0.1075)
μ_2	5	4.8909(0.2634)	4.8748(0.3603)	5.0827(0.2188)
σ_1	1	1.0026(0.0900)	0.9926(0.0992)	1.0124(0.0855)
σ_2	2	2.0854(0.1875)	2.0838(0.2433)	1.9256(0.1589)

Table 2.6. The estimated semiparametric component density estimates for a randomly selected simulated dataset from Model (2.71). The table shows the means(standard errors) of the estimates.

The next set of simulations is also from a normal location mixture model with the following parameters:

$$g(x) = \lambda_1 N(x; 0, 1) + \lambda_2 N(x; 2, 1) + \lambda_3 N(x; 3, 1) \tag{2.72}$$

which $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$. We are interesting in seeing how well the three methods perform when there are three components and when two of the components are not well separated. We generated 200 simulations of sample size $n = 500$. In Table 2.7, we

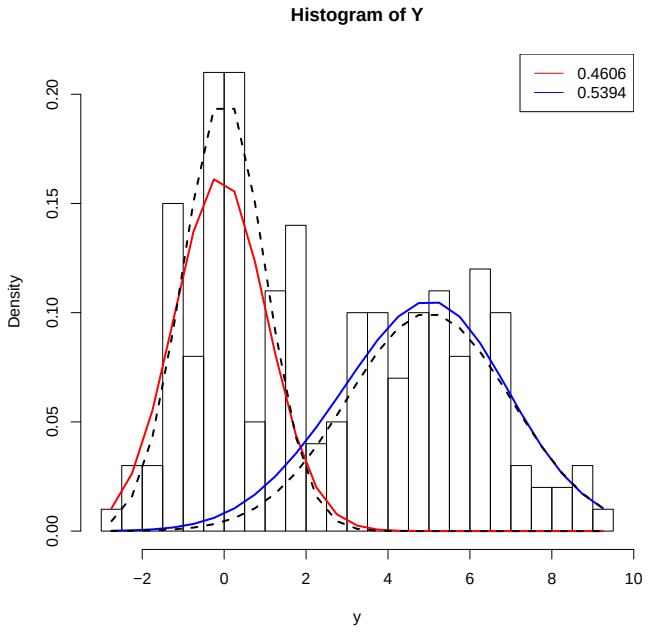


Fig. 2.6. The semiparametric component density estimates for a randomly selected simulated dataset from Model (2.71) which is a normal mixture model with $\lambda = 0.5$, $\mu = (0, 5)$, and $\sigma^2 = (1, 4)$.

display the estimated means(standard errors) for the component means and standard deviations for the tilted method (sp.density), the tilted method with a symmetric carrier (sp.density*), the normal method (normal), and the nonparametric method (np). The plot of the estimated component densities for a randomly selected simulation is shown in Figure 2.7. From the table, the tilted method performs similarly to the normal method, with the normal method performing slightly better. For the Bordes et al. (2007) method, the estimated component means and standard deviations are not as close to the true estimates as with the other methods. This maybe due to not being able to start the algorithm with different starting values (using the default bandwidth). In the figure, the estimates are close to the true density functions even when the distributions are not well separated.

For the next example, we will consider models that are not symmetric. As we have stated before, the function to calculate the estimates for the Benaglia et al. (2009a) method is not capable of handling data that is not symmetric. Although we cannot

	True	sp.density	sp.density*	normal	np
λ_1	$\frac{1}{3}$	0.2762(0.0789)	0.2771(0.0792)	0.3258(0.0942)	0.2707(0.0630)
λ_2	$\frac{1}{3}$	0.3439(0.1197)	0.3556(0.1255)	0.2655(0.2232)	0.4480(0.0797)
λ_3	$\frac{1}{3}$	0.3823(0.1333)	0.3699(0.1409)	0.4086(0.2248)	0.2814(0.0566)
μ_1	0	0.2508(0.5193)	0.2185(0.5500)	-0.0366(0.3333)	1.4725(0.4759)
μ_2	2	1.6720(0.4051)	1.6981(0.3887)	1.7598(0.6377)	1.7360(0.2415)
μ_3	3	2.6348(0.4503)	2.6753(0.3469)	2.9489(0.5175)	1.7361(0.2415)
σ_1	1	1.0934(0.2700)	1.1127(0.2895)	0.9702(0.1605)	1.5620(0.1321)
σ_2	1	1.5338(0.3102)	1.5087(0.3044)	0.7685(0.4954)	1.5564(0.1294)
σ_3	1	1.0838(0.2932)	1.0115(0.2490)	0.8948(0.2988)	1.5564(0.1294)

Table 2.7. The estimated semiparametric component density estimates for a randomly selected simulated dataset from Model (2.72). The table shows the means(standard errors) of the estimates. The column labeled sp.density* has a symmetric carrier.

determine at this time whether the parameters in our model are identifiable, we know a sufficient condition for identifiability in the Benaglia et al. (2009a) method is that the distributions have to be symmetric and it has to be a location only mixture.

We simulated 200 samples of size $n = 300$ from the following gamma mixture model:

$$g(x) = \lambda G(x; 2, 2) + (1 - \lambda) G(x; 5, 2) \tag{2.73}$$

where $\lambda = 0.5$ and $G(\cdot; \alpha, \beta)$ is a gamma density function with mean $\alpha\beta$ and variance $\alpha\beta^2$. The results from the tilted method as well as the normal mixture method are shown in Table 2.8 and estimated densities for a randomly selected simulated dataset is shown in Figure 2.8. Since the normal mixture model is not the correct model for this dataset, we do not expect the method to perform well. There is not, however, any other method to compare the tilted method to as all the other methods require the symmetric/location assumption. The results show that the estimates for the component means and standard deviations have smaller standard errors than those of the normal mixture model with the tilted method using the symmetric carrier performing slightly better.

The second gamma mixture we simulated was from a two-component mixture with $\lambda = 0.5$ and $\alpha = (2, 5)$ and $\beta = (2, 1)$ (or with component means $(4, 5)$ and component

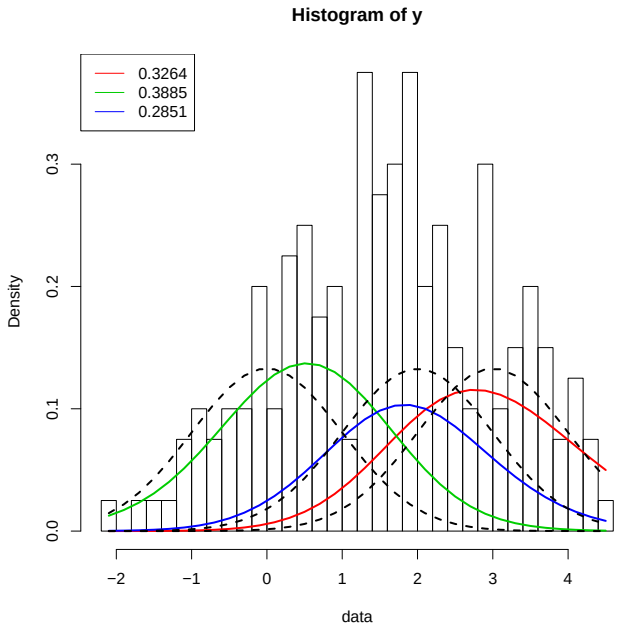


Fig. 2.7. The semiparametric component density estimates for a randomly selected simulated dataset from Model (2.72) which is a normal mixture model with three components.

standard deviations $(2.8284, 2.8296)$). We generated 200 simulations with sample size $n = 500$. The results for the tilted method and the normal mixture method are found in Table 2.9 and the component density estimates for a randomly selected simulation are plotted in Figure 2.9. The two component densities in this model are not well separated. From the Table, we can see that the tilted method using the symmetric carrier is not able to find the two components. As for the tilted method with the unrestricted carrier, the standard errors of the estimates are much smaller than those of the normal mixture method and the means are either better or just as good as those from the normal.

The last set of simulations are from a two-component Laplace distribution with the form:

$$h(y) = \lambda \mathcal{L}(y; \mu_1, b_1) + (1 - \lambda)\mathcal{L}(y; \mu_2, b_2) \tag{2.74}$$

where $\mathcal{L}(\cdot; \mu, b)$ is the Laplace density function with mean μ and variance $2b^2$. For this example, we generated 200 dataset of sample size of $n = 500$. The first set we considered

Parameter	True	sp.density	sp.density*	normal
λ	0.5	0.4075(0.2540)	0.3546(0.1590)	0.3540(0.3272)
μ_1	2	2.1082(0.7683)	1.6779(0.5794)	1.1193(1.0869)
μ_2	4.5	3.877(0.3500)	4.0442(0.2741)	4.6843(2.0459)
σ_1	2	2.2417(0.3482)	2.0475(0.3896)	0.7241(0.6999)
σ_2	1.5	1.7002(0.368)	1.6757(0.2375)	1.8775(0.6983)

Table 2.8. The estimated semiparametric component density estimates for a randomly selected simulated dataset from Model (2.73). The table shows the means(standard errors) of the estimates.

Parameter	True	sp.density	sp.density*	normal
λ	0.5	0.4453(0.1227)	0.9299(0.1712)	0.6767(0.1294)
μ_1	4	3.2558(0.3302)	4.3153(0.3946)	3.412(0.3807)
μ_2	5	5.4960(0.3312)	13.5624(4.6914)	6.8706(0.7269)
σ_1	2.8284	1.5172(0.2555)	2.3778(0.3048)	1.6096(0.2273)
σ_2	2.2361	2.8296(0.2086)	2.2824(1.2285)	2.8417(0.4084)

Table 2.9. The estimated semiparametric component density estimates for a randomly selected simulated dataset from second gamma mixture. The table shows the means(standard errors) of the estimates.

has $\lambda = 0.5$, $\mu = (\mu_1, \mu_2) = (0, 3)$, and $\sigma = (1, 1)$ (or $b = (1, 1)$). The results for the tilted and normal mixture methods are shown in Table 2.10.

The final model has the same form as model (3.82) but has $\lambda = 0.5$, $\mu = (0, 5)$, and $b = (2, 1)$ (or has component means $\mu = (0, 5)$ and component standard deviations $\sigma = (2\sqrt{2}, \sqrt{2})$). There were 200 simulations with sample size $n = 500$. The results are shown in Table 2.11.

Overall, the tilted method seems to be a robust and flexible method. The results from the tilted method for the normal simulations are close to those when using the normal mixture method. Even when the component distributions were not symmetric and the exponential tilted assumption is not valid, the tilted method overall produces better estimates than choosing using a mixture of normals.

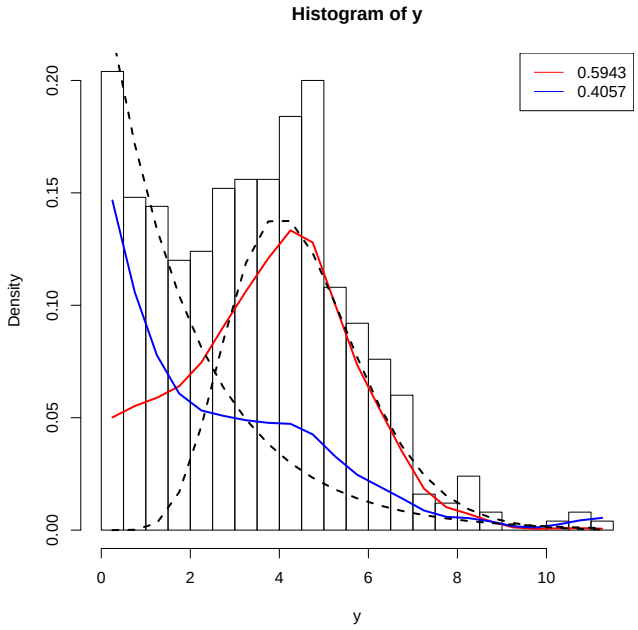


Fig. 2.8. The semiparametric component density estimates for a randomly selected simulated dataset from Model (2.73) which is a gamma mixture model with two components.

Parameter	True	sp.density	sp.density*	normal	np
λ	0.5	0.4872(0.3451)	0.4708(0.3738)	0.4856(0.3176)	0.5040(0.0526)
μ_1	0	0.8118(0.7587)	0.5933(0.7647)	0.3881(0.6772)	0.7476(0.6678)
μ_2	3	2.1549(0.8263)	2.5039(1.2208)	2.6418(0.9897)	2.247(0.6646)
σ_1	1.4142	2.6841(1.9691)	2.5576(1.9483)	1.3529(0.8118)	1.7743(0.2683)
σ_2	1.4142	2.3811(1.5927)	2.1465(1.4617)	1.4082(0.7743)	1.7685(0.2721)

Table 2.10. The estimated semiparametric component density estimates for a randomly selected simulated dataset from Model (3.82). The table shows the means(standard errors) of the estimates for Model (3.82).

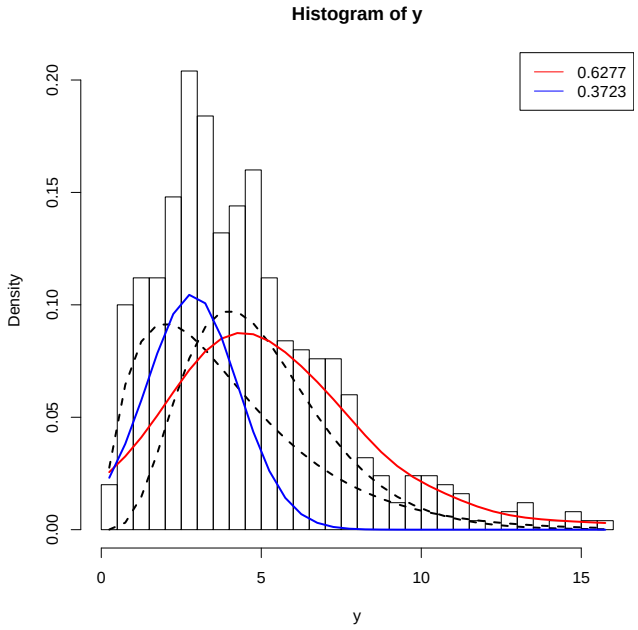


Fig. 2.9. The semiparametric component density estimates for a randomly selected simulated dataset from the second gamma mixture model with two components.

Parameter	True	sp.density	sp.density*	normal
λ	0.5	0.6867(0.1827)	0.3684(0.3256)	0.6478(0.1666)
μ_1	0	1.5865(0.5235)	0.2709(1.4972)	1.179(0.5687)
μ_2	5	4.7912(0.7813)	3.8133(1.1745)	4.9547(0.6009)
σ_1	2.8284	3.6972(1.611)	4.366(3.0531)	3.1596(0.7617)
σ_2	1.4142	1.0397(0.8029)	2.1148(1.2057)	0.921(0.6679)

Table 2.11. The estimated semiparametric component density estimates for a randomly selected simulated dataset from second Laplace mixture. The table shows the means(standard errors) of the estimates.

2.8 Example

The real data example we will analyze with the tilted method proposed here is data collected from Old Faithful Geyser in Yellowstone National Park, Wyoming (see Azzalini and Bowman (1990) and Hunter et al. (2007)). The waiting times between eruptions (minutes) and the duration of the eruptions were recorded for 272 occasions ($n = 272$). We will analyze the waiting times between eruptions. The dataset of the waiting times for eruptions can be found in the standard version of `R` under the dataset *faithful*.

The results in Table 2.12 show the results using the tilted method (sp.density), the tilted method using a symmetric carrier (sp.density*), the normal mixture method (normal), the normal mixture method with equal variances (normal*) and the nonparametric method proposed by Bordes et al. (2007). From the table, we can see that the estimates of the component means and standard deviations are similar for each of the parameters in all five cases. The estimated component densities are plotted in Figure 2.10.

Table 2.12. Results for the Old Faithful Data

Method	$\hat{\mu}_1$	$\hat{\sigma}_1$	$\hat{\mu}_2$	$\hat{\sigma}_2$	$\hat{\lambda}$
sp.density	54.9046	6.5440	79.7910	6.4342	0.3574
sp.density*	53.8530	5.4245	79.3541	6.7267	0.3316
normal	54.6148	5.8712	80.0911	5.8678	0.3609
normal*	54.6136	5.8691	80.0903	5.8691	0.3608
np	54.6853	6.2804	79.7584	6.4074	0.3534

2.9 Conclusion

In this chapter, we presented a semiparametric method, which includes an exponential tilt, to estimate the component densities in finite univariate mixture models. We did not show the parameters in the model were identifiable. Although we did not show the identifiability, the method works very well in our simulations estimating component

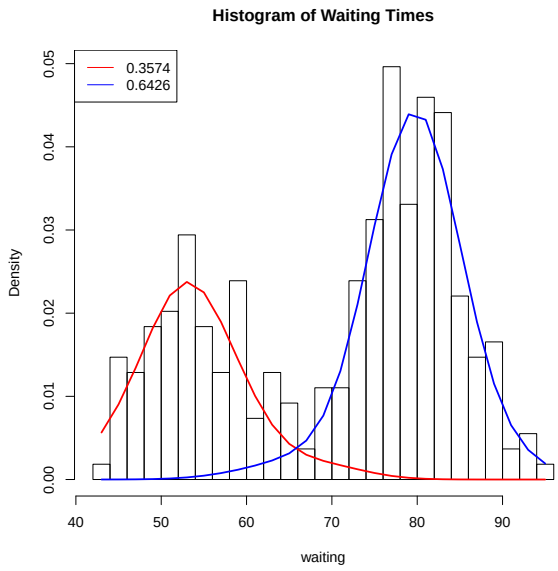


Fig. 2.10. The semiparametric component density estimates for the waiting times for eruptions for Old Faithful.

densities, the mixing proportions, and the component means and standard deviations. This is also the case when the exponential tilt assumption was not valid for the model we chose. This seems to have an advantage over the normal mixture method as well as the nonparametric method.

There is one limitation to extending this method to the multivariate case. Suppose we were considering a two-component, three-coordinate mixture model. Suppose as with the simulations in this chapter, we choose 30 breaks for each coordinate. Then we would have 30^3 cells, a problem that R simply cannot handle. We could use another computer language, such as C, to handle that kind of data.

With this limitation, we wanted to develop another method that could handle the exponential tilt assumption but also be able to incorporate the multivariate case. We wanted the method to be less computationally intensive than the one described in this chapter. We present this method in the next chapter.

Chapter 3

Multivariate Mixtures of Exponentially Tilted Models

t

3.1 Introduction

In this chapter we propose a method for estimating the component CDFs (and component marginal CDFs) for a class of semiparametric multivariate mixture models. Similar to Chapter 2, the component density functions are unspecified but are assumed to be related by an exponential tilt. For the parameters in a multivariate mixture model to be identifiable, assumptions must be made (see Section 1.2). For our model, we assume the coordinates (or repeated measure) are independent conditioned on the component membership.

Our method is a generalization of the estimation method of Leung and Qin (2006). We improve their method by relaxing the assumption of conditional i.i.d. coordinates to conditional independence. This is an important generalization because it gives the model more flexibility. For example, in experimental psychology, the assumption of conditional i.i.d. repeated measures may not be valid. Our method also handles an arbitrary number of components ($m > 1$) and for an arbitrary number of coordinates (or repeated measures) ($k > 1$). To find the estimates of the exponential tilt parameters, we develop an EM algorithm. It is, however, difficult to maximize the log likelihood in our model. Instead, we maximize the log profile likelihood since the profile likelihood behaves very similarly to the likelihood. We will begin the chapter by describing the method and how to use the EM to find estimates for the parameters of the exponential tilt. Next, we discuss some modifications to handle the independent and identically distributed case or cases that require a blocking structure. In Section 3.7, we present likelihood ratio tests, using the profile likelihood, to test for a particular block structure. We also develop a

model selection method based on the Bayesian Information Criterion (BIC); see Section 3.9. In Section 3.8 we present some simulations to evaluate how the well the method performs. Lastly, real data examples are presented in Sections 3.10.1 and 3.10.2.

3.2 Semiparametric Finite Mixture Model

Suppose there is a set of n multivariate vectors $X_1, \ldots, X_n$ from a finite mixture distribution with $m > 1$ components, where $X_i' = (x_{i1}, \ldots, x_{ik})$, for $i = 1, \ldots, n$. In addition to the conditional independence assumption for the coordinates, we assume that the marginal component densities are related by an exponential tilt. Let $X' = (x_1, \ldots, x_k)$ be a generic observation from a m component, k dimensional multivariate mixture, which has density

$$h(X) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} g_{lj}(x_j), \tag{3.1}$$

where λ_1 represents the mixing proportion of component one (the baseline distribution) and λ_l represents the mixing proportion of component l and $\sum_{l=1}^{m} \lambda_l = 1$. Also, f_j and g_{lj} represent univariate PDFs (or probability mass function, PMF). Let H, F_j, and G_{lj} denote the CDFs corresponding to h, f_j, and g_{lj}, respectively. Finally, let $f = \prod_{j=1}^{k} f_j$, $g_l = \prod_{j=1}^{k} g_{lj}$, $F = \prod_{j=1}^{k} F_j$, and $G_l = \prod_{j=1}^{k} G_{lj}$ be the multivariate PDFs and CDFs of the m components. For sake of clarity, Table 3.2 displays the notation used in this chapter. The notation $l = 1$ refers to the baseline distribution, $f_j(x)$, and its associated parameter, λ_1. Note for model (3.1), other than the exponential tilt relationship between components, there are no other parametric assumptions on the coordinates, hence the semiparametric model.

Following Anderson (1979), let f_j and g_{lj} be related by a quadratic exponential tilt model

$$g_{lj}(x_j) = \exp(\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2) f_j(x_j), \tag{3.2}$$

Component	$l = 2, \ldots, m$
Repeated Measure	$j = 1, \ldots, k$
Observation	$i = 1, \ldots, n$

Table 3.1. Notation for the Semiparametric Finite Mixture Model

where α_{lj}, β_{lj} and γ_{lj} are the unknown exponential tilt parameters. Note that the model can be extended to include higher order terms but in this thesis we consider only to the order two. The PDF (3.1) can be re-written as

$$h(X) = \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\sum_{j=1}^{k} \alpha_{lj} + \sum_{j=1}^{k} \beta_{lj} x_j + \sum_{j=1}^{k} \gamma_{lj} x_j^2\right\}\right] \prod_{j=1}^{k} f_j(x_j). \tag{3.3}$$

Theorem 8 of Allman, Matias, and Rhodes (2009) states that the parameters in (3.1) are uniquely identifiable up to label switching provided that $k \geq 3$ and, for each $j = 1, \ldots, k$, the m distributions $\{f_{lj}\}_{1 \leq l \leq m}$ are linearly independent (see Section 1.2). This result makes sense since linear independence precludes expressing any one of the coordinate distributions as a linear combination of the other $m - 1$ distributions. Since, in our case, in an m-component mixture, $\sum \lambda_l = 1$ and $\lambda_l > 0$, and for each $j = 1, \ldots, k$

$$\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{\alpha_{lj} + \beta_{lj} x_j + \gamma_{lj} x_j^2\right\} \neq 0 \text{ for } -\infty < x_j < \infty$$

identifiability follows for the parameters in (3.3).

Let $\theta'_{lj} \equiv (\alpha_{lj}, \beta_{lj}, \gamma_{lj})$ and write $\tilde{x}'_{ij} \equiv (1, x_{ij}, x_{ij}^2)$ and $\delta' = (\lambda_1, \ldots, \lambda_m, \theta_{21}, \ldots, \theta_{mk})$. Then the likelihood based on the observed data is

$$L(\delta, F_1, \ldots, F_k) = \prod_{i=1}^{n} \left[\left\{\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lj}\right)\right\} \prod_{j=1}^{k} dF_j(x_{ij})\right].$$

The maximizing F_j only jumps at each observed x_{ij} (Owen, 1988). Let the jump sizes be p_{ij}. Then the log likelihood is

$$\ell(\delta, p_{ij}, i \in [1:n],\ j \in [1:k]) = \sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lj} \right) \right\} + \sum_{j=1}^{k} \sum_{i=1}^{n} \log p_{ij}.$$

For fixed δ, $\ell(\delta, p_{ij}; i \in [1:n], j \in [1:k])$ can be maximized with respect to the p_{ij}s subject to the constraints

$$\sum_{i=1}^{n} p_{ij} - 1 = 0, \quad p_{ij} \geq 0, \quad \sum_{i=1}^{n} p_{ij} \left\{ \exp\left(\tilde{x}'_{ij} \theta_{lj} \right) - 1 \right\} = 0. \tag{3.4}$$

for $j = 1, \ldots, k$, $l = 2, \ldots, m$, and where $\theta_{1j} = 0$.

The last k constraints in (3.4) come from the exponential tilt model (3.2) and are responsible for ensuring that the resulting g_{lj} are proper PDFs. The constrained maximization can be accomplished using a Lagrange multiplier argument (Leung and Qin, 2006). Therefore, we maximize the following likelihood with respect to p_{ij}.

$$\begin{aligned} \ell = & \sum_{i=1}^{n} \log \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lj} \right) \right] + \sum_{i=1}^{n} \sum_{j=1}^{k} \log(p_{ij}) \\ & - \sum_{j=1}^{k} \eta_{1j} \left(\sum_{i=1}^{n} p_{ij} - 1 \right) - \sum_{l=2}^{m} \sum_{j=1}^{k} \eta_{lj} \left[\sum_{i=1}^{n} p_{ij} \left\{ \exp\left(\tilde{x}'_{ij} \theta_{lj} \right) - 1 \right\} \right] \end{aligned} \tag{3.5}$$

The derivatives with respect to p_{ij}, η_{1j}, and η_{lj} are shown below:

$$\frac{\partial \ell}{\partial p_{ij}} = \frac{1}{p_{ij}} - \eta_{1j} - \sum_{l=2}^{m} \eta_{lj} \left\{ \exp\left(\tilde{x}'_{ij} \theta_{lj} \right) - 1 \right\} \tag{3.6}$$

$$\frac{\partial \ell}{\partial \eta_{1j}} = 0 \Rightarrow \sum_{i=1}^{n} p_{ij} = 1 \tag{3.7}$$

$$\frac{\partial \ell}{\partial \eta_{lj}} = 0 \Rightarrow \sum_{i=1}^{n} p_{ij} \left\{ \exp\left(\tilde{x}'_{ij} \theta_{lj} \right) - 1 \right\} = 0. \tag{3.8}$$

Setting (3.6) equal to zero and solving for p_{ij} leads to

$$p_{ij} = \frac{1}{\eta_{1j} + \sum_{l=2}^{m} \eta_{lj} \left\{ \exp(\tilde{x}'_{ij} \theta_{lj}) - 1 \right\}}, \tag{3.9}$$

where $\eta \equiv (\eta_{21}, \ldots, \eta_{mk})$ are Lagrange multipliers determined by the equations

$$\sum_{i=1}^{n} \frac{\exp(\tilde{x}'_{ij} \theta_{lj}) - 1}{1 + \frac{1}{n} \sum_{l=2}^{m} \eta_{lj} \{ \exp(\tilde{x}'_{ij} \theta_{lj}) - 1 \}} = 0 \tag{3.10}$$

where $l = 1, \ldots, m$ and $j = 1, \ldots, k$.

Using (3.7) with (3.9) gives

$$\eta_{ij} = n - \sum_{j=1}^{k} \eta_{jl}. \tag{3.11}$$

Putting (3.11) and (3.9) together, we obtain

$$p_{ij} = \frac{1}{n} \left(\frac{1}{1 + \frac{1}{n} \sum_{l=2}^{m} \eta_{lj} \{ \exp(\tilde{x}'_{ij} \theta_{lj}) - 1 \}} \right). \tag{3.12}$$

Note that if the exponential tilt parameters $\theta'_{lj} = (0, 0, 0)$, then (3.12) would simply be the weights found for the empirical distribution, namely $1/n$.

Substituting the p_{ij}s back into the log likelihood gives the following semiparametric log profile likelihood

$$\begin{aligned} \ell_P(\delta) = & \sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp \left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lj} \right) \right\} \\ & - \sum_{j=1}^{k} \sum_{i=1}^{n} \log \left[1 + \frac{1}{n} \sum_{l=2}^{m} \eta_{lj} \left\{ \exp(\tilde{x}'_{ij} \theta_{lj}) - 1 \right\} \right] - kn \log(n). \end{aligned} \tag{3.13}$$

The underlying vector of parameters can be estimated by maximizing $\ell_P(\delta)$ with respect to δ. Denote the maximum semiparametric likelihood estimate, $\hat{\delta}$; see the next section. Assuming a sufficiently smooth model for f_j, $j = 1, \ldots, k$ in (3.3), the results on asymptotic theory of profile likelihoods in Qin (1999) and Murphy and van der Vaart (2000) imply that $\hat{\delta}$ is consistent and asymptotically normally distributed.

3.3 The EM Algorithm

In general it is difficult to directly maximize the empirical likelihood. In this chapter, we adapt the EM algorithm (Dempster et al., 1977) developed for the full parametric model to the semiparametric model (3.3). Let the d_{il} be a Bernoulli random variable indicating that observation i comes from component l and $\sum_{l=1}^{m} d_{il} = 1$. Note that d_{i1} will be 1 if the ith observation comes from F, the baseline distribution, and 0 otherwise. In practice, d_{il} is not observed. However, if $(d_{i1}, \ldots, d_{im}, x_i)$, for $i = 1, ..., n$ were observed, then the complete log likelihood would be

$$\ell_c(\delta) = \sum_{i=1}^{n}\sum_{l=1}^{m} d_{il} \log \lambda_l + \sum_{i=1}^{n}\sum_{j=1}^{k} \log p_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=2}^{m} d_{il}\tilde{x}'_{ij}\theta_{lj}.$$

Given the current estimate $\delta^{(t)}$ and conditioning on the observed data gives

$$\begin{aligned}\mathbb{E}_{\delta^{(t)}}\{\ell_c(\delta)|\text{data}\} = &\sum_{i=1}^{n}\sum_{l=1}^{m} w_{il}^{(t)} \log \lambda_l^{(t)} + \sum_{j=1}^{k}\sum_{i=1}^{n} \log p_{ij} \\ &+ \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=2}^{m} w_{il}^{(t)}\tilde{x}'_{ij}\theta_{lj}^{(t)}, \end{aligned} \tag{3.14}$$

where

$$w_{il}^{(t)} = \frac{\lambda_l^{(t)} \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij}\theta_{lj}^{(t)}\right)}{\sum_{l=1}^{m} \lambda_l^{(t)} \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij}\theta_{lj}^{(t)}\right)} \tag{3.15}$$

is the conditional probability of $d_{il} = 1$ given the data, for $l = 1, \ldots, m$. By imposing the constraints

$$\sum_{i=1}^{n} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \sum_{i=1}^{n} p_{ij}\left\{\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t)}\right) - 1\right\} = 0,$$

the profiled log likelihood, given $w_{il}^{(t)}$ is

$$\ell_P(\delta) = \sum_{i=1}^{n}\sum_{l=1}^{m} w_{il}^{(t)} \log \lambda_l + \sum_{j=1}^{k}\sum_{i=1}^{n}\sum_{l=2}^{m} w_{il}^{(t)} \tilde{x}'_{ij}\theta_{lj} - \sum_{j=1}^{k}\sum_{i=1}^{n} \log\left[1 + \frac{1}{n}\sum_{l=2}^{m}\eta_{lj}\{\exp(\tilde{x}'_{ij}\theta_{lj}) - 1\}\right] - nk\log(n). \tag{3.16}$$

Given $w_{il}^{(t)}$, maximizing ℓ_P with respect to λ_l and θ_{lj} yields

$$\frac{\partial \ell_P}{\partial \lambda_l} = 0 \Rightarrow \lambda_l^{(t+1)} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n}, \tag{3.17}$$

$$\frac{\partial \ell_P}{\partial \alpha_{lj}} = \sum_{i=1}^{n} w_{il}^{(t)} - \sum_{i=1}^{n} \frac{\frac{1}{n}\eta_{lj}\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right)}{1 + \frac{1}{n}\sum_{l=2}^{m}\eta_{lj}\left\{\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right) - 1\right\}} = 0 \tag{3.18}$$

$$\frac{\partial \ell_P}{\partial \beta_{lj}} = \sum_{i=1}^{n} w_{il}^{(t)} x_{ij} - \sum_{i=1}^{n} \frac{\frac{1}{n}\eta_{lj}\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right) x_{ij}}{1 + \frac{1}{n}\sum_{l=2}^{m}\eta_{lj}\left\{\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right) - 1\right\}} = 0 \tag{3.19}$$

$$\frac{\partial \ell_P}{\partial \gamma_{lj}} = \sum_{i=1}^{n} w_{il}^{(t)} x_{ij}^2 - \sum_{i=1}^{n} \frac{\frac{1}{n}\eta_{lj}\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right) x_{ij}^2}{1 + \frac{1}{n}\sum_{l=2}^{m}\eta_{lj}\left\{\exp\left(\tilde{x}'_{ij}\theta_{lj}^{(t+1)}\right) - 1\right\}} = 0 \tag{3.20}$$

for $l = 2, \ldots, m$ and $j = 1, \ldots, k$. If we consider (3.18) and (3.12) we obtain

$$n\left(\frac{\eta_{lj}^{(t+1)}}{n}\right) - \sum_{i=1}^{n} w_{il}^{(t)} = 0 \Rightarrow \frac{\eta_{lj}^{(t+1)}}{n} = \frac{\sum_{i=1}^{n} w_{il}^{(t)}}{n} = \lambda_l^{(t+1)} \tag{3.21}$$

The above estimating equations can be solved numerically to obtain $\theta_{lj}^{(t+1)}$ for $l = 2, \ldots, m$ and $j = 1, \ldots, k$. Alternatively, $\lambda_l^{(t+1)}$ and $\frac{1}{n}\eta_{lj}^{(t+1)} = \lambda_l^{(t+1)}$ may be

substituted in l_P, which can then be maximized with respect to θ_{lj} by using the simplex method, which can be found in many commonly used software packages, such as the `optim` function in R (R Development Core Team (2009)). The updated values $\delta^{(t+1)}$ are then substituted back into (3.14) and the algorithm iterated until convergence. We recommend using different starting values and comparing log profile likelihood values to check that the algorithm does not stop at a local maximum. The algorithm stops when the value of the log profile likelihood does not change more than a predetermined value. Since the exponential tilt parameters may be difficult to interpret, it may be challenging to find initial values for them at the start of the algorithm. We recommend generating an $n \times m$ matrix of initial values of the $w_{il}^{(t)}$s and starting the algorithm at the M-step. The initial matrix may be generated randomly or found using a clustering function. In the software `R`, there is a clustering function, `kmeans`, that we use in our function.

3.4 Modifications of the General Model and Algorithm

In this section we upgrade the model and the EM algorithm to make the method a bit more flexible. Benaglia et al. (2008) incorporated into their model a *block* structure. This block structure will allow for models where some, or all, of the repeated measures are conditionally i.i.d. This added feature of the model is quite useful. In the water-level example, described in Section 3.10.2, the angular tilt of the vessel is the same, say for clock orientations 11 and 5 o'clock. This may indicate that the measurements come from the same distribution. Therefore, we will have the ability to to allow some of the coordinates to be conditionally independent and others to be conditionally independent and identically distributed. It also gives the advantage of using more of the data to estimate the unknown parameters in the model. In the following section, we adjust the model and the EM algorithm to include the block structure. In Section 3.8, we perform simulations to show the performance of the method when the model contains the block structure.

3.4.1 Modeling with a Block Structure

This section is basically an analog to Section 3.2. In this section, however, we upgrade the model and method to handle the block structure. The notation in this section changes. Again, for sake of clarity, the notation is given in the table below (Table 3.2).

Table 3.2. Notation of the model with the block structure.

Component	$l = 1, \ldots, m$
Repeated Measure	$j = 1, \ldots, k$
Observation	$i = 1, \ldots, n$
Block	$a = 1, \ldots, B$

Let $X' = (x_1, \ldots, x_k)$ be an observation from the following multivariate mixture,

$$h(X) = \lambda_1 \prod_{j=1}^{k} f_{b_j}(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} g_{lb_j}(x_j) \tag{3.22}$$

where the notation b_j denotes the block to which the j-th repeated measure (coordinate) belongs, where $1 \leq b_j \leq B$ and B is the total number of blocks. If $b_j = j$ for all $j \in 1, \ldots, k$, then we have the independence case as is Section 3.2. If $b_j = 1$ for all j, then we would have the case where all the repeated measurements are conditionally independent and identically distributed. Similarly, $g_{lb_j} = f_{b_j}(x_{ij}) \exp(\alpha_{lb_j} + \beta_{lb_j} x_j + \gamma_{lb_j} x_j^2)$. The density can be rewritten as

$$h(X) = \left[\lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left\{ \sum_{j=1}^{k} \alpha_{lb_j} + \sum_{j=1}^{k} \beta_{lb_j} x_j + \sum_{j=1}^{k} \gamma_{lb_j} x_j^2 \right\} \right] \prod_{j=1}^{k} f_{b_j}(x_j). \tag{3.23}$$

The likelihood based on the observed data is

$$L(\delta) = \prod_{i=1}^{n} \left[\left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lb_j} \right) \right\} \prod_{j=1}^{k} dF_{b_j}(x_{ij}) \right] \tag{3.24}$$

where $\theta'_{lb_j} = (\alpha_{lb_j}, \beta_{lb_j}, \gamma_{lb_j})$ and $\delta' = (\lambda_1, \ldots, \lambda_m, \theta_{l1}, \ldots, \theta_{lB})$. If we let the jumps size be p_{ib_j} then the log likelihood becomes

$$\ell(\delta) = \sum_{i=1}^{n} \log \left\{ \lambda_1 + \sum_{l=2}^{m} \lambda_l \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \theta_{lb_j} \right) \right\} + \sum_{i=1}^{n} \sum_{j=1}^{k} \log(p_{ib_j}). \tag{3.25}$$

For fixed δ, $\ell(\delta)$ can be maximized with respect to p_{ib_j} subject to the following constraints:

$$\sum_{i=1}^{n} p_{ib_j} = 1, \quad p_{ib_j} \geq 0, \quad \sum_{i=1}^{n} p_{ib_j} \left[\exp(\tilde{x}'_{ij} \theta_{ib_j}) \right] = 0 \tag{3.26}$$

for $1 \leq b_j \leq B$.

From this point on, the notation in the section becomes messy. We will describe the notation and what each variable represents in as much detail as possible. There is at least one coordinate associated with each of the 1 through B blocks. We will denote the number of coordinates in the ath block as $\mathcal{C}_a = \sum_{j=1}^{k} I\{b_j = a\}$, for $a = 1, \ldots, B$. For the calculations, we concatenate the coordinates such that $b_j = a$ for each a. For example, suppose we have a model with four coordinates. If coordinate 1 and coordinate 2 are conditionally i.i.d., then for that block, $\mathcal{C}_a = 2$. Once we concatenate the coordinates, the new notation becomes

$$y'_{c_a a} = (x_{11}, \ldots, x_{n1}, x_{12}, \ldots, x_{n2})$$

where $c_a = 1, \ldots, n\mathcal{C}_a$. Similarly, as in the non-block notation, we let

$$\tilde{y}'_{c_a a} = (1, y_{c_a a}, y^2_{c_a a}). \tag{3.27}$$

Using this notation, we estimate the jumps as:

$$p_{c_a a} = \frac{1}{nC_a} \left[\frac{\exp\left(\tilde{y}'_{c_a a} \theta_{la} \right)}{1 + \sum_{l=2}^{m} \lambda_l \left\{ \exp\left(\tilde{y}'_{c_a a} \theta_{la} \right) - 1 \right\}} \right]. \tag{3.28}$$

For calculations, we will need different notation for the estimated posterior probabilities. Therefore, when we use $w_{il}^{(t)}$, we will refer to the usual definition as shown in

Section 3.3. When we use $w^{(t)}_{c_a l}$ it will denote an observation in the concatenated version where we concatenate the original $w^{(t)}_{il}$, $\mathcal{C}_a$ times. In the example above with two coordinates in the block, the concatenated version of the posterior probabilities would be

$$(w_{il}, \ldots, w_{nl}, w_{il}, \ldots, w_{nl}).$$

We start the EM algorithm with an initial $n \times m$ matrix of posterior probabilities, $\mathbf{W}^{(0)}$, where $w^{(0)}_{il}$ is an initial posterior probability of the i-th observation and the l-th component. Recall that $\sum_{l=1}^{m} w^{(t)}_{il} = 1$ for all $i \in [1, n]$ and for all t. Then we calculate the following parameters for the M-step:

- The mixing proportions, $l = 1, \ldots, m$

$$\hat{\lambda}^{(t+1)}_{l} = \frac{\sum_{i=1}^{n} \hat{w}^{(t)}_{il}}{n} \tag{3.29}$$

- The exponential tilt parameter, $\hat{\alpha}^{(t+1)}_{la}$, is the solution to:

$$\sum_{c_a=1}^{n\mathcal{C}_a} w^{(t)}_{c_a l} - \sum_{c_a=1}^{n\mathcal{C}_a} \left[\frac{\lambda^{(t)}_{l} \exp\left(\tilde{y}'_{c_a a} \hat{\theta}^{(t)}_{la} \right)}{1 + \sum_{l=2}^{m} \lambda^{(t)}_{l} \left\{ \exp\left(\tilde{y}'_{c_a a} \hat{\theta}^{(t)}_{la} \right) - 1 \right\}} \right] = 0 \tag{3.30}$$

- The exponential tilt parameter, $\hat{\beta}^{(t+1)}_{la}$, is the solution to:

$$\sum_{c_a=1}^{n\mathcal{C}_a} w^{(t)}_{c_a l} y_{c_a a} - \sum_{c_a=1}^{n\mathcal{C}_a} \left[\frac{\lambda^{(t)}_{l} y_{c_a a} \exp\left(\tilde{y}'_{c_a a} \hat{\theta}^{(t)}_{la} \right)}{1 + \sum_{l=2}^{m} \lambda^{(t)}_{l} \left\{ \exp\left(\tilde{y}'_{c_a a} \hat{\theta}^{(t)}_{la} \right) - 1 \right\}} \right] = 0 \tag{3.31}$$

- The exponential tilt parameter, $\hat{\gamma}_{la}^{(t+1)}$ is the solution to:

$$\sum_{c_a=1}^{n\mathcal{C}_a} w_{c_a l}^{(t)} y_{c_a a}^2 - \sum_{c_a=1}^{n\mathcal{C}_a} \left[\frac{\lambda_l^{(t)} y_{c_a a}^2 \exp\left(\tilde{y}'_{c_a a} \hat{\theta}_{la}^{(t)} \right)}{1 + \sum_{l=2}^{m} \lambda_l^{(t)} \left\{ \exp\left(\tilde{y}'_{c_a a} \hat{\theta}_{la}^{(t)} \right) - 1 \right\}} \right] = 0. \tag{3.32}$$

Once the exponential tilt parameters are found for each block, we set $\hat{\theta}_{lj} = \hat{\theta}_{la}$ if $b_j = a$ for $l = 1, \ldots, m$, $j = 1, \ldots, k$, and $a = 1, \ldots, B$. After the above are found, we proceed to the E-step to find the estimated posterior probabilities:

$$\hat{w}_{il}^{(t+1)} = \frac{\hat{\lambda}_l^{(t)} \exp\left(\sum_{j=1}^{k} \tilde{x}'_{ij} \hat{\theta}_{lj}^{(t)} \right)}{\sum_{l'=1}^{m} \hat{\lambda}_{l'}^{(t)} \exp\left(\tilde{x}'_{ij} \hat{\theta}_{l'j}^{(t)} \right)} \tag{3.33}$$

where $\hat{\theta}_{1j} = 0$ for all $j = 1, \ldots, k$.

The steps are then iterated until the change in the log profile likelihood is no greater than a predetermined convergence criterion. Similarly to Section 3.3, we recommend using different starting values and comparing the log profile likelihoods as a check that the algorithm does not stop at a local maximum. We find the easiest way to start the algorithm is to provide an initial $n \times m$ matrix of posterior probabilities. The posterior probabilities are then concatenated to fit the number of components in each block for the above calculations in the EM.

3.5 Estimation of features in the component distributions

3.5.1 Conditionally independent case

In this section, we discuss estimation of features in the component distributions. We also identify a moment matching property similar to one found by Efron and Tibshirani (1996) for the univariate non-mixture case and in the univariate mixture case as discussed in Section 2.2.4.

Let the final value of the EM-algorithm upon convergence, $\delta^{(t+1)}$, be $\hat{\delta}$. Once we have the estimates for the parameters, we can use them to find the estimate of the component CDFs. Define the estimates to be

$$\hat{p}_{ij} = \frac{1}{n}\left(\frac{1}{1+\sum_{l=2}^{m}\hat{\lambda}_l\{\exp(\tilde{x}'_{ij}\hat{\theta}_{lj})-1\}}\right)$$

and

$$\hat{q}_{lij} = \frac{1}{n}\left(\frac{\exp(\tilde{x}'_{ij}\hat{\theta}_{lj})}{1+\sum_{l=2}^{m}\hat{\lambda}_l\{\exp(\tilde{x}'_{ij}\hat{\theta}_{lj})-1\}}\right).$$

The estimates resemble the empirical CDF with the weights given by the estimated jumps. If the exponential tilt parameters $\theta_{lj} = 0$ for $l = 2, \ldots, m$ then $\hat{q}_{lij} = \frac{1}{n}$.

The CDF of the mixture distribution, H, can be estimated by

$$\hat{H}(x_1, \ldots, x_k) = \hat{\lambda}_1\hat{F}(x_1, \ldots, x_j) + \sum_{l=2}^{m}\hat{\lambda}_l\hat{G}_l(x_1, \ldots, x_j),$$

where

$$\hat{F}(x_1, \ldots, x_k) = \prod_{j=1}^{k}\left(\sum_{i=1}^{n} I(x_{ij} \le x_j)\hat{p}_{ij}\right),$$

$$\hat{G}_l(x_1, \ldots x_k) = \prod_{j=1}^{k}\left(\sum_{i=1}^{n} I(x_{ij} \le x_j)\hat{p}_{ij}\exp\left(\tilde{x}'_{ij}\hat{\theta}_{lj}\right)\right)$$

are the estimated CDFs of the component distributions. Furthermore, the marginal CDFs, F_j and G_{lj}, can be estimated by

$$\hat{F}_j(x_j) = \sum_{i=1}^{n} I(x_{ij} \leq x_j)\hat{p}_{ij} \tag{3.34}$$

$$\hat{G}_{lj}(x_j) = \sum_{i=1}^{n} I(x_{ij} \leq x_j)\hat{p}_{ij} \exp(\tilde{x}'_{ij}\hat{\theta}_{lj}). \tag{3.35}$$

In Section 3.8, we show examples of the estimated component marginal CDFs. They are very close to the true CDFs.

We can also find estimates of the marginal PDFs. In the R package `mixtools` there is a function, weighted kernel density estimate (`wkde`), that allows us to do this quite easily (see Young et al., 2008). The function calculates a kernel density estimate with weights at each of the observations. We can use this function with posterior probabilities, w_{il}, as the weights. The estimated PDFs are

$$\hat{g}_{lj}(u) = \frac{1}{h}\sum_{i=1}^{n} \frac{w_{il}}{\sum_{i'=1}^{n} w_{i'l}} \Phi\left(\frac{u - x_{ij}}{h}\right), \text{ for } l = 1, \ldots, m \tag{3.36}$$

for $l = 1, \ldots, m$, $j = 1, \ldots, k$, where h is a bandwidth, and $\Phi(\cdot)$ is the standard normal PDF.

An interesting result from the EM algorithm is a moment matching property. For the general model (3.22), if we differentiate the log profile likelihood (3.25) with respect to β_{lj} and γ_{lj}, we obtain, from (3.31), (3.73), and (3.29),

$$\frac{\partial \ell_P}{\partial \beta_{lj}} = 0 \Rightarrow \sum_{i=1}^{n} w_{il}x_{ij} - \sum_{i=1}^{n} \frac{\lambda_l x_{ij} \exp\left(\tilde{x}'_{ij}\theta_{lj}\right)}{1 + \sum_{l=2}^{m} \lambda_l \left(\exp(\tilde{x}'_{ij}\theta_{lj}) - 1\right)} = 0, \tag{3.37}$$

$$\frac{\partial \ell_P}{\partial \gamma_{lj}} = 0 \Rightarrow \sum_{i=1}^{n} w_{il}x_{ij}^2 - \sum_{i=1}^{n} \frac{\lambda_l x_{ij}^2 \exp\left(\tilde{x}'_{ij}\theta_{lj}\right)}{1 + \sum_{l=2}^{m} \lambda_l \left(\exp(\tilde{x}'_{ij}\theta_{lj}) - 1\right)} = 0. \tag{3.38}$$

Combining (3.37) and (3.38) with (3.33) gives

$$\sum_{i=1}^{n} x_{ij} q_{lij} = \frac{\sum_{i=1}^{n} w_{il} x_{ij}}{\sum_{i=1}^{n} w_{il}}, \tag{3.39}$$

$$\sum_{i=1}^{n} x_{ij}^2 q_{lij} = \frac{\sum_{i=1}^{n} w_{il} x_{ij}^2}{\sum_{i=1}^{n} w_{il}}. \tag{3.40}$$

The last two equations match the weighted moments using the posterior probabilities to the tilted component moments; see Efron and Tibshirani (1996) for an example in the univariate non-mixture case and Section 2.2.4 for the univariate mixture case.

The component moments are estimated using the following equations,

$$\hat{\mu}_{lj} = \sum_{i=1}^{n} x_{ij} \hat{q}_{lij}, \tag{3.41}$$

$$\hat{\sigma}_{lj}^2 = \sum_{i=1}^{n} x_{ij}^2 \hat{q}_{lij} - (\hat{\mu}_{lj})^2, \tag{3.42}$$

for $l = 1, \ldots, m$ and $j = 1, \ldots, k$. Note that $q_{1ij} = p_{ij}$.

3.5.2 Case with blocks of conditionally i.i.d. coordinates

This section is an analog to the previous section for the case when we have B blocks of conditionally i.i.d. coordinates. Let the final value of the EM-algorithm upon convergence, $\delta^{(t+1)}$, be $\hat{\delta}$. Once we have the estimates for the parameters we can use those to find the estimate of the component CDFs. Define the estimates to be

$$\hat{p}_{c_a a} = \frac{1}{n\mathcal{C}_a} \left(\frac{1}{1 + \sum_{l=2}^{m} \hat{\lambda}_l \{\exp(\tilde{y}'_{c_a a} \hat{\theta}_{la}) - 1\}} \right)$$

and

$$\hat{q}_{lc_a a} = \frac{1}{n\mathcal{C}_a}\left(\frac{\exp(\tilde{y}'_{c_a a}\hat{\theta}_{la})}{1+\sum_{l=2}^{m}\hat{\lambda}_l\{\exp(\tilde{y}'_{c_a a}\hat{\theta}_{la})-1\}}\right).$$

where $\mathcal{C}_a = \sum_{j=1}^{k} I\{b_j = a\}$ is the number of coordinates that belong to each block, $a = 1,\ldots,B$, B is the total number of blocks, and $c_a = 1,\ldots,n\mathcal{C}_a$. The estimates resemble the empirical CDF with the weights given by the estimated jumps. If the exponential tilt parameters $\theta_{la} = 1$ for $l = 2,\ldots,m$ then $\hat{q}_{lc_a a} = \frac{1}{n\mathcal{C}_a}$.

The CDF of the mixture distribution, H, and the marginal CDFs, F_a, and G_a can be estimated by

$$\hat{H}(y_1,\ldots,y_B) = \hat{\lambda}_1\hat{F}(y_1,\ldots,y_B) + \sum_{l=2}^{m}\hat{\lambda}_l\hat{G}_l(y_1,\ldots,y_B),$$

where

$$\hat{F}(y_1,\ldots,y_B) = \prod_{a=1}^{B}\left(\sum_{c_a=1}^{n\mathcal{C}_a} I(y_{c_a a}\le y_a)\hat{p}_{c_a a}\right),$$

$$\hat{G}_l(y_1,\ldots,y_B) = \prod_{a=1}^{B}\left(\sum_{c_a=1}^{n\mathcal{C}_a} I(y_{c_a a}\le y_a)\hat{p}_{c_a a}\exp\left(\tilde{y}'_{c_a a}\hat{\theta}_{la}\right)\right)$$

and

$$\hat{F}_a(y_a) = \sum_{c_a=1}^{n\mathcal{C}_a} I(y_{c_a a}\le y_a)\hat{p}_{c_a a} \tag{3.43}$$

$$\hat{G}_{la}(y_a) = \sum_{c_a=1}^{n\mathcal{C}_a} I(y_{c_a a}\le y_a)\hat{p}_{c_a a}\exp(\tilde{y}'_{c_a a}\hat{\theta}_{la}). \tag{3.44}$$

Similarly to the conditionally independence case, we show examples in Section 3.8 of the estimated component marginal CDFs and they are close the the true CDFs.

We are also able to use the weighted kernel density estimate, `wkde`, in the R package in `mixtools` to estimate the marginal PDFs using the posterior probabilities, $w_{c_a l}$ as the weights. The estimated PDFs are

$$\hat{g}_{la}(u) = \frac{1}{h} \sum_{c_a=1}^{n\mathcal{C}_a} \frac{w_{c_a l}}{\sum_{c'_a}^{n\mathcal{C}_a} w_{c'_a l}} \Phi\left(\frac{u - y_{c_a a}}{h}\right), \text{ for } l = 1, \ldots, m \tag{3.45}$$

where h is a bandwidth, $\Phi(\cdot)$ is the standard normal PDF.

Finally, incorporating the block structure into the model still retains the moment matching property discussed in the previous section. The steps are similar to the conditionally independent case. The component moments are estimated using the following equations,

$$\hat{\mu}_{la} = \sum_{c_a=1}^{n\mathcal{C}_a} y_{c_a a} \hat{q}_{l c_a a}, \tag{3.46}$$

$$\hat{\sigma}^2_{la} = \sum_{c_a=1}^{n\mathcal{C}_a} y^2_{c_a a} \hat{q}_{l c_a a} - (\hat{\mu}_{la})^2, \tag{3.47}$$

for $l = 1, \ldots, m$ and $j = 1, \ldots, k$. Note that $q_{1 c_a a} = p_{c_a a}$.

3.5.3 Identifiability of the Exponential Tilt Parameters and Label Switching

In section 1.2 we discussed identifiability up to label switching. In this section, we would like to discuss non-identifiability due to label switching of the parameters in the exponential tilt. In McLachlan and Peel (2000) label switching is discussed for the mixture model. While label switching issues apply to our model with regard to the components, there is one more complication. To illustrate this, suppose we have a two component mixture model with three coordinates where the coordinates are conditionally independent ($k = 3$, $m = 2$). For ease of notation we will use the notation in Section 3.3 instead of that of Section 3.4. The likelihood for n observation vectors is:

$$L = \prod_{i=1}^{n} \left[\lambda_1 + (1 - \lambda_1) e^{\tilde{X}'_{i1} \theta_{21}} e^{\tilde{X}'_{i2} \theta_{22}} e^{\tilde{X}'_{i3} \theta_{23}} \right] f_1(x_{i1}) f_2(x_{i2}) f_3(x_{i3}) \tag{3.48}$$

where $\tilde{X}'_{ij} = (1, x_{ij}, x^2_{ij})$, $\theta'_{2j} = (\alpha_{2j}, \beta_{2j}, \gamma_{2j})$ and $j = 1, 2, 3$. The model specifies a common distribution for each coordinate and that the distributions for each component in the mixture are related to this common distribution by an exponential tilt. We call the common distribution the baseline distribution. While the EM algorithm estimates the parameters in the exponential tilt, the function does not, however, always estimate the baseline distribution, f_j, as seen in (3.48).

The likelihood can be rewritten to have the following possible baselines:

$$\prod_{i=1}^{n} \left[\lambda_1 + (1 - \lambda_1) e^{\sum_{j=1}^{k} \tilde{X}'_{ij} \theta_{2j}} \right] f_1(x_{i1}) f_2(x_{i2}) f_3(x_{i3}) \tag{3.49}$$

$$\prod_{i=1}^{n} \left[\lambda_1 e^{\sum_{j=1}^{k} -\tilde{X}'_{ij} \theta_{2j}} + (1 - \lambda_1) \right] e^{\sum_{j=1}^{k} \tilde{X}'_{ij} \theta_{2j}} f_1(x_{i1}) f_2(x_{i2}) f_3(x_{i3}) \tag{3.50}$$

In (3.49), the baseline distribution for coordinate one is $f_1(x)$. It is common to both component distributions. In (3.50), the baseline distribution for coordinate one is $g_1(x) = f_1(x) e^{\tilde{X}'_1 \theta_{21}}$. Note that in both situations the distributions are related by an exponential

tilt, either by $e^{\tilde{X}'_1\theta_{21}}$ or $e^{-\tilde{X}'_1\theta_{21}}$. Hence, the baseline distribution is not identifiable without additional restrictions.

Suppose we generate observations from the two-component conditionally independent normal mixture model

$$h(x_1, x_2, x_3) = 0.3 \prod_{j=1}^{3} \phi\left(x_j; 0, 1\right) + 0.7 \prod_{j=1}^{3} \phi\left(x_j; \mu_j, \sigma_j^2\right), \tag{3.51}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. For the second component, $(\mu_1, \mu_2, \mu_3) = (2, 2.5, 3)$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1.5, 2, 1)$. This model is examined in more detail in Section 3.8.5. If we consider coordinate one, then one possible baseline distribution is $N(0, 1)$ and the parameters in the exponential tilt would be $\theta'_{21} = (-1.5361, 1.3333, 0.1667)$ (see Section 1.4). Another possible baseline would be $N(2, 1.5)$ and the parameters would be $\theta'^{*}_{21} = (1.5361, -1.3333, -0.1667)$. A similar situation could occur for θ_{22} and θ_{23}.

To illustrate this we generated 100 random samples from model (3.51). The results are shown in Table 3.3. The means of estimates tend to be near zero and the standard errors are large. This is due to the non-identifiability of the baseline distribution. Figure 3.1 shows plots of the estimates of the parameters for each coordinate for each simulation. If we focus on coordinate one, there appears to be two groups of numbers, the positive estimates and the negative estimates. One group has estimates corresponding to the case where the baseline is $N(0, 1)$ and the other with the baseline $N(2, 1.5)$. A similar explanation can be given for coordinates two and three.

We propose a way to eliminate this label switching issue. That is to adjust the parameters of the exponential tilts after the EM algorithm converges so that the component with the smallest mixing proportion contains the baseline distributions. In our function, we implemented this calculation. Table 3.4 shows the means and standard errors of the estimates with this new adjustment. Figure 3.2 shows the plots of these new estimates. As you can see from the table, the means are closer to the true estimates and also the standard errors have decreased. In the figure, there no longer appears

Parameter		True	Mean	Std. Error
θ_{21}	α_{21}	-1.5361	0.0221	1.5949
	β_{21}	1.3333	-0.0151	1.4758
	γ_{21}	0.1667	0.0132	0.2139
θ_{22}	α_{22}	-1.9091	0.0886	1.9988
	β_{22}	1.2500	-0.0823	1.4968
	γ_{22}	0.2500	-0.0137	0.2879
θ_{23}	α_{23}	-4.5000	0.2978	5.0330
	β_{23}	3.0000	-0.2671	3.5981
	γ_{23}	0.0000	0.0225	0.5041

Table 3.3. Mean (stdev) of the Semiparametric Estimates Based on 100 Simulations of Sample Size 500 from a Mixture of Normals

to be two groups of estimates. Although the standard errors have decreased with the adjustment, they are still rather large. We do find that for some datasets the function produces estimates that are "far" from the true values. We examined these datasets and estimates in detail and found that these estimates are valid. This means that they do satisfy the constraints in the model.

Whether or not we force the baseline to be the component with the smallest mixing proportion, the estimated component means and standard deviations are the same for both cases. Even though the means and standard deviations are calculated after the EM and are not parameters in the model, the method provides good estimates due to the moment matching property discussed in Section 3.5 (see Table 3.5). We also find that the exponential tilt parameters are difficult to interpret even for a mixture of normal distributions. In Section 3.8, we examine models where the exponential tilt assumption is not valid. In these cases, the parameters are not interpretable. Therefore, for the rest of this thesis, we choose to report the estimated component means and standard deviations instead of the parameters in the exponent.

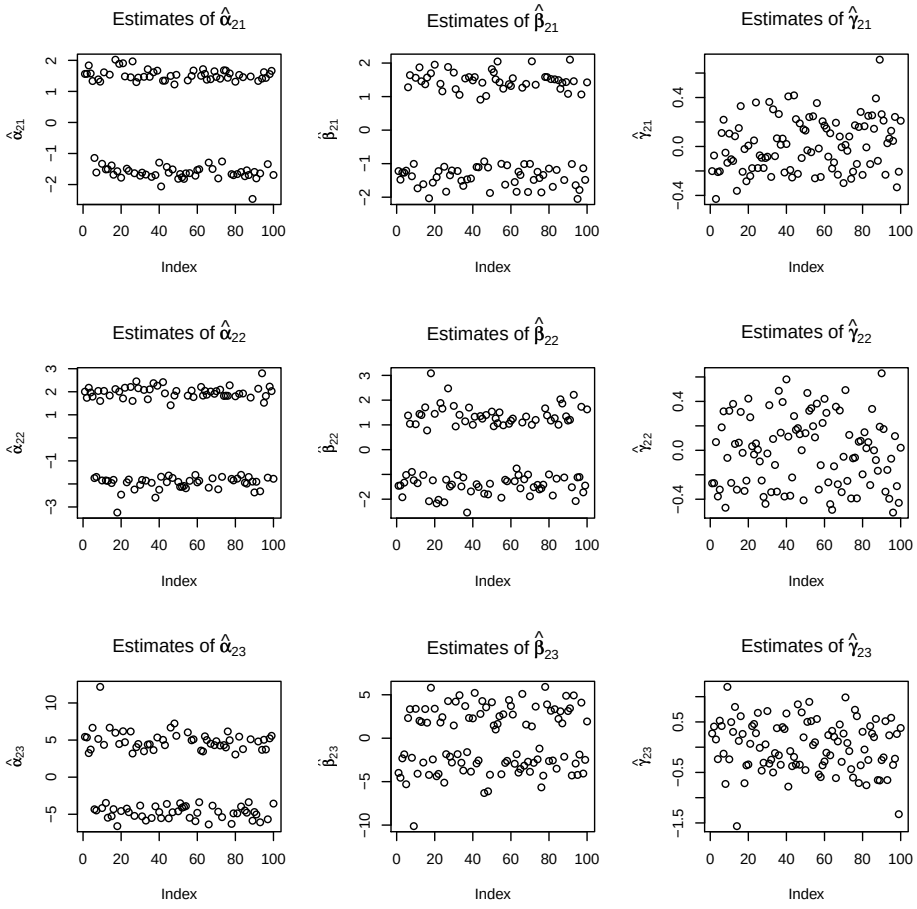


Fig. 3.1. Estimates of the exponential tilt parameters for $n = 100$ simulations the conditionally independent normal mixture described in (3.51)

Parameter		True	Mean	Std. Error
θ_{21}	α_{21}	-1.5361	-1.5749	0.1969
	β_{21}	1.3333	1.4407	0.2855
	γ_{21}	0.1667	0.1559	0.1463
θ_{22}	α_{22}	-1.9091	-1.9723	0.2713
	β_{22}	1.2500	1.4372	0.4013
	γ_{22}	0.2500	0.2096	0.1967
θ_{23}	α_{23}	-4.5000	-4.8761	1.1850
	β_{23}	3.0000	3.3128	1.3903
	γ_{23}	0.0000	-0.0159	0.5044

Table 3.4. Mean (stdev) of the Semiparametric Estimates Based on 100 Simulations of Sample Size 500 from a Mixture of Normals with Adjustments

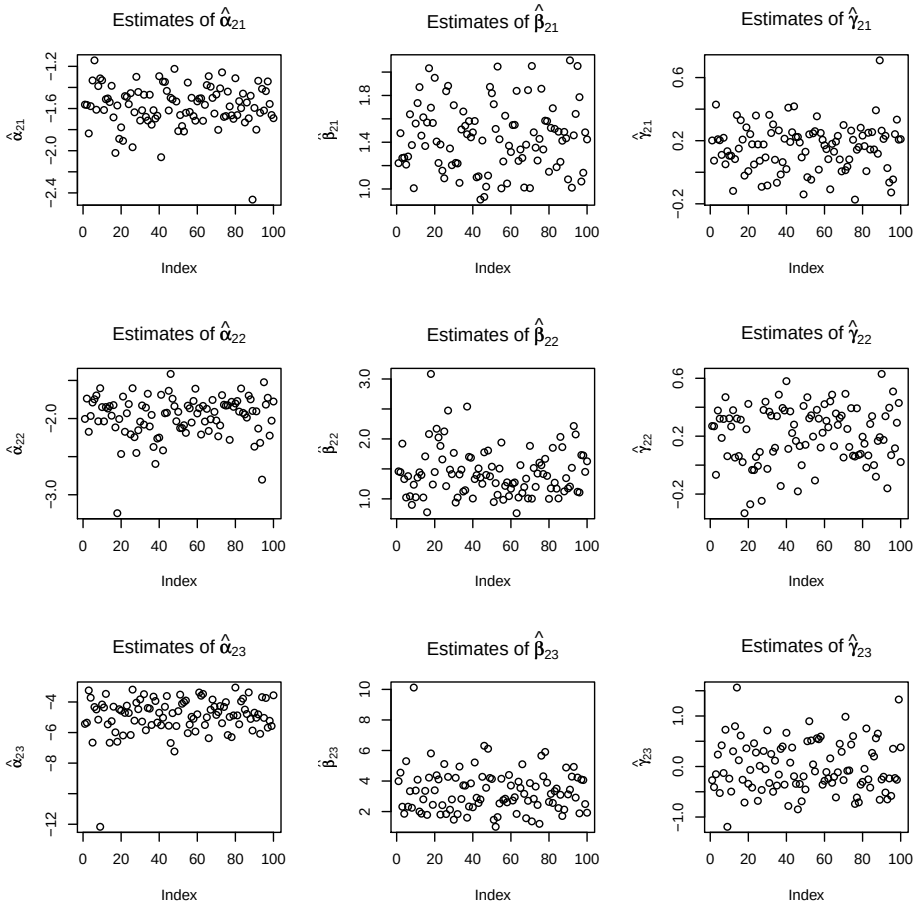


Fig. 3.2. Estimates of the exponential tilt parameters for $n = 100$ simulations the conditionally independent normal mixture described in (3.51) when the baseline is forced to be the component with the smallest mixing proportion.

Parameter	True	Mean	Std. Error
λ	0.3	0.2970	0.0274
μ_{11}	0	-0.0193	0.1088
μ_{12}	0	-0.0037	0.1173
μ_{13}	0	0.0147	0.1389
μ_{21}	2	1.9878	0.0831
μ_{22}	2.5	2.4901	0.0989
μ_{23}	3	3.0089	0.0779
σ_{11}	1	0.9954	0.0769
σ_{12}	1	0.9859	0.0880
σ_{13}	1	1.0177	0.1116
σ_{21}	1.2247	1.2097	0.0545
σ_{22}	1.4142	1.4052	0.0702
σ_{23}	1	0.9959	0.0522

Table 3.5. The estimates of the component means and standard deviations based on 100 simulations from the conditionally independent normal mixture model 3.51.

3.6 Checking the Independence Assumption

3.6.1 Introduction

The most important assumption in this method is that of conditional independence. The assumption means that conditioning on the component membership, the coordinates or repeated measures are independent. If this assumption is not valid, the parameters in the model are not identifiable in general (see Section 1.2). In practice, it is beneficial to have a way of checking this assumption. In this section, we present a technique to examine the validity of this assumption.

The formulas derived in this section are for a finite mixture model with two components ($m = 2$). They may be generalized for an arbitrary number of components ($m > 1$). Let x_j be the jth coordinate of the vector $X = (x_1, \cdots, x_k)$ which comes from the following distribution:

$$h(X) = \lambda f_1(X) + (1 - \lambda) f_2(X) \tag{3.52}$$

where $X \in \mathbb{R}^k$ and $f_l(\cdot)$ is the lth component multivariate density function. If we assume the coordinates are conditionally independent, the model (3.52) becomes:

$$h(X) = \lambda \prod_{j=1}^{k} f_j(x_j) + (1 - \lambda) \prod_{j=1}^{k} g_j(x_j). \tag{3.53}$$

The technique we suggest depends on the calculations of the correlations between the coordinates both with and without the conditional independence assumption. Therefore, we need to find the expected values and variances for both mixture models ((3.52) and (3.53)). For model (3.52), the unconditional expectation and variance equal:

$$\mathbb{E}(x_j) = \lambda \mu_{1j} + (1 - \lambda) \mu_{2j} \tag{3.54}$$

and

$$\mathrm{Var}(x_j) = \lambda \sigma_{1j}^2 + (1 - \lambda) \sigma_{2j}^2 + \lambda(1 - \lambda) \left(\mu_{1j} + \mu_{2j} \right)^2 \tag{3.55}$$

where μ_{lj} represent the expected value for the lth component and the jth coordinate. There is a similar representation for σ^2_{lj}. Note that (3.54) and (3.55) are also the expectation and variance under the conditional independence assumption.

For the correlation, we also need to calculate the unconditional covariance. For this, we need

$$\mathbb{E}(x_{j}x_{j'}) = \lambda\left(\sigma^{(1)}_{jj'} + \mu_{1j}\mu_{1j'}\right)(1-\lambda)\left(\sigma^{(2)}_{jj'} + \mu_{2j}\mu_{2j'}\right) \tag{3.56}$$

where $\sigma^{(l)}_{jj'} = \mathrm{Cov}(X_{j}, X_{j'}|l)$ is the conditional covariance between X_{j} and $X_{j'}$ given the component membership $l = 1, 2$ for $1 \leq j,\ j' \leq k$ and $j \neq j'$. Thus, the unconditional covariance is

$$\begin{aligned}\mathrm{Cov}(x_{j}, x_{j'}) =& \mathbb{E}(x_{j}x_{j'}) - \mathbb{E}(x_{j})\mathbb{E}(x_{j'}) \\ =& \lambda\sigma^{(1)}_{jj'} + (1-\lambda)\sigma^{(2)}_{jj'} + \lambda(1-\lambda)\left(\mu_{1j} - \mu_{2j}\right)\left(\mu_{1j'} - \mu_{2j'}\right).\end{aligned} \tag{3.57}$$

Under the conditional independence assumption, $\sigma^{(l)}_{jj'} = 0$ for all $j \neq j'$, $l = 1, 2$. Therefore, the conditional covariance becomes

$$\mathrm{Cov}(x_{j}, x_{j'}) = \lambda(1-\lambda)\left(\mu_{1j} - \mu_{2j}\right)\left(\mu_{1j'} - \mu_{2j'}\right). \tag{3.58}$$

We use the above equations to calculate the correlation using

$$\rho_{jj'} = \frac{\mathrm{Cov}(x_{j}, x_{j'})}{\sqrt{\mathrm{Var}(x_{j})\mathrm{Var}(x_{j'})}}. \tag{3.59}$$

3.6.2 Technique for Conditionally Independent Coordinates

We recommend a technique to help examine the validity the conditional independence assumption. For this proposed method, we compare the sample correlations

between the coordinates with the sample correlations under the conditional independence assumption using (3.58) and (3.59). If the values are "close", then it is reasonable to say that the correlation in the data is not much more than we would see from the conditional independence mixture model.

The first step is to calculate the sample correlations between the coordinates. Denote the sample correlations as $r_{jj'}$ for $1 \leq j, j' \leq k$ and $j \neq j'$. Next, calculate the sample correlations with the formula for the conditional covariances using (3.58) and (3.59). Denote these correlations as $r^*_{jj'}$.

Once we have both sets of correlations, we transform the correlations using Fisher's transformation:

$$Z_{jj'} = \frac{1}{2} \log \left(\frac{1 + r_{jj'}}{1 - r_{jj'}} \right) \tag{3.60}$$

$$Z^*_{jj'} = \frac{1}{2} \log \left(\frac{1 + r^*_{jj'}}{1 - r^*_{jj'}} \right). \tag{3.61}$$

Then plot, $Z_{jj'}$ against $Z^*_{jj'}$, for all $j \neq j'$. Finally, we put boundaries of $\pm 2\frac{1}{\sqrt{n-3}}$ around each of the sample correlations. We choose two standard deviations instead of a multiplier taking into account multiple comparison because we allow for error and choose more conservative intervals. If all of the sample correlations calculated under the conditional independence assumption fall within these bounds, then it suggests that the correlations are what we expect to see in the conditional independence mixture model.

3.6.3 Examples with Simulated Data

We have three examples with a mixture of normal distributions. The first example is generated with a model of conditionally independent coordinates and the second does not assume conditional independence. The two models have the following form:

$$h(X) = \sum_{l=1}^{m} \lambda_l \text{MVN}\left(X; \boldsymbol{\mu}_l, \Sigma_l\right) \tag{3.62}$$

where $X \in \mathbb{R}^k$, $0 \leq \lambda_l \leq 1$ for all $l = 1, \ldots, m$, $\sum_{l=1}^{m} \lambda_l = 1$, and $\text{MVN}(\cdot; \boldsymbol{\mu}, \Sigma)$ is the multivariate normal density function with mean $\boldsymbol{\mu}$ and variance-covariance matrix, Σ.

For the first normal mixture example, we let $m = 2$, $\lambda = 0.3$, $k = 3$, $\Sigma_1 = \Sigma_2 = I_k$ where I_k is the identity matrix, $\boldsymbol{\mu}_1 = (0, 0, 0)'$, and $\boldsymbol{\mu}_2 = (2, 2, 3)'$). We generated $n = 300$ observations. We will compare 3 correlations so they are easy to compare without the use of a plot. The sample correlations are $r_{12} = 0.4693$, $r_{13} = 0.5867$, and $r_{23} = 0.5849$. The respective lower and upper bounds are $(0.3536, 0.5850)$, $(0.4710, 0.7024)$, and $(0.4692, 0.7006)$. The sample correlations under the conditional independence assumption are: 0.4950, 0.5714, and 0.5760. Since the values fall within the lower and upper bounds of the sample correlations, it is reasonable to say the coordinates are consistent with conditional independence.

The second normal mixture followed (3.62) with $m = 3$, $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$, $n = 500$, $\boldsymbol{\mu}_1 = (0, 0, 0, 0, 0)$, $\boldsymbol{\mu}_2 = (2, 2, 2, 2, 2)$, $\boldsymbol{\mu}_3 = (4, 3, 4, 3, 4)$, $\Sigma_1 = \Sigma_2 = I_5$ where I_5 is the identity matrix, and $\Sigma_3 = diag(1, 2, 1, 2, 1)$ where $diag(\cdot)$ is a diagonal matrix. The plot of the sample correlations are shown in Figure 3.3. Since the sample correlations under the conditional independence assumption fall within the bounds, it is reasonable to assume that the coordinates are uncorrelated.

The last example involves a normal mixture with coordinates that are not conditionally independent. We let $m = 2$, $\lambda = 0.5$, $k = 5$, $n = 500$, $\boldsymbol{\mu}_1 = (0, 0, 0, 0, 0)$, $\boldsymbol{\mu}_2 = (2, 2, 2, 2, 2)$, $\Sigma_1 = I_5$ and $\Sigma_2 = diag(0.7) + 0.3$ such that $\sigma_j^2 = 1$ and $\sigma_{jj'} = 0.3$ for all $j \neq j'$. Figure 3.4 show the plot of the correlations. In this plot, the sample correlations calculated under the conditional independence assumption do not fall within the bounds. Therefore, as expected, it is not reasonable to assume the assumption is valid. From the figure, all of the correlations are biased in the same direction. This is because the correlations under the conditionally independent assumption underestimated the correlation in the data.

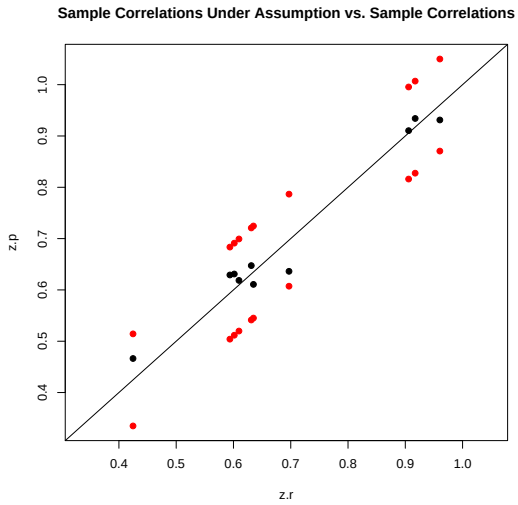


Fig. 3.3. The transformed sample correlations under the conditional independence assumption plotted against the transformed sample correlations for the first example. The red points are the bounds of plus or minus $2/\sqrt{n-3}$.

3.6.4 Handling Blocks of Conditionally i.i.d. Coordinates

The technique described above explains what to do when we have conditionally independent coordinates. In this section, we will discuss how to handle data that contains blocks of conditionally i.i.d. coordinates (see Section 3.4). The sample correlation between two blocks is calculated using

$$r_{aa'} = \frac{\widehat{\mathrm{Cov}}(X_a, X_{a'})}{\sqrt{\widehat{\mathrm{Var}}(X_a)\widehat{\mathrm{Var}}(X_{a'})}} \tag{3.63}$$

where

$$\widehat{\mathrm{Cov}}(X_a, X_{a'}) = \hat{\lambda}(1-\hat{\lambda})\left(\hat{\mu}_{1a} - \hat{\mu}_{2a}\right)\left(\hat{\mu}_{1a'} - \hat{\mu}_{2a'}\right) \tag{3.64}$$

$$\widehat{\mathrm{Var}}(X_a) = \hat{\lambda}\hat{\sigma}^2_{1a} + (1-\hat{\lambda})\hat{\sigma}^2_{2a} + \hat{\lambda}(1-\hat{\lambda})\left(\hat{\mu}_{1a} - \hat{\mu}_{2a}\right)^2 \tag{3.65}$$

$\hat{\mu}_{la}$ is the estimated l-th component mean for block a, and $\hat{\sigma}_a$ is the estimated l-th component standard deviation for block a, for $1 \leq a, a' \leq B$, $a \neq a'$, and $l = 1, \cdots, m$.

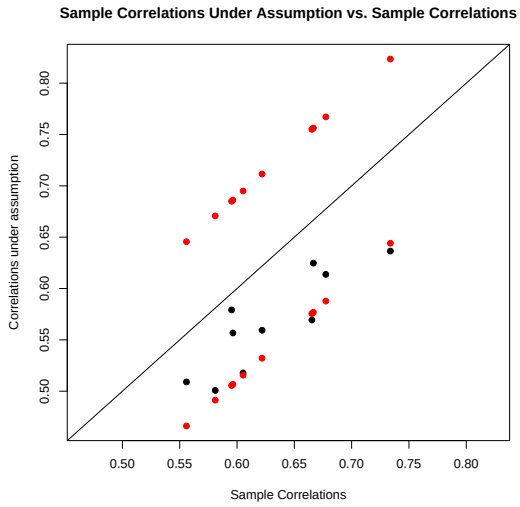


Fig. 3.4. The transformed sample correlations under the conditional independence assumption plotted against the transformed sample correlations for the second example. The red points are the bounds of plus or minus $2/\sqrt{n-3}$.

The next step is to calculate the sample correlations without assuming conditional independence. The difference between this case and the case were we have conditional independence is that we assume some of the coordinates are conditionally i.i.d. Therefore, we will have multiple sample correlations estimating the same population correlation. For example, suppose we have a mixture model with two components and five conditionally conditionally independent coordinates. Now, assume that coordinates 1, 2, and 3 are i.i.d. and make up one block and coordinates 4 and 5 make up the second block. The correlation matrix would look like

$$\begin{bmatrix} 1 & & & & \\ \rho_{12} & 1 & & & \\ \rho_{12} & \rho_{12} & 1 & & \\ \rho^*_{12} & \rho^*_{12} & \rho^*_{12} & 1 & \\ \rho^*_{12} & \rho^*_{12} & \rho^*_{12} & \rho_{45} & 1 \end{bmatrix}, \tag{3.66}$$

where ρ^*_{12} is the correlation between block 1 and block 2 and $\rho_{jj'}$ is the correlation between coordinate j and j', where $j \neq j'$. Note that $\rho_{12} = \rho_{13} = \rho_{23}$ since the coordinates are conditionally i.i.d.. What we propose is to calculate the sample correlation matrix and average over the values that are used to estimate the same parameter. If we let $r_{jj'}$ be the sample correlation for $j \neq j'$, then in our example, $\hat{\rho}_{12} = (r_{12} + r_{13} + r_{23})/3$.

3.6.4.1 Simulated Examples

In this section, we use two simulated datasets, one from a model with a block structure and the other from a model with conditionally i.i.d. coordinates, and we demonstrate how to use the technique to check if the conditional assumption is reasonable.

The first simulated dataset was generated from a model with the following form

$$h(X) = \lambda \prod_{j=1}^{k} \phi\left(x_j; 0, 1\right) + (1-\lambda) \prod_{j=1}^{k} \phi\left(x_j; 2, 1\right) \tag{3.67}$$

with $\lambda = 0.5$, $k = 7$, sample size $n = 500$, and $\phi\left(\cdot; \mu, \sigma^2\right)$ is the normal density function with mean, μ, and variance, σ^2. In the conditional i.i.d. case, we are estimating only one correlation parameter. In this example, the true correlation between the coordinates is 0.50. The Fisher's transformation of the the sample correlation is 0.6508 (or a correlation of 0.5722) and of the sample correlation assuming conditional i.i.d. is 0.6688 (or a correlation of 0.5642) which is within the bounds of plus or minus $\frac{2}{\sqrt{n-3}} = 0.0895$. Therefore, the conditionally i.i.d. assumption does not seem to be violated for this example.

The second simulated example comes from the model

$$\begin{aligned} h(X) =& \lambda_1 \prod_{j=1}^{2} \phi(x_j; 0, 1) \prod_{j=3}^{3} \phi(x_j; 3, 1) \\ &+ (1-\lambda) \prod_{j=1}^{2} \phi(x_j; 2, 1) \prod_{j=3}^{3} \phi(x_j; 5, 1), \end{aligned}$$

where $\lambda_1 = 0.5$ and $n = 500$ observations were generated. Since there are two blocks in this example, there are three correlation parameters we are trying to estimate: the correlations between all the coordinates in each block and the correlation between the two blocks. The set of sample correlations is (0.4694, 0.5151) and the set of correlations under the assumption is (0.5059, 0.5162). The transformed sample correlations without the conditionally independent assumption and with the assumption are (0.5093, 0.5711) and (0.5572, 0.5126), respectively. For this example, the bounds are $\pm 2\frac{1}{\sqrt{n-3}} = 0.0897$. Since all the sample correlations under the assumption are within the bounds, the conditional independence assumption does not seem not violated.

3.7 Likelihood Ratio Tests and Conditional Independence

3.7.1 Introduction

In Section 3.6 we introduced a technique to help us determine the validity of the conditional independence (or conditional i.i.d.) assumption. The most important benefit of using our semiparametric method is the profile likelihood. Since the profile likelihood behaves similarly to a likelihood, we have the advantage of using likelihood ratio tests. In this section, we will use the profile likelihood to perform a test for conditional i.i.d. coordinates versus conditional independent coordinates and also a test for a particular block structure. Once we introduce the tests, we present simulations to show how well the likelihood ratio test using the profile likelihood performs.

Since we have the advantage of a profile likelihood in this method, the test can be carried out using a likelihood ratio test. We have the following test statistic:

$$\Lambda = \frac{L_{P}(\hat{\delta})}{L_{P}(\tilde{\delta})} \tag{3.68}$$

where $L_{P}(\cdot)$ is the profile likelihood, $\hat{\delta}$ is the MLE for $\delta = (\lambda_{1}, \ldots, \lambda_{m}, \theta_{21}, \ldots, \theta_{lk})'$ under the null hypothesis, and $\tilde{\delta}$ is the MLE for δ, under the full model. Likelihood theory suggests that for large samples, $-2\ln\Lambda$ has approximately a chi-square distribution with r degrees of freedom where r is the difference in the number of parameters for the two models.

3.7.2 Likelihood Ratio Tests for Conditionally i.i.d. Blocks Coordinates

The test we consider in this section is for testing conditionally independent coordinates against a model that consists of blocks of conditionally i.i.d. coordinates. In

this situation, we have the following two models:

$$h_1(X) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} f_j(x_j) \exp\left(\tilde{x}_j' \theta_{lj}\right) \tag{3.69}$$

$$h_2(X) = \lambda_1 \prod_{j=1}^{k} f_{b_j}(x_j) + \sum_{l=2}^{m} \lambda_l \left(\prod_{j=1}^{k} f_{b_j}(x_j)\right) \exp\left(\sum_{j=1}^{k} \tilde{x}_{ij}' \theta_{lb_j}\right). \tag{3.70}$$

For model (3.69) there are k coordinates and m components. Therefore we have $(2k+1)(m-1)$ parameters in the model. For model (3.70) there are B blocks and m components. In this model, we have $(2B+1)(m-1)$ parameters. For large samples, the test statistic, (3.68), will have an approximate chi-square distribution with $(2k+1)(m-1) - (2B+1)(m-1) = 2(k-B)(m-1)$ degrees of freedom, where B is the number of blocks.

In Section 3.4 we present the formula for the estimated "jumps", (3.28), for each observation in the block of conditionally iid coordinates. In (3.28), there are nc_a observations where C_a is the number of coordinates that are in block a, for $a = 1, \cdots, B$ and B is the number of blocks. When comparing the log profile likelihood for the case with the blocks and the case without the block, the calculation for the log profile likelihood should be the same. Therefore, in the simulations for the next section, we calculate the log profile likelihood for the block case using equation (3.13) with $\theta_{lj} = \theta_{lb_j}$ for $j = 1, \cdots, k$, where b_j is denotes the block to which the j-th coordinate belongs.

3.7.3 Simulations

In this section, we present simulations to show how well the likelihood ratio test chooses the correct model from either (3.70) or (3.69). For each model we consider, we will show rejection rates at various significance levels as well as a Q-Q plot of the simulated likelihood ratio test statistics.

The first model we consider in a two component ($m = 2$), three coordinate ($k = 3$) normal mixture model

$$h(x_1, x_2, x_3) = \lambda \prod_{j=1}^{2} N\left(x_j; 0, 1\right) N\left(x_k; 0, 1\right) + (1 - \lambda) \prod_{j=1}^{2} N\left(x_j; 2, 1\right) N\left(x_k; 4, 1\right) \tag{3.71}$$

where $\lambda = 0.3, 0.5$, $N(\cdot; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. The results for 300 simulations of sample size $n = 500$ are shown in Table 3.6 for $\lambda = 0.5$. The degrees of freedom for the likelihood statistic is $df = 2(3-2)(2-1) = 2$. Therefore, we compare our likelihood statistics with a chi-square distribution with 2 degrees of freedom. The Q-Q plot for the likelihood statistics is shown in Figure 3.5. The results when $\lambda = 0.3$ are shown in Appendix A. The results are similar to those shown here.

Significance Level (α)	Rejection Rate
0.01	0.0100
0.05	0.0533
0.10	0.0833
0.25	0.2167

Table 3.6. The rejection rates for 300 simulations of sample size $n = 500$ from Model (3.71) at various significance levels.

The second model we considered is a three component ($m = 3$), five coordinate normal mixture model with two block ($B = 2$) of i.i.d. coordinates.

$$\begin{aligned} h(X) =& \lambda_1 \prod_{j=1}^{5} N\left(x_j; 0, 1\right) + \lambda_2 \prod_{j=1}^{3} N\left(x_j; 3, 1\right) \prod_{j=4}^{5} N\left(x_j; 4, 1\right) \\ &+ \lambda_3 \prod_{j=1}^{3} N\left(x_j; 6, 1\right) \prod_{j=4}^{5} N\left(x_j; 7, 1\right). \end{aligned} \tag{3.72}$$

We ran a set of 100 simulations of sample size $n = 500$ from this model. For the first set (results are shown below), $\lambda = (1/3, 1/3, 1/3)$. For large samples, the likelihood ratio

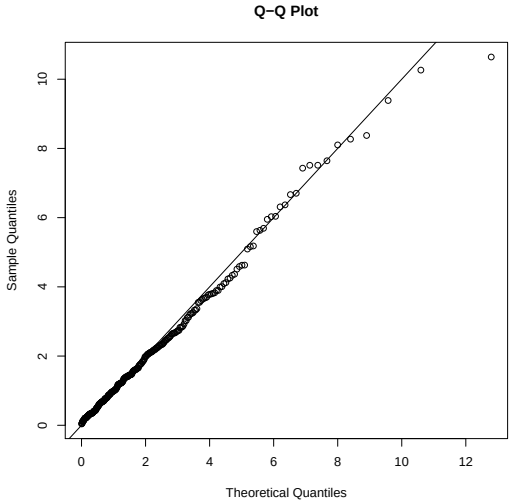


Fig. 3.5. The Q-Q plot for the test statistics from the simulations of Model (3.71). The theoretical distribution is a chi-square with degrees of freedom equal to 2.

statistics should follow a chi-square distribution with $2(5-2)(3-1) = 12$. The rejection rates for four significance levels are shown in Table 3.7 and the Q-Q plot of the likelihood ratio statistics is shown in Figure 3.6.

Significance Level (α)	Rejection Rate
0.01	0.000
0.05	0.050
0.10	0.090
0.25	0.210

Table 3.7. The rejection rates for 100 simulations of sample size $n = 500$ from Model (3.72) at various significance levels.

The next example is from a gamma mixture of the following form:

$$h(X) = \lambda G(2,2,)G(2,2)G(2,2) + (1-\lambda)G(10,0.5)G(10,0.5)G(10,1) \tag{3.73}$$

where $\lambda = 0.5$ and $G(\alpha, \beta)$ is the Gamma density function with shape parameter, α, and scale, β (or with mean $\alpha\beta$ and variance $\alpha\beta^2$). We generated 100 simulations of sample size $n = 300$. The rejection rates for various significance levels are shown in Table 3.8.

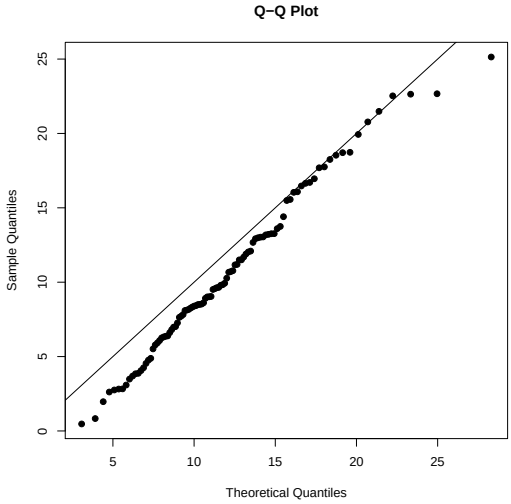


Fig. 3.6. The Q-Q plot for the test statistics from the simulations of Model (3.71). The theoretical distribution is a chi-square with degrees of freedom equal to 12.

The Q-Q plot for the likelihood ratio test statistic against the theoretical values of a chi-square distribution with two degrees of freedom. We can see from the graph that the test statistic follows closely to the chi-square. Also, the rejection rates are close to the significance levels.

Significance Level (α)	Rejection Rate
0.01	0.01
0.05	0.07
0.10	0.12
0.25	0.27

Table 3.8. The rejection rates for 100 simulations of sample size $n = 500$ from Model (3.73) at various significance levels.

3.7.4 Likelihood Ratio Tests for Conditionally i.i.d. Coordinates

In this section, we present the likelihood ratio test for conditionally i.i.d. coordinates. This is a special case of the test in Section 3.7.2 when we have one block ($B = 1$). Consider the following two models. The first model assumes conditional independence

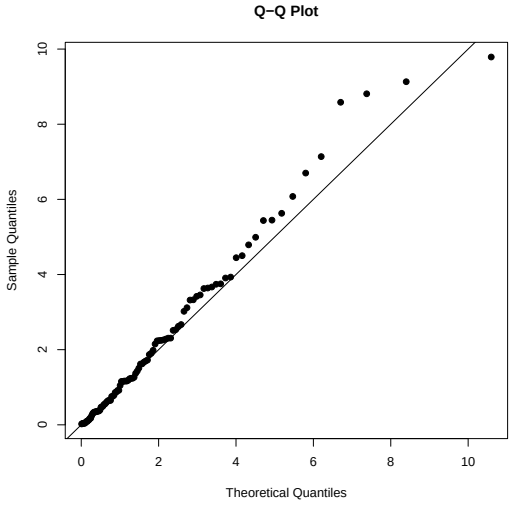


Fig. 3.7. The Q-Q plot for the test statistics from the simulations of Model (3.73). The theoretical distribution is a chi-square with degrees of freedom equal to 2.

of the coordinates and the second assumes the coordinates are conditionally i.i.d.

$$h_1(X) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} f_j(x_j) \exp\left(\tilde{x}_j' \theta_{lj}\right) \tag{3.74}$$

$$h_2(X) = \lambda_1 \prod_{j=1}^{k} f_j(x_j) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} f_j(x_j) \exp\left(\tilde{x}_j' \theta_{l}\right) \tag{3.75}$$

The only difference between (3.74) and (3.75) is that in (3.75) we assume that $\theta_{l1} = \theta_{l2} = \cdots = \theta_{lk}$, for $l = 2, \ldots, m$.

We want to test:

$$H_0 : \theta_{l1} = \ldots = \theta_{lk}, \text{ for } l = 2, \cdots, m$$

$$H_1 : \theta_{lj} \neq \theta_{lj'}, \text{ for } 2 \leq l \leq m \text{ and for some } j \neq j'.$$

In other words, we want to test the null hypothesis that we have model (3.75) against the alternative that we have model (3.74).

For model (3.74), we have k coordinates (or repeated measures) and m components ($m > 1$). The total number of parameters for this model is $(2k + 1)(m - 1)$. For

model (3.75), we have $2(m-1)+(m-1) = 3(m-1)$ parameters. Therefore, the degrees of freedom for this test will be

$$r = (2k+1)(m-1) - 3(m-1) = 2(k-1)(m-1). \tag{3.76}$$

3.7.4.1 Simulations

In this section, we present some simulations to determine how well this method works. The simulations in this section have large sample size, $n = 500$. For simulations with smaller sample size see Section 3.8. The first three examples have the following form

$$h(X) = \sum_{l=1}^{m} \lambda_l \text{MVN}_k \left(X; \boldsymbol{\mu}_l, \Sigma_l \right) \tag{3.77}$$

where $\text{MVN}\left(\cdot, \boldsymbol{\mu}_l, \Sigma_l\right)$ is a multivariate normal distribution with k coordinates and mean $\boldsymbol{\mu}_l$ and variance-covariance matrix Σ_l, for all $l = 1, \ldots, m$.

The first model (Model (1)) is a mixture model of the form (3.77) with two components ($m = 2$), three coordinates ($k = 3$), $\lambda_1 = \lambda_2 = 0.50$, $\boldsymbol{\mu}_1 = (0,0,0)'$, $\boldsymbol{\mu}_2 = (2,2,2)'$, and $\Sigma_1 = \Sigma_2 = I_3$, where I_3 is the identity matrix with three rows. We generated 200 simulations of sample size $n = 500$. We calculated the profile likelihoods under both models then calculated the chi-square statistic. The test statistic should follow a chi-square distribution with four degrees of freedom. Figure 3.8 shows the quantiles for the test statistic against those of a chi-square random variable with $df = 4$. The values should fall on the line if the distributions are the same. In Table 3.9 we table the rejection rates for various significance values. From both the table and the figure, we can see that the distribution of the test statistic falls very close to that of a chi-square statistic with the appropriate degrees of freedom, in this case $df = 4$.

The second set of simulations (Model (2)) has the form of (3.77) with $m = 2$ components, $k = 3$ coordinates, $\lambda_1 = \lambda_2 = 0.50$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = (0,0,0)'$, $\Sigma_1 = I_3$, and $\Sigma_2 = 100 I_3$, where I_3 is the identity matrix with three rows. We generated 200 simulations of sample size $n = 500$ and calculated the test statistics. Under the null hypothesis, the test statistics should follow a chi-square distribution with four degrees

Significance Level (α)	Rejection Rate
0.01	0.005
0.05	0.045
0.10	0.070
0.25	0.225

Table 3.9. The rejection rates for Model (1) at various significance levels.

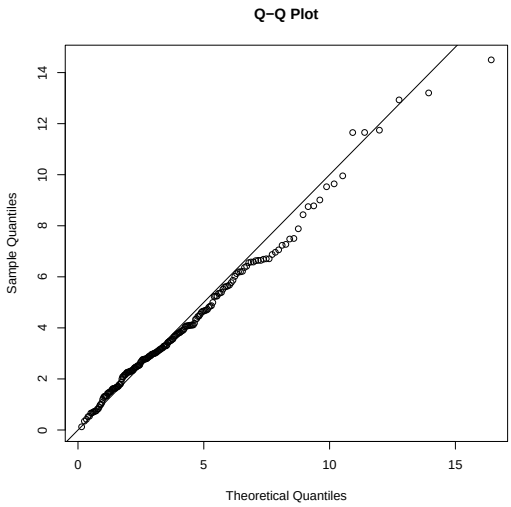


Fig. 3.8. The Q-Q plot for the test statistics from the simulations of Model (1). The theoretical distribution is a chi-square with degrees of freedom equal to 4.

of freedom. In Table 3.10, we display the rejection rates for the test statistics at four different significance values. The rejection rates are close to the significance values. In Figure 3.9 is a Q-Q plot for the test statistics. The values fall close to the theoretical values. We can also see that the this method works well even for a scale mixture.

The third model we consider has the following attribution in (3.77): $m = 3$, $k = 7$, $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$, $\boldsymbol{\mu}_1 = (0,0,0,0,0,0,0)'$, $\boldsymbol{\mu}_2 = (2,2,2,2,2,2,2)'$, $\boldsymbol{\mu}_3 = (4,4,4,4,4,4,4)'$, and $\Sigma_1 = \Sigma_2 = \Sigma_3 = I_7$. We generated 500 simulations of sample size $n = 500$. For this example, we have $2(k-1)(m-1) = 2(6)(2) = 24$ parameters and thus $df = 24$ for the test statistic. The quantile plot is shown in Figure 3.10 and the rejection rates are tabled in Table 3.11.

Significance Level (α)	Rejection Rate
0.01	0.010
0.05	0.075
0.10	0.145
0.25	0.260

Table 3.10. The rejection rates for Model (2) at various significance levels.

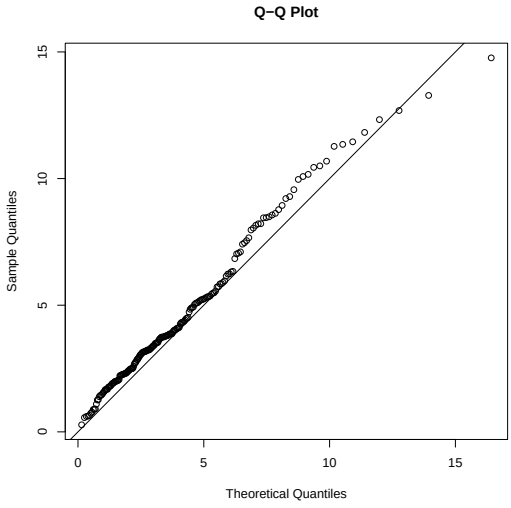


Fig. 3.9. The Q-Q plot for the test statistics from the simulations of Model (2). The theoretical distribution is a chi-square with degrees of freedom equal to 4.

Next, we will consider an example where the true densities are not symmetric. Consider the following model:

$$h(x_1, x_2, x_3) = \lambda \prod_{j=1}^{3} G(2,2) + (1-\lambda) \prod_{j=1}^{3} G(10,1) \tag{3.78}$$

where $\lambda = 0.5$ and $G(\alpha, \beta)$ is the Gamma density function with parameters, α and β (or mean equal to $\alpha\beta$ and a variance of $\alpha\beta^2$). We generated 100 simulations of sample size $n = 500$. The degrees of freedom for the likelihood ratio test is $df = 4$. Table 3.12 shows the rejection rates for various significance levels and Figure 3.11 displays the Q-Q plot for the likelihood test statistics for 100 simulated datasets. The rejection rates are close

Significance Level (α)	Rejection Rate
0.01	0.010
0.05	0.070
0.10	0.115
0.25	0.225

Table 3.11. The rejection rates for Model (3) at various significance levels.

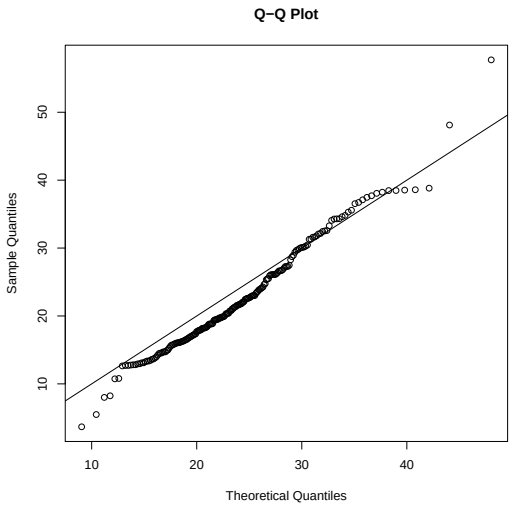


Fig. 3.10. The Q-Q plot for the test statistics from the simulations of Model (3). The theoretical distribution is a chi-square with degrees of freedom equal to 24.

to the significance levels and the test statistics seem to follow the chi-square distribution with 4 degrees of freedom.

From the four examples shown in this section, the likelihood ratio test using the profile likelihoods seem to follow the theorized distribution of the likelihood ratio statistic using the actual likelihood. We have shown that the likelihood ratio test could be used as a tool to help determine an appropriate model.

Significance Level (α)	Rejection Rate
0.01	0.00
0.05	0.02
0.10	0.06
0.25	0.22

Table 3.12. The rejection rates for Model (3.78) at various significance levels.

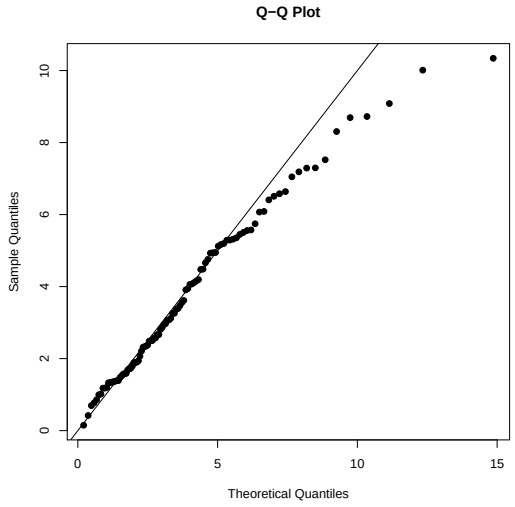


Fig. 3.11. The Q-Q plot for the test statistics from the simulations of Model (3.78). The theoretical distribution is a chi-square with degrees of freedom equal to 4.

3.8 Monte Carlo simulations of the component means and standard deviations

3.8.1 Introduction

In this section, we provide the results of various simulation studies. The observations were drawn from the following mixture model

$$h(x_{1},\ldots,x_{k}) = \lambda_{1}\prod_{j=1}^{k} f_{j}(x_{j}) + \sum_{l=2}^{m}\lambda_{l}\prod_{j=1}^{k} g_{lj}(x_{j}) \tag{3.79}$$

where λ_{l} is the mixing proportion and f_{j} and g_{lj} are density functions for $j = 1,\ldots,k$ and $l = 1,\ldots,m$. The nine different simulation studies will be of this form with various density functions, components, and coordinates. The simulation results from the tilted method are compared to the conditionally independent normal mixture model and the nonparameteric method proposed by Benaglia et al. (2009a). In cases where applicable, we compare our results to others in the literature, such as the cut-point model (see Hettmansperger and Thomas (2000); Elmore (2003); Cruz-Medina et al. (2004)).

There are a variety of models used for the simulations. We selected several normal mixture models, one pure location mixture, one pure scale mixture, and one location/scale mixture. In section 1.4, we showed how we can use the exponential tilt model (Anderson, 1979) to find relationships between normal density functions. We will also consider other models that do not satisfy the exponential tilt assumption. Since the interpretation of the quadratic exponent is not straightforward, we report results for the component means and standard deviations (see Section 3.5.3). We will also provide examples of the estimated CDFs and PDFs for each component.

All of the results from the semiparametric method in this section were calculated using the function in Appendix B. The results from the nonparametric method proposed by Benaglia et al. (2009a) and the EM for the conditionally independent normal mixture were calculated using the package `mixtools` in `R`.

3.8.2 Normal Location Mixture with Conditionally i.i.d. Coordinates

For this simulation study, we generated observations from the following normal location mixture model

$$h(x_1, \ldots, x_k) = \lambda \prod_{j=1}^{k} \phi\left(x_j; 0, 1\right) + (1-\lambda) \prod_{j=1}^{k} \phi\left(x_j; \mu, 1\right) \tag{3.80}$$

with $\phi(\cdot; \mu, \sigma^2)$ the normal density function, $\lambda = 0.5$, and $\mu = 2$. We chose this particular model to compare the results reported in Benaglia (2008) in which they reported results using the cut-point model purposed in Elmore (2003) and those of the normal mixture model.

We generated 1000 simulations of sample size $n = 100$ from (3.80). The results are shown in Table 3.13. The table includes the results from Benaglia (2008), which include the estimated component means and variances from the nonparametric method (npEM), from the cut-point model, the normal mixture with i.i.d. components, and from the exponential tilt model (tiltedEM). We calculated the estimated MSE for comparative purposes. The results from the normal model should perform better than the other methods since it is the true model. As you can see from the table, the estimates are close to the true values in all four methods and all of the MSEs are of the same magnitude. For an arbitrary simulation, we plotted the estimated component densities and the estimated CDFs based on the tilted model in Figure 3.12. The estimated densities are very close to the true densities. We expected our model to produce good results in this situation because normal densities fit the exponential tilt assumption.

For the next normal location mixture simulations, we set $\lambda = 0.5$, $\sigma_1^2 = \sigma_2^2 = 1$, and $\mu_1 = 1$ and $n = 100$. Again, this mixture model is identical to the one used in Benaglia (2008) in order to compare the tiltedEM, the normal mixture, the cut-point method, and the npEM. The mixture for these simulations is not well separated (only one standard deviation apart). Since it is the true model, we expect the normal mixture to perform better than the semiparametric methods and the nonparametric method. The estimated component means and standard deviations, along with their standard errors

		Parameters				
Method		λ	μ_1	σ^2_1	μ_2	σ^2_2
npEM	Mean	0.49666	0.00174	0.99378	2.00139	0.99630
	Std.Error	0.05188	0.05263	0.07040	0.05298	0.07497
	MSE	0.00278	0.00277	0.00499	0.00281	0.00563
Cut-Point	Mean	0.49988	-0.00016	0.99873	2.00020	0.99770
	Std.Error	0.04970	0.05158	0.07369	0.05148	0.07649
	MSE	0.00247	0.00266	0.00543	0.00265	0.00586
Normal	Mean	0.49665	0.00160	0.99353	2.00147	0.99613
	Std.Error	0.05193	0.05259	0.07024	0.05304	0.07475
	MSE	0.00271	0.00277	0.00498	0.00282	0.00560
TiltedEM	Mean	0.50017	-0.00315	0.99840	1.99859	0.99983
	Std.Error	0.05282	0.05139	0.07257	0.05201	0.07350
	MSE	0.00279	0.00265	0.00527	0.00271	0.00540

Table 3.13. Two component normal mixture with i.i.d. coordinates with $\lambda = 0.5$, $\mu_1 = 0$, $\mu_2 = 2$, and $\sigma^2_1 = \sigma^2_2 = 1$. The results are based on 1000 simulations each with sample size $n = 100$.

and MSEs, are shown in Table 3.14. We randomly selected one of the simulated datasets and plotted the semiparametric estimates for the component PDFs and CDFs. They are shown in Figure 3.13. The solid lines are the estimated density functions and the dashed lines are the true ones. We can see from the figure that the exponential tilt method produces estimates that are close to the true functions.

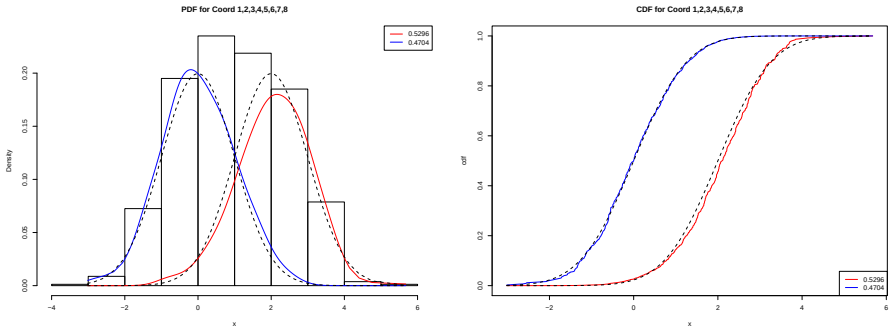


Fig. 3.12. Estimated PDF and CDF for a two component normal mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.5$, $\mu_1 = 0$, $\mu_2 = 2$, and $\sigma_1^2 = \sigma_2^2 = 1$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

		Parameters				
Method		λ	μ_1	σ_1^2	μ_2	σ_2^2
npEM	Mean	0.25048	0.00013	1.01758	1.00110	0.99382
	Std.Error	0.06179	0.11090	0.12219	0.05394	0.06062
	MSE	0.00382	0.01230	0.01524	0.00291	0.00371
Cut-Point	Mean	0.25950	0.00738	0.99959	1.00252	0.99454
	Std.Error	0.07176	0.12369	0.12394	0.05710	0.06626
	MSE	0.00524	0.01535	0.01536	0.00327	0.00442
Normal	Mean	0.25207	-0.00409	0.99442	1.00377	0.99432
	Std.Error	0.06242	0.11898	0.12568	0.05666	0.06329
	MSE	0.00390	0.01417	0.01583	0.00322	0.00404
TiltedEM	Mean	0.25424	-0.00437	0.99095	1.00207	0.99740
	Std.Error	0.06412	0.11323	0.13419	0.05305	0.06778
	MSE	0.00413	0.01284	0.01809	0.00282	0.00460

Table 3.14. Two component normal mixture with i.i.d. coordinates with $\lambda = 0.5$, $\mu_1 = 0$, $\mu_2 = 1$, and $\sigma_1^2 = \sigma_2^2 = 1$. The results are based on 1000 simulations each with sample size $n = 100$.

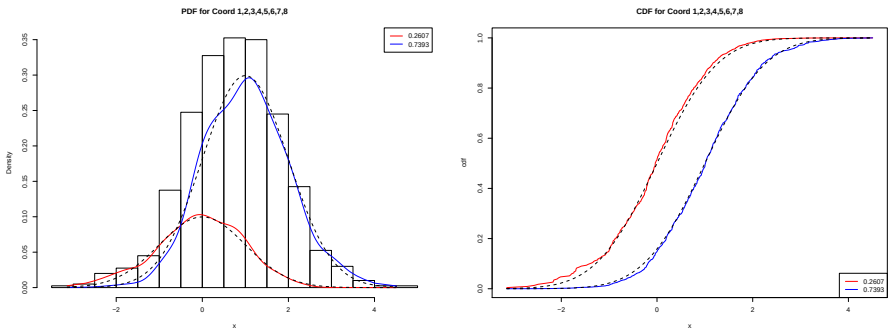


Fig. 3.13. Estimated PDF and CDF for a two component normal mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.5$, $\mu_1 = 0$, $\mu_2 = 1$, and $\sigma_1^2 = \sigma_2^2 = 1$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

3.8.3 Normal Scale Mixture with Conditionally i.i.d. Coordinates

The simulations in this subsection are from the following normal scale mixture with i.i.d. coordinates

$$h(x_1, \ldots, x_k) = \lambda \prod_{j=1}^{k} \phi\left(x_j; 0, 1\right) + (1-\lambda) \prod_{j=1}^{k} \phi\left(x_j; 0, \sigma^2\right) \tag{3.81}$$

where $k = 3$ is the number of coordinates, and $\phi(\cdot, 0, \sigma^2)$ is the normal density function with mean 0 and variance σ^2. For the simulations, we generated 1000 simulations for sample size $n = 100$ from (3.81) with $m = 2$, $k = 8$, $\lambda = 0.5$, and $\sigma^2 = 9$. Again, we ran these simulations to compare to the results found in Benaglia (2008) and Elmore (2003).

The results from the exponential tilt model along with those found in Benaglia (2008) are shown in Table 3.15. The table contains the means, standard errors, and the estimated MSEs for all four models. The normal mixture should perform better than the other methods. We can see from the table that in general, the normal has the smallest MSEs with the tiltedEM having values very close to those of the normal. This is not surprising due to the exponential tilt assumption. The tiltedEM does seem to perform slightly better than the nonparametric and the cut-point method. Although scale mixtures can be difficult to estimate, the tiltedEM performs very well. A plot of the estimated component PDFs and CDFs are shown in Figure 3.14. The component PDFs and CDFs are the solid lines and the true densities are the dashed lines. The estimated densities are very close to the true densities.

		Parameters				
Method		λ	μ_1	σ^2_1	μ_2	σ^2_2
npEM	Mean	0.53787	-0.00003	1.25462	-0.00077	9.30732
	Std.Error	0.05153	0.05691	0.23016	0.16183	0.68219
	MSE	0.00409	0.00324	0.11780	0.02619	0.55983
Cut-Point	Mean	0.50853	0.00124	1.08239	0.00370	9.04498
	Std.Error	0.05541	0.05385	0.12530	0.16078	0.71188
	MSE	0.00314	0.00290	0.02249	0.02586	0.50880
Normal	Mean	0.50003	-0.00095	0.99463	-0.00012	8.96253
	Std.Error	0.04959	0.05076	0.07914	0.14874	0.66406
	MSE	0.00246	0.00258	0.00629	0.02212	0.44238
TiltedEM	Mean	0.50010	0.00006	0.99998	0.00533	8.97602
	Std.Error	0.05231	0.05177	0.08047	0.15332	0.65568
	MSE	0.00274	0.00268	0.00648	0.02354	0.43049

Table 3.15. Two component normal scale mixture with i.i.d. coordinates with $\lambda = 0.5$, $\mu_1 = \mu_2 = 0$, $\sigma^2_1 = 1$, and $\sigma^2_2 = 9$. The results are based on 1000 simulations each with sample size $n = 100$.

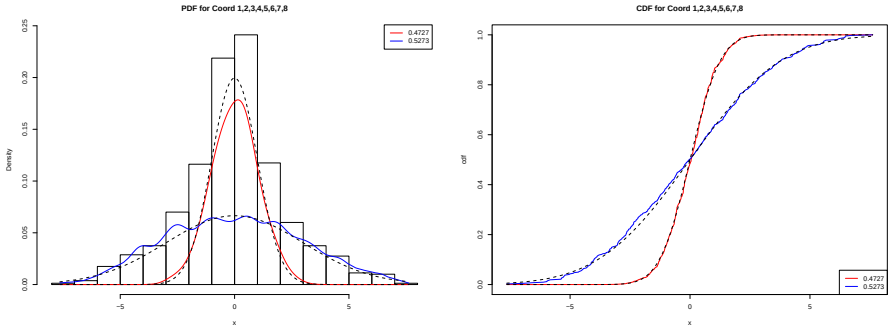


Fig. 3.14. Estimated PDF and CDF for a two component normal scale mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.5$, $\mu_1 = \mu_2 = 0$, $\sigma^2_1 = 1$ and $\sigma^2_2 = 9$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

3.8.4 Laplace Location Mixture with Conditionally i.i.d. Coordinates

This model is a two component mixture of Laplace (double exponential) distributions with conditionally i.i.d. coordinates. The model is of the form

$$h(x_1, \cdots, x_k) = \lambda \prod_{j=1}^{k} \mathcal{L}(x_j; 0, 1) + (1 - \lambda) \prod_{j=1}^{k} \mathcal{L}(x_j; \mu, 1), \tag{3.82}$$

where $\mathcal{L}(\cdot; \mu, \beta)$ is a Laplace distribution with mean μ and variance 2β. For this location model, we chose the parameters to match those of Benaglia (2008) for comparison. Therefore, we have $\lambda = 0.25$, $\mu = 1$, and $\beta = 1$, which implies that $\sigma_1^2 = \sigma_2^2 = 2$. We generated 1000 simulations of sample size $n = 100$.

We compare the results for these simulations using the tiltedEM, the nonparametric method, the cut-point method, a normal mixture with conditionally i.i.d. coordinates. We do not expect the normal mixture EM to perform well since it is not the correct model. The nonparametric method, npEM, should do better since it is a fully nonparametric approach. It is an interesting model to consider since the exponential tilt assumption is not valid for this model. We expect it to estimate the component means and standard deviations well, however, due to the moment matching property of the method.

Table 3.16 shows the results as shown in Benaglia (2008) as well as those from the tiltedEM. We included the estimated MSEs instead of the bias to obtain a better idea of how well methods estimate the component means, standard deviations, and mixing parameters. Of all four methods, the normal procedure performs the worst. The estimates from the nonparametric method and the exponential tilt method provide the best results. It is worth noting, even though the exponential tilt assumption is not valid, the tiltedEM provides good estimates.

		Parameters				
Method		λ	μ_1	σ^2_1	μ_2	σ^2_2
npEM	Mean	0.27193	0.03381	2.07990	1.00896	1.95796
	Std.Error	0.07995	0.22308	0.47448	0.10196	0.19025
	MSE	0.00687	0.05091	0.23152	0.01048	0.03796
Cut-Point	Mean	0.25958	-0.00348	1.96639	1.00967	1.97900
	Std.Error	0.09970	0.20167	0.41106	0.09088	0.20959
	MSE	0.01003	0.04068	0.17010	0.00835	0.04437
Normal	Mean	0.31326	0.44078	2.29526	0.84752	2.00739
	Std.Error	0.10931	0.48929	1.33621	0.20984	0.60117
	MSE	0.01595	0.43369	1.87258	0.06728	0.36146
TiltedEM	Mean	0.23076	-0.10382	1.69741	0.98991	2.04994
	Std.Error	0.04679	0.15113	0.43492	0.07720	0.22287
	MSE	0.00256	0.03362	0.28072	0.00606	0.05217

Table 3.16. Two component Laplace location mixture with i.i.d. coordinates with $\lambda = 0.25$, $\mu_1 = 0$, $\mu_2 = 1$, and $\sigma^2_1 = \sigma^2_2 = 2$. The results are based on 1000 simulations each with sample size $n = 100$.

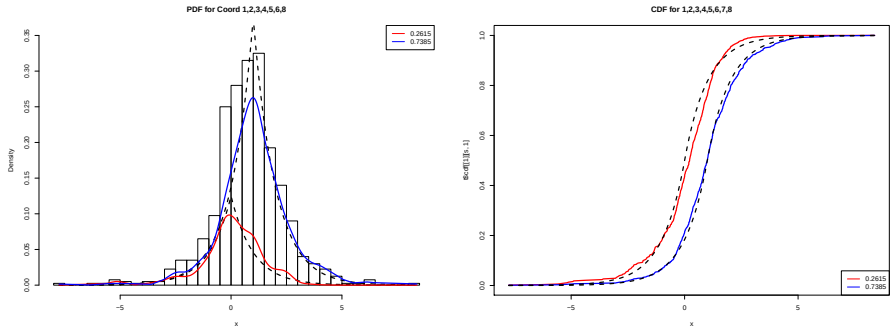


Fig. 3.15. Estimated PDF and CDF for a two component Laplace location mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.25$, $\mu_1 = 0$, $\mu_2 = 1$, $\sigma^2_1 = \sigma^2_2 = 2$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

3.8.5 Normal Mixtures with Conditionally Independent Coordinates

We considered two trivariate conditionally independent normal mixture models of the form

$$\lambda \text{MN}(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2 \mathbf{I}_k) + (1-\lambda)\text{MN}(\mathbf{x}; \boldsymbol{\mu}_2, \boldsymbol{\sigma}_2^2 \mathbf{I}_k). \tag{3.83}$$

where MN is the multivariate normal density, $\boldsymbol{\mu}_l$ is the mean and the variance-covariance matrix $\boldsymbol{\sigma}_l^2 \mathbf{I}_k$ for $l = 1, 2$, and k is the number of coordinates, and λ is the mixing proportion.

For the first set of simulations, we use model (3.83) with $k = 3$, $\boldsymbol{\mu}_1 = (0, 0, 0)'$, $\boldsymbol{\mu}_2 = (1, 1.5, 2.5)$, $\boldsymbol{\sigma}_1^2 = (1, 1, 1)'$, and $\boldsymbol{\sigma}_2^2 = (1, 1, 1)$. Three values of $\lambda = 0.3,\ 0.5,\ 0.8$ and a sample size of $n = 500$. For each of the combinations of λ, 500 simulations were carried out. The results are shown in Table 3.17 where we show the estimated component means, standard deviations, and mixing proportion for the tiltedEM, the normal mixture, and the nonparametric method, npEM. All the methods perform similarly and produce good estimates for the component means and variances and for the mixing proportion. For this model, we can no longer compare our results to the cut-point model since the cut-point model requires the assumption of the coordinates being conditionally i.i.d. Using an arbitrary simulated dataset, the estimated component PDFs and CDFs are produced in Figure 3.16. The solid lines are the estimates and the dashed lines are the true density functions. We can see that the estimates are close to the true densities. The results for $\lambda = 0.5, 0.8$ produced similar results for both sets of simulations. The tabled values can be found in Tables B.2 and B.3 in Appendix B.

For the second set of simulations, we use model (3.83) with $k = 3$, $\boldsymbol{\mu}_1 = (0, 0, 0)'$, $\boldsymbol{\mu}_2 = (2, 2.5, 3)$, $\boldsymbol{\sigma}_1^2 = (1, 1, 1)'$, and $\boldsymbol{\sigma}_2^2 = (1.5, 2, 1)$. Three values of $\lambda = 0.3,\ 0.5,\ 0.8$ and a sample size of $n = 500$. For each of the combinations of λ, 500 simulations were carried out. The results are shown in Table 3.18 where we show the estimated component means, standard deviations, and mixing proportion for the tiltedEM, the normal mixture, and the nonparametric method, npEM. Again, the normal method, the

tiltedEM, and the npEM perform very well but this time the normal method performs slightly better. This may be due to the component densities being fairly close together. Using an arbitrary simulated dataset, the estimated component PDFs and CDFs are produced in Figure 3.17. The solid lines are the estimates and the dashed lines are the true density functions. Again, we can see that the estimates are close to the true densities. The results for $\lambda = 0.5, 0.8$ produced similar results for both sets of simulations. The tabled values can be found in Tables B.4 and B.5 Appendix B.

Table 3.17. Two component normal mixture with conditionally independent coordinates with $\lambda = 0.3$, $\boldsymbol{\mu}_1 = (0, 0, 0)$, $\boldsymbol{\mu}_2 = (1, 1.5, 2.5)$, and $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2 = (1, 1, 1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$.

Parameter	True	TiltedEM	Normal	npEM
λ	0.3	0.3024(0.0330)	0.3004(0.0297)	0.2986(0.0302)
μ_{11}	0	-0.0033(0.0967)	-0.0014(0.0950)	-0.0096(0.0946)
μ_{12}	0	0.0001(0.1006)	0.0012(0.0968)	-0.0081(0.0953)
μ_{13}	0	0.0308(0.1466)	0.0042(0.1223)	0.0605(0.1285)
μ_{21}	1	0.9988(0.0571)	0.9952(0.0556)	0.9962(0.0554)
μ_{22}	1.5	1.5012(0.0682)	1.4968(0.0656)	1.4970(0.0646)
μ_{23}	2.5	2.4904(0.0680)	2.4959(0.0656)	2.4654(0.0645)
σ_{11}	1	0.9909(0.0658)	0.9915(0.0658)	0.9918(0.0640)
σ_{12}	1	0.9931(0.0724)	0.9944(0.0679)	0.9932(0.0660)
σ_{13}	1	1.0235(0.1322)	0.9929(0.0876)	1.0706(0.1098)
σ_{21}	1	0.9966(0.0422)	0.9986(0.0406)	0.9967(0.0402)
σ_{22}	1	0.9959(0.0455)	0.9993(0.0426)	0.9973(0.0419)
σ_{23}	1	1.0006(0.0603)	0.9979(0.0497)	1.0280(0.0549)

Parameter	True	TiltedEM	Normal	npEM
λ	0.3	0.2997(0.0221)	0.2986(0.0220)	0.3026(0.0225)
μ_{11}	0	-0.0053(0.0938)	-0.0066(0.0929)	0.0033(0.0928)
μ_{12}	0	0.0020(0.0839)	0.0006(0.0833)	0.0124(0.0845)
μ_{13}	0	0.0090(0.0910)	-0.0034(0.0874)	0.0400(0.0927)
μ_{21}	2	1.9981(0.0675)	1.9955(0.0667)	2.0027(0.0680)
μ_{22}	2.5	2.5079(0.0786)	2.5044(0.0787)	2.5137(0.0785)
μ_{23}	3	2.9981(0.0566)	2.9988(0.0557)	2.9970(0.0560)
σ_{11}	1	0.9934(0.0653)	0.9916(0.0647)	1.0026(0.0655)
σ_{12}	1	0.9940(0.0671)	0.9929(0.0650)	1.0118(0.0749)
σ_{13}	1	1.0176(0.0840)	0.9980(0.0665)	1.0594(0.0847)
σ_{21}	1.2247	1.2239(0.0514)	1.2260(0.0511)	1.2209(0.0508)
σ_{22}	1.4142	1.4060(0.0587)	1.4084(0.0576)	1.4006(0.0580)
σ_{23}	1	0.9994(0.0428)	0.9999(0.0407)	1.0021(0.0410)

Table 3.18. Two component normal mixture with conditionally independent coordinates with $\lambda = 0.3$, $\boldsymbol{\mu}_1 = (0,0,0)$, $\boldsymbol{\mu}_2 = (2,2.5,3)$, $\boldsymbol{\sigma}_1^2 = (1,1,1)$, and $\boldsymbol{\sigma}_2^2 = (1.5,2,1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$.

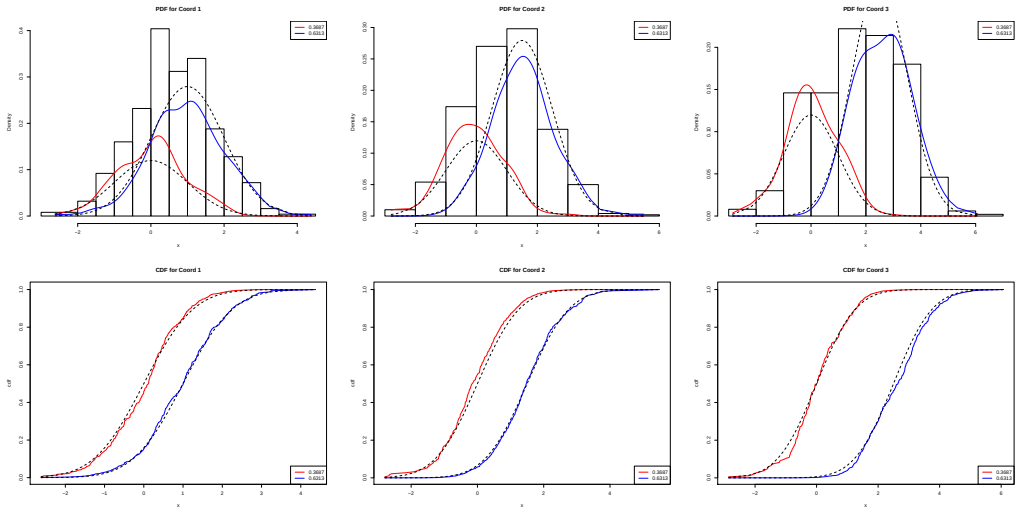


Fig. 3.16. Estimated PDF and CDF for a two component normal mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.3$, $\boldsymbol{\mu}_1 = (0,0,0)'$, $\boldsymbol{\mu}_2 = (1,1.5,2.5)'$, and $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2 = (1,1,1)'$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

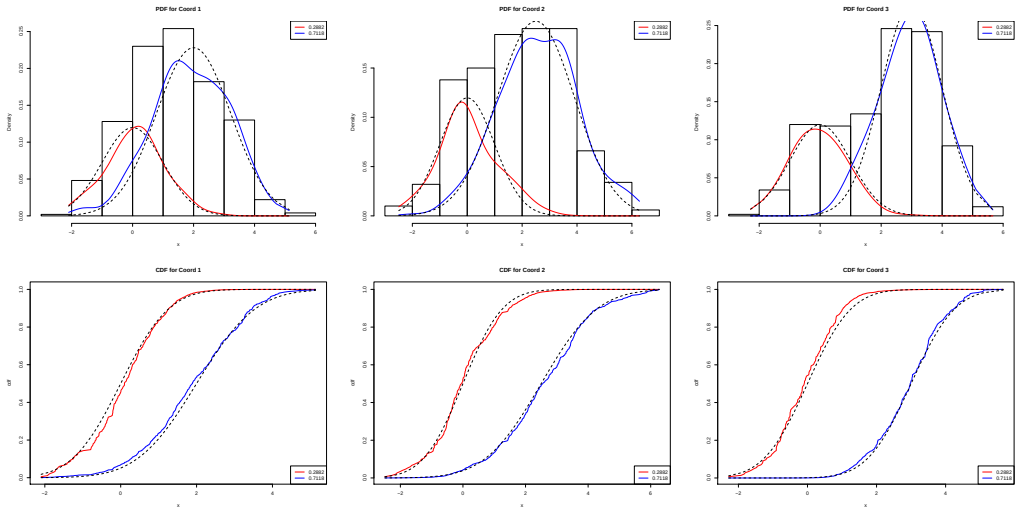


Fig. 3.17. Estimated PDF and CDF for a two component normal mixture model with i.i.d. coordinates. The parameters are $\lambda = 0.3$, $\boldsymbol{\mu}_1 = (0,0,0)'$, $\boldsymbol{\mu}_2 = (1,2.5,3)'$, and $\boldsymbol{\sigma}_1^2 = (1,1,1)'$, and $\boldsymbol{\sigma}_2^2 = (1.5,2,1)'$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

3.8.6 Gamma Mixture with Conditionally Independent Coordinates

Our model not only provides good estimates when marginals are normal but also for other distributions where the exponential tilt assumption may not be strictly valid. In Section 1.4, we show how two gamma distributions with the same shape parameter are related by an exponential tilt. We were interested in checking the performance of the semiparametric method when the assumption of the exponential tilt relationship is not valid and when the underlying distributions are not symmetric. Therefore, in this section, we ran simulations from the following model

$$\lambda g(x_1; 2, 2)g(x_2; 2, 2)g(x_3; 2, 2) + (1 - \lambda)g(x_1; 5, 2)g(x_2; 10, 1)g(x_3; 10, 0.5) \quad (3.84)$$

where $\lambda = 0.4$ and $g(\cdot; \alpha, \beta)$ is a gamma density function with parameters α and β such that the mean is $\alpha\beta$ and the variance is $\alpha\beta^2$. We used $\lambda = 0.4$ and 1000 simulations of sample sizes $n = 50$ and 300 were carried out. Note that these densities do not have a common shape parameter (α) and therefore do not satisfy the exponential tilt assumption.

Table 3.19 shows the semiparametric estimates based on 1000 simulations from (3.84) for both sample sizes. When the sample size is small ($n = 50$), the standard errors of the estimates are large. But as the sample size increases, the standard errors also decrease. We also calculated the estimates from a normal mixture and the nonparametric method; see Table 3.20. Although the results are not tabled, we calculated the MSEs for all of the estimates for both sample sizes. The plots are shown in Figure 3.19. We expect that the nonparametric method will produce better estimates since the model is misspecified using the normal mixture and/or the tiltedEM. For sample size $n = 50$, the performances of the three different methods are very similar with npEM containing slightly smaller MSEs than the tiltedEM and the normal. When the sample size increases, however, the tiltedEM has MSEs closer to those of the nonparametric method. Even though the assumptions are not valid for the gamma distributions, the exponential tilt method still produces decent estimates.

For one of the simulated datasets, we plotted the estimated component PDFs and CDFs; see Figure 3.18. The solid lines are the estimated density functions and the dashed lines are the true ones. The components are not well separated and asymmetric and yet the method does a good job estimating the true densities.

Parameter	True	$n = 50$	$n = 300$
λ	0.4	0.3867(0.1193)	0.3731(0.0402)
μ_{11}	4	3.8665(1.0225)	3.8497(0.3563)
μ_{12}	4	4.7867(2.5215)	3.7885(0.5376)
μ_{13}	4	4.0312(0.8921)	3.9709(0.3074)
μ_{21}	10	9.9907(1.1238)	9.8466(0.3866)
μ_{22}	10	9.0397(1.7791)	9.8656(0.4154)
μ_{23}	5	4.8488(0.4480)	4.9617(0.1335)
σ_{11}	2.8284	2.4731(1.0585)	2.6509(0.4470)
σ_{12}	2.8284	2.4781(1.0094)	2.5349(0.5008)
σ_{13}	2.8284	2.4283(0.8997)	2.8429(0.3273)
σ_{21}	4.4721	4.2317(0.8020)	4.4886(0.2984)
σ_{22}	3.1623	3.2063(0.6265)	3.2180(0.2276)
σ_{23}	1.5811	1.6973(0.4666)	1.5999(0.1474)

Table 3.19. Two component gamma mixture with conditionally independent coordinates with $\lambda = 0.4$, $\boldsymbol{\alpha}_1 = (2, 2, 2)$, $\boldsymbol{\alpha}_2 = (5, 10, 10)$, $\boldsymbol{\beta}_1 = (2, 2, 2)$, and $\beta_2 = (2, 1, 0.5)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.84) with sample sizes $n = 50$ and $n = 300$.

Parameter	True	Normal	npEM
λ	0.4	0.3182(0.0458)	0.3603(0.0369)
μ_{11}	4	3.4184(0.3679)	3.7848(0.3126)
μ_{12}	4	3.3216(0.4566)	3.7192(0.3400)
μ_{13}	4	4.0684(0.3565)	3.9597(0.2977)
μ_{21}	10	9.5657(0.3920)	9.7648(0.3677)
μ_{22}	10	9.5983(0.3219)	9.7934(0.2807)
μ_{23}	5	4.8400(0.1322)	4.9517(0.1244)
σ_{11}	2.8284	1.9662(0.2811)	2.6228(0.4085)
σ_{12}	2.8284	1.9050(0.3321)	2.5480(0.3825)
σ_{13}	2.8284	2.9100(0.3906)	2.8085(0.3271)
σ_{21}	4.4721	4.5843(0.2888)	4.5026(0.2886)
σ_{22}	3.1623	3.4283(0.2330)	3.2722(0.2017)
σ_{23}	1.5811	1.7174(0.1365)	1.6750(0.1225)

Table 3.20. Results from the normal and nonparametric method based on 1000 simulations of sample size $n = 300$ from model 3.84. Compare these values to those of Table 3.19

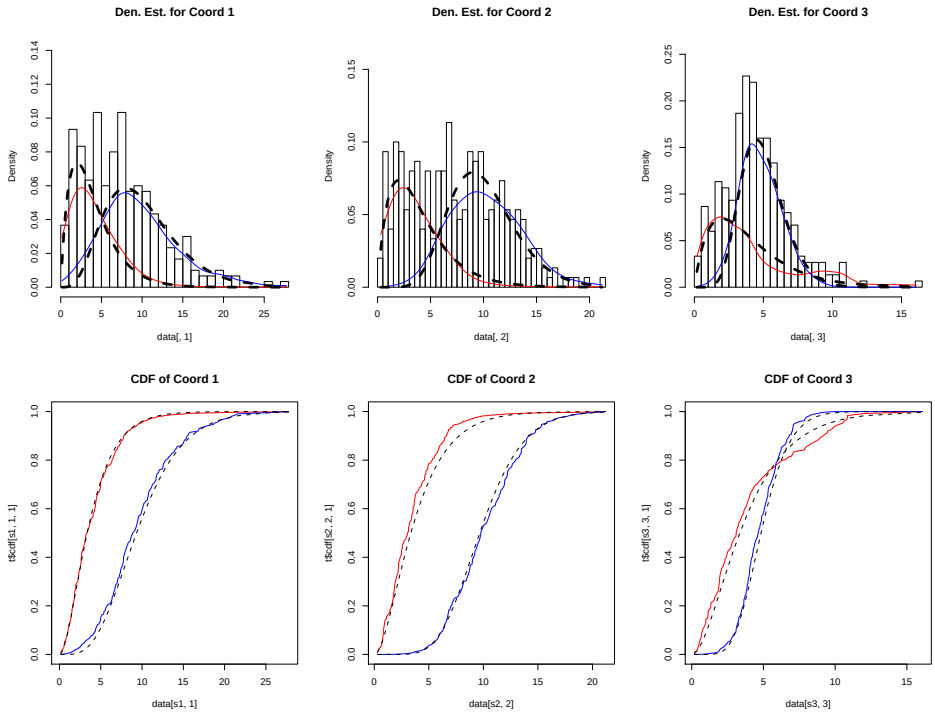


Fig. 3.18. Estimated PDF and CDF for a two component gamma mixture model with conditionally independent coordinates with $\lambda = 0.4$, $\boldsymbol{\alpha}_1 = (2, 2, 2)$, $\boldsymbol{\alpha}_2 = (5, 10, 10)$, $\boldsymbol{\beta}_1 = (2, 2, 2)$, and $\beta_2 = (2, 1, 0.5)$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

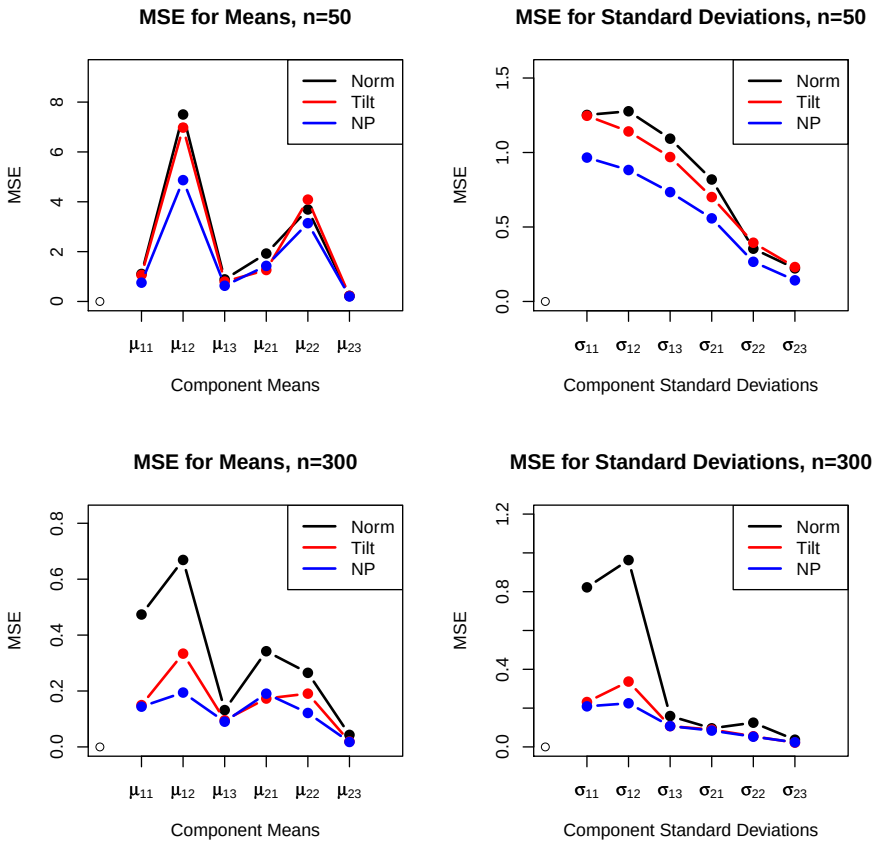


Fig. 3.19. Plots of the MSEs for the tiltedEM, npEM, and the normal mixture for the gamma distributions.

3.8.7 Mixture of Different Distributions

Another interesting example is one where the marginals are not from the same family of distributions. We sampled $n = 50$ and $n = 300$ observations from the following model:

$$\lambda f(x_1; 0, 1)g(x_2; 2, 2)f(x_3; 0, 1) + (1 - \lambda)g(x_1; 2, 2)h(x_2; 0, 1)h(x_3; 0, 4) \tag{3.85}$$

where $\lambda = 0.3$, $f(\cdot; \mu, \sigma^2)$ is an normal distribution, $g(\cdot; \alpha, \beta)$ is a gamma distribution, and $h(\cdot; \mu, b)$ is a double exponential density. We generated 1000 simulations of sample sizes $n = 50$ and $n = 300$ from this model.

The estimates for the means and standard deviations from the tiltedEM are shown in Table 3.21. The standard errors of the estimates decrease significantly as the sample size increases. The results using the normal method and the nonparametric method for $n = 300$ are shown in Table 3.22. Since the normal procedure does not correctly specify this model, we do not expect it to perform well but we do expect the npEM to produce good estimates. We calculated the estimated MSEs for all the estimates from the three methods. Figure 3.21 displays the results for both sample sizes. When the sample size is small, the tiltedEM performs slightly better than the other two. When the sample size increases, however, we see that the npEM and the tiltedEM are very similar for nearly every estimate, and better in one case. When sample size is $n = 300$, the normal method has a high MSE for the mean for the gamma coordinate in the first component. The bias of this estimate is the cause of the high MSE. Interestingly, the MSE for the standard deviation of the first component and third coordinate is higher than those from the tilted and normal methods. We ran another set of these simulations to verify the results. They are shown in Tables B.6 and B.7 in Appendix B. The MSEs are plotted in Figure B.2. The results are similar as those shown in this section.

The plots of the estimated marginal CDFs and PDFs for one of the simulation are shown in Figure 3.20. Again, these distributions do not satisfy the exponential tilt assumption but the estimates are good.

Parameter	True	$n = 50$	$n = 300$
λ	0.3	0.3315(0.0802)	0.3074(0.0310)
μ_{11}	0	1.1281(1.7738)	0.0914(0.2696)
μ_{12}	4	3.0295(1.7461)	3.9223(0.3587)
μ_{13}	0	-0.0201(1.7105)	0.0073(0.1953)
μ_{21}	4	3.6031(1.0114)	4.0041(0.2227)
μ_{22}	0	0.3320(0.8854)	-0.0071(0.1231)
μ_{23}	0	-0.0176(1.0880)	-0.0051(0.4027)
σ_{11}	1.0000	1.6858(1.1363)	1.2945(0.5892)
σ_{12}	2.8284	2.3249(0.9161)	2.7939(0.3370)
σ_{13}	1.0000	1.8732(1.6381)	1.0938(0.2168)
σ_{21}	2.8284	2.6159(0.6092)	2.7695(0.2253)
σ_{22}	1.4142	1.5968(0.6973)	1.4079(0.1617)
σ_{23}	5.6569	5.1092(1.3985)	5.6617(0.4376)

Table 3.21. Two component mixture of different distributions with conditionally independent coordinates with $\lambda = 0.3$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.85) with sample sizes $n = 50$ and $n = 300$.

Parameter	True	Normal	npEM
λ	0.3	0.3444(0.0368)	0.3257(0.0325)
μ_{11}	0	0.2593(0.1984)	0.2525(0.2031)
μ_{12}	4	3.4755(0.3907)	3.8017(0.3383)
μ_{13}	0	0.0053(0.1133)	0.0062(0.2060)
μ_{21}	4	4.1354(0.2317)	4.0330(0.2220)
μ_{22}	0	0.0143(0.1198)	-0.0527(0.1032)
μ_{23}	0	-0.0054(0.4234)	-0.0063(0.4196)
σ_{11}	1.0000	1.1055(0.1346)	1.5225(0.4571)
σ_{12}	2.8284	3.0406(0.3489)	2.8300(0.3338)
σ_{13}	1.0000	1.0332(0.2904)	1.7460(0.8121)
σ_{21}	2.8284	2.8723(0.2274)	2.7910(0.2215)
σ_{22}	1.4142	1.3378(0.1313)	1.3410(0.1112)
σ_{23}	5.6569	5.8176(0.4974)	5.6290(0.4341)

Table 3.22. Results from the normal and nonparametric method based on 1000 simulations of sample size $n = 300$ from model (3.85)

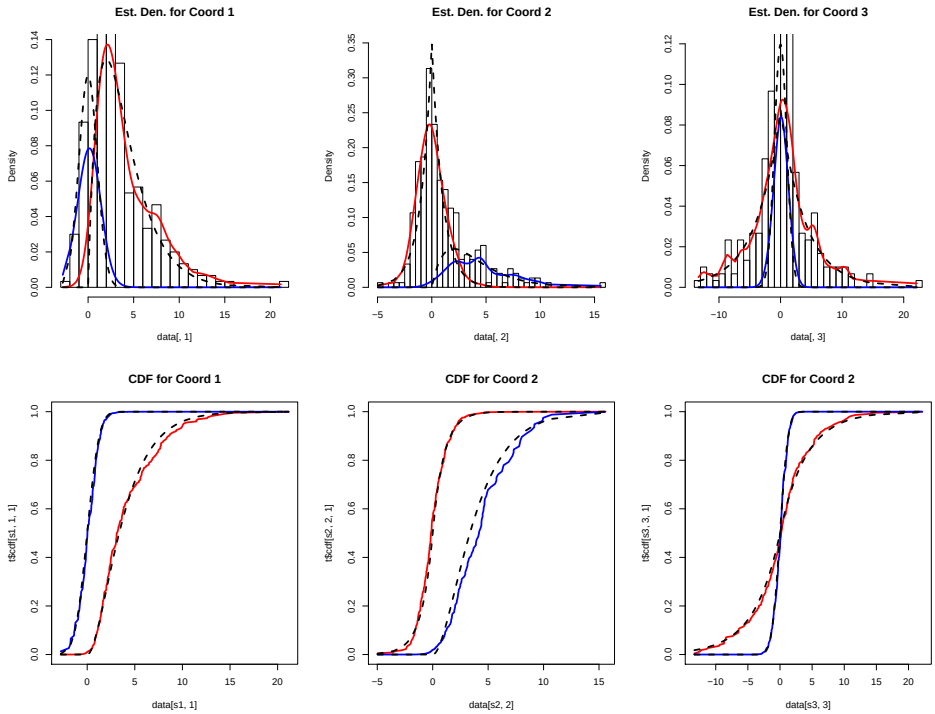


Fig. 3.20. Estimated PDF and CDF for a two component mixture model of different distributions with conditionally independent coordinates with $\lambda = 0.3$. The dashed lines represent the true density functions and the solid lines are the estimates found by the tiltedEM function.

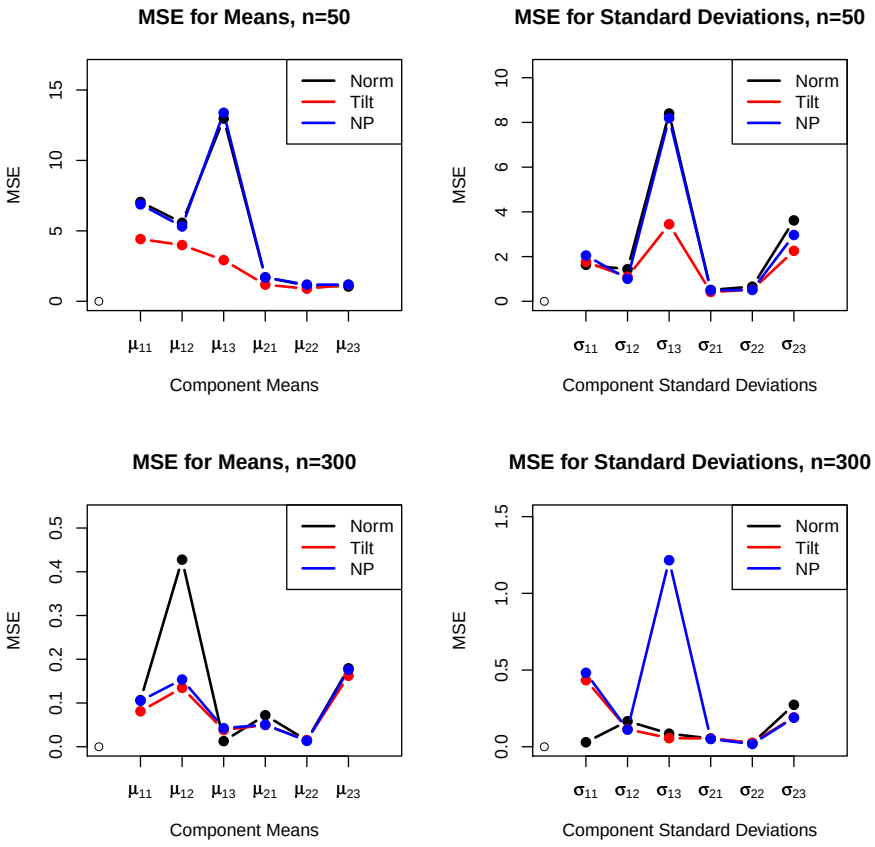


Fig. 3.21. Plots of the MSEs for the tiltedEM, npEM, and the normal mixture for the different distributions.

3.8.8 Mixture of Blocks of Conditionally i.i.d. Coordinates

In this section, we will consider a three component mixture with five coordinates. The first three coordinates are i.i.d and makeup the first block and the last two coordinates are i.i.d and makeup the second block. We generated $n = 300$ observations from the following mixture model

$$h(x_1, \cdot, x_5) = \lambda_1 \prod_{j=1}^{3} \phi\left(x_j; 0, 1\right) \prod_{j=4}^{5} \phi\left(x_j; 0, 4\right) + \lambda_2 \prod_{j=1}^{3} \phi\left(x_j; 2.5, 1\right) \prod_{j=4}^{5} g\left(x_j; 10, 0.5\right)$$
$$+ \lambda_3 \prod_{j=1}^{3} \phi\left(x_j; 5, 1\right) \prod_{j=4}^{5} \mathcal{L}\left(x_j; 10, 1\right), \tag{3.86}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2, $g(\cdot; \alpha, \beta)$ is the gamma density function with mean $\alpha\beta$ and variance $\alpha\beta^2$, and $\mathcal{L}(\cdot; \mu, \beta)$ is the Laplace density function with mean μ and variance 2β.

For this model, we will compare the tiltedEM to the npEM and the normal mixture. The normal mixture model is an incorrect model for this data and the exponential tilt assumption is not valid. Even though the assumptions are not satisfied, we hope the tilted model performs well.

The semiparametric block component means and variance are shown in Table 3.23. The results from Benaglia (2008) using the npEM and a normal mixture are also shown for comparison. We should note that the cut-point model discussed in Elmore (2003) is not capable of handling data beyond the conditionally i.i.d. structure. The three methods presented in the table produce results that are very similar. The methods produce results that are very close to the true values.

A plot of the block component PDFs and CDFs for a randomly selected simulation are shown in Figure 3.22. The semiparametric estimates provide good fits for the true density functions.

		npEM	Normal	tiltedEM
Param.	True	Est.	Est.	Est.
λ_1	0.25	0.2404(0.0244)	0.2503(0.0243)	0.2489(0.0255)
λ_2	0.35	0.3506(0.0276)	0.3509(0.0276)	0.3518(0.0274)
λ_3	0.40	0.3990(0.0280)	0.3988(0.0280)	0.3993(0.0277)
μ_{11}	0	-0.0063(0.0670)	-0.0070(0.0667)	0.0006(0.0710)
μ_{12}	0	-0.0024(0.1661)	-0.0029(0.1656)	-0.0065(0.1692)
μ_{21}	2.50	2.4998(0.0566)	2.5018(0.0576)	2.5019(0.0577)
μ_{22}	5.00	4.9974(0.1106)	4.9949(0.1110)	4.9960(0.1050)
μ_{31}	5.00	4.9976(0.0532)	4.9971(0.0532)	4.9990(0.0546)
μ_{32}	10.00	9.9948(0.0937)	9.9993(0.0930)	10.0028(0.0932)
σ_{11}	1.00	0.9958(0.0970)	0.9952(0.0971)	0.9937(0.0825)
σ_{12}	4.00	3.9784(0.4647)	3.9774(0.4609)	3.9580(0.4699)
σ_{21}	1.00	0.9980(0.0791)	1.0023(0.0807)	0.9975(0.0825)
σ_{22}	2.50	2.4936(0.2804)	2.4930(0.2830)	2.4847(0.2827)
σ_{31}	1.00	0.9971(0.0751)	0.9988(0.0753)	0.9994(0.0758)
σ_{33}	2.00	1.9871(0.2887)	1.9602(0.2787)	1.9910(0.2933)

Table 3.23. Three component mixture with two blocks of conditionally i.i.d. coordinates. The parameters are $\boldsymbol{\lambda} = (0.25, 0.35, 0.40)$. The results are based on 1000 simulations of sample size $n = 300$.

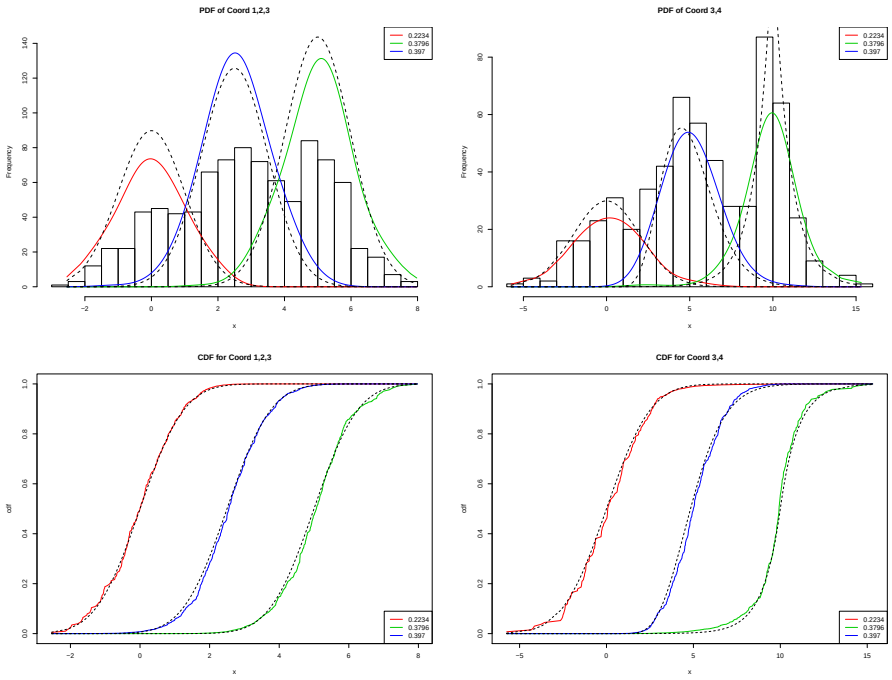


Fig. 3.22. Plots of the semiparametric estimates of the PDFs and CDFs for the three component mixture with two blocks of conditionally i.i.d. coordinates. The dashed lines represent the true density functions and the solid lines are the semiparametric estimates of the density functions.

3.8.9 Conclusion and Discussion

In this section we present various models to judge the performance of the exponential tilt method. For the normal mixture simulations the method shows results that are similar to the true model. This is expected since the exponential tilt assumption is valid for normal density functions. The method gives good results for normal scale mixtures with conditionally i.i.d. coordinates, for normal location models with conditionally i.i.d. coordinations as well as for conditionally independent coordinates. In the normal cases, the method showed similar results compared to the nonparametric method and the cut-point method.

We also included some models that did not fit the exponential tilt assumption such as a mixture of gamma distributions and a mixture of different coordinate distributions. For the gamma mixture with large sample size, it performed better than the incorrect model, the normal mixture, and nearly as good as the purely nonparametric method. For the mixture of different coordinate distributions, the tiltedEM seemed to do at least as good as the nonparametric method when estimating the component means and standard deviations based on the MSEs. The advantage of this method is the moment matching property discussed in Section 3.5. From the plots of the estimates PDFs and CDFs, we can see that they are very close to the true density and cumulative distribution.

Both the tilted method and the nonparametric method are capable of handling block structure in the data. The results of the simulations that included blocks were similar for those two methods. This is an advantage over models that only assume conditionally i.i.d. coordinates. By having the blocking structure capability, this semiparametric method will be able to produce more efficient estimates since it will have more data to estimate the parameters.

The method performs well when the assumption is valid as well as when it is not. Overall it appears that the semiparametric exponential tilt model provides a very flexible model fitting methodology.

3.9 Model Selection

3.9.1 Introduction

So far in this thesis, we have assumed the number of components in the mixture, m, is fixed and known. For some real life situations, the number of components is assumed from theory. In other situations, however, the number of components is unknown and must be determined. Therefore, it is beneficial to have a method to help determine the number of components in the mixture. Our model has the advantage of having a profile likelihood. Since the semiparametric profile likelihood behaves similarly to a parametric likelihood (Murphy and van der Vaart (2000)), we suggest a model selection method based on the Bayesian Information Criterion (BIC, Schwartz (1978)) using the profile likelihood instead of a parametric likelihood, called pBIC.

3.9.2 Bayesian Information Criterion (BIC) and pBIC

It is possible to increase the likelihood by increasing the number of parameters in the model resulting in "over-fitting". This calls for a model selection method that will help determine the "best" subset of parameters for a particular model based on the data. Many model selection methods contain a penalty term to counteract the number of parameters such as BIC, Akaike Information Criterion (AIC), and many others. The basic form of these methods is:

$$-2\ln L + \text{penalty} \tag{3.87}$$

where L is the maximized value of the likelihood. Most of the well known model selection methods differ only by the penalty term. For example, if the penalty equals $2s$, where s is the number of parameters, it is the same as AIC and when it equals $s\ln(n)$, where n is the number of observations, it is equal to BIC (McLachlan and Peel, 2000, pp. 209-210).

Since the semiparametric profile likelihood behaves similarly to a parametric likelihood (Murphy and van der Vaart, 2000), we replace the maximized likelihood in the formula for BIC with the semiparametric profile likelihood and call it the profile BIC

(pBIC):

$$-2\ln L_{P} + s\ln(n) \tag{3.88}$$

where L_{P} is the semiparametric profile likelihood. Since mixture models do not satisfy all of the regularity conditions in Schwartz (1978) we turn to simulations to study the criteria (McLachlan and Peel, 2000).

3.9.3 Simulations of Normal Mixtures

In this section, we use simulations conducted in Benaglia (2008) for a comparison of our model selection method. Benaglia (2008), has a nonparametric multivariate density estimation method which does not contain parameters or a likelihood. Therefore, she uses a selection criterion based on a penalized minimum-distance (MD) (see Chen and Kalbfleisch, 1996) using the Kolmogorov-Smirnov distance. In that thesis, n observations are generated from the following conditionally independent multivariate normal mixture

$$h(x_{1},\ldots,x_{k}) = \sum_{l=1}^{m} \lambda_{l} N\left(\boldsymbol{\mu}_{l}, \sigma_{l}\mathbf{I}_{k}\right) \tag{3.89}$$

where m is the number of components, k is the number of repeated measures, $\boldsymbol{\mu}_{l}$ is the component mean, σ_{l} is the component standard deviation, and $\mathbf{I}_{k}$ is the $k \times k$ identity matrix.

The following three models were used to show the performance of these model selection methods. Each of the models below consisted of 100 simulations. All of the simulations will be special cases of (3.89) varying in sample size (n) and the number of repeated measures (k). For each model considered, the mixing proportions of the components are equal, *i.e.*, for a model with m components, $\lambda_{1} = \lambda_{2} = ... = \lambda_{m}$.

Model 1: Normal location mixtures with two, three, and four components ($m_{0} = 2, 3, 4$). There were $k = 7$ repeated measures with $n = 300$, $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$, and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 7$.

Model 2: Normal location mixtures with two, three, and four components ($m_0 = 2, 3, 4$). There were $k = 10$ repeated measures with $n = 1000$, $(m_{1j}, m_{2j}, m_{3j}, m_{4j}) = (0, 2, 4, 6)$, and $s_{lj} = 1$ for $l = 1, \ldots, m; j = 1, \ldots, 10$.

Model 3: Normal scale mixtures with two and three components ($m_0 = 2, 3$). There were $k = 5$ repeated measures with $n = 500$, $(m_{1j}, m_{2j}, m_{3j}) = (0, 0, 0)$, and $(s_{1j}, s_{2j}, s_{3j}) = (0, 10, 50)$ for $j = 1, \ldots, 5$.

Table 3.24 displays the proportion of times pBIC and MD selected the correct number of components. In the table $s = (2k + 1)(m - 1)$ is the number of parameters estimated in each semiparametric model, k is the number of repeated measure, n is the sample size, m_0 is the true number of components in the mixture, m is the number of components assumed in the calculation of the pBIC, and p is the proportion of times the particular method chose the correct number of components. We can see pBIC performs very well. For the locations mixtures, Models 1 and 2, with the number of components $m_0 = 2$ and $m_0 = 3$, pBIC and MD have accuracy of 100%. When the number of components for these models increases to four, however, the accuracy decreases. pBIC still performs better than the MD with 96% for four components for Model 1 compared to the 81% by MD. For Model 2, pBIC is 98% with MD being 70%. For Model 3, the model with pure scale mixtures, the accuracy decreases but, again, pBIC performs better than MD.

The same models were used for another set of simulations but the sample sizes were decreased. The results are shown in Table 3.25. This small simulation study suggests that pBIC was effective for estimating the number of components in the semiparametric mixture. Our methods does just as good or better than the method used in Benaglia (2008). For Model 3 with smaller sample size, pBIC did not perform well for the scale mixture with $m = 3$. However, when the sample size was increased to 500, the proportion of correct selection increased to 0.90. Overall, the simulations show that the pBIC performed reasonably well.

For each of the three models above, one dataset was chosen from the simulations. We use these datasets to illustrate how to use the pBIC to select the number of components in the mixture. Similarly to the BIC method, the model with the smallest pBIC is

Table 3.24. pBIC simulations results compared with Minimum Distance (MD) method for Models 1-3. The table displays the proportion of times pBIC chose the correct number of components (p).

					$m_0 = 2$		$m_0 = 3$		$m_0 = 4$	
Model	n	k	Method	Compared	s	p	s	p	s	p
1	300	7	pBIC	$1 \leq m \leq 4$	15	1.00	30	1.00	45	0.96
			MD			1.00		1.00		0.81
2	1000	10	pBIC	$1 \leq m \leq 5$	21	1.00	42	1.00	63	0.98
			MD			1.00		1.00		0.70
3	500	5	pBIC	$1 \leq m \leq 3$	11	1.00	22	0.90	–	–
			MD			0.99		0.66		–

Table 3.25. pBIC simulations results for Models 1-3 with smaller sample sizes, where n is the number of observations. The table displays the proportion of times pBIC chose the correct number of components.

		$m_0 = 2$			$m_0 = 3$			$m_0 = 4$		
			Proportion			Proportion			Proportion	
Model	k	s	$n = 100$	$n = 200$	s	$n = 100$	$n = 200$	s	$n = 100$	$n = 200$
1	7	15	1.00	1.00	30	1.00	1.00	45	0.96	0.98
2	10	21	1.00	1.00	42	0.95	0.99	63	0.91	0.97
3	5	11	0.94	0.97	22	0.65	0.67	–	–	–

used to determine how many components should be in the model. Tables 3.26, 3.27, and 3.28 show the pBIC for each of the selected simulations. In the table, ℓ_P represents the maximized value for the log profile likelihood and 's' represents the number of parameters estimated. The row in the table that has bold text is the number of components that corresponds to the model with the smallest pBIC. For Model 1 with two components (Table 3.26), the model with two components gives a better fit than that of three or four components. In Table 3.27, we can see the pBICs for two, three, and four components are close but the value for two components is the smallest. Finally for Model 3 in Table 3.28, the values for pBIC are very close for two and three components. The smallest value is for two components and thus is chosen as the best model.

Table 3.26. Selecting the number of components using pBIC for Model 1 with two components

Components	ℓ_P	s	pBIC
1	-11977.94	0	23955.89
2	**-11430.19**	**15**	**22945.94**
3	-11418.18	30	23007.47
4	-11404.60	45	23065.86

Table 3.27. Selecting the number of components using pBIC for Model 2 with three components

Components	ℓ_P	s	pBIC
1	-69077.55	0	138155.1
2	-65724.25	21	131593.6
3	**-63979.77**	**42**	**128249.7**
4	-63954.51	63	128344.2
5	-63938.87	84	128458.0

3.9.4 Simulations of Gamma Mixtures

We showed above that pBIC performs well for model selection when we have a mixture of normals. Now, we would like to see how well it performs when the mixture is not made of normal distributions but rather mixtures of gamma distributions. The models we simulated from are of the following form:

$$h(X) = \lambda_1 \prod_{j=1}^{k} g(x_j; 2, 2) + \sum_{l=2}^{m} \lambda_l \prod_{j=1}^{k} g(x_j; \alpha_{lj}, \beta_{lj}) \tag{3.90}$$

where $g(\cdot; \alpha, \beta)$ is a gamma density function with mean $\alpha\beta$ and variance $\alpha\beta^2$. The mixing proportions were chosen such that $\lambda_1 = \ldots, = \lambda_m$. We considered the following three models

Model 1 Gamma mixture model with two component ($m_0 = 2$) and three coordinates ($k = 3$). There were 100 simulations of sample size $n = 300$ with $\boldsymbol{\alpha} = (5, 10, 10)'$ and $\boldsymbol{\beta} = (2, 1, 0.5)'$.

Table 3.28. Selecting the number of components using pBIC for Model 3 with two components

Components	ℓ_P	No.	pBIC
1	-15536.52	0	31073.04
2	**-14787.92**	11	**29644.19**
3	-14777.00	22	29690.73

Model 2 Gamma mixture model with three components ($m_0 = 3$) and five coordinates ($k = 5$). There were 100 simulations of sample size $n = 500$ with $\boldsymbol{\alpha}_2 = (10, 10, 10, 20, 7)$, $\boldsymbol{\alpha}_3 = (1, 1, 1, 1, 1)$, $\boldsymbol{\beta}_2 = (1, 1, 0.5, 0.5, 1)$, and $\boldsymbol{\beta}_3 = (2, 2, 2, 2, 2)$.

Model 3 Gamma mixture model with four components ($m_0 = 4$) and five coordinates ($k = 5$). There were 100 simulations of sample size $n = 700$ with $\boldsymbol{\alpha}_2 = (10, 10, 10, 20, 7)$, $\boldsymbol{\alpha}_3 = (1, 1, 1, 1, 1)$, $\boldsymbol{\alpha}_4 = (5, 5, 5, 5, 5)$, $\boldsymbol{\beta}_2 = (1, 1, 0.5, 0.5, 1)$, $\boldsymbol{\beta}_3 = (2, 2, 2, 2, 2)$, and $\boldsymbol{\beta}_4 = (2, 2, 2, 2, 2)$.

The results from the simulations are shown in Table 3.29. The column m_0 represents the true number of components for each model and p is the proportion of times pBIC showed the correct number of components. For Model 1, pBIC produced the correct number of components in 99% of the simulations. For Model 2, it chose the correct number of components 81% of the time and for Model 4 it was correct 74% of the time. It is worth noting that the gamma mixtures of Models 1-3 do not satisfy the exponential tilt assumption but the method still performs fairly well.

Model	Compared	m_0	p
1	$1 \leq m \leq 4$	2	0.99
2	$1 \leq m \leq 5$	3	0.81
3	$1 \leq m \leq 5$	4	0.74

Table 3.29. pBIC simulations results for Model 1-3 using (3.90). The table displays the proportion of times pBIC chose the correct number of components (p)

3.10 Real Data Examples

In this section, we present the analysis of two real datasets. The first is the Reaction Time dataset as described in Cruz-Medina et al. (2004) and the second is the Water level dataset described by Benaglia et al. (2008). For both examples, we compare results from previous works where appropriate. Also presented are bootstrap standard errors of the component means and standard deviations. We used two methods to produce the bootstrap samples. The first method is the nonparametric bootstrap and the second is the weighted bootstrap as described in Benaglia (2008).

We considered the standard errors estimates based on the formula found in Elmore (2003):

$$\widehat{\mathrm{SE}}(\hat{\mu}_{la}) = \sqrt{\frac{\hat{\sigma}^2_{la}}{nC_a\hat{\lambda}_l}}. \tag{3.91}$$

where $\hat{\mu}_{la}$ is the l-th estimated component mean for the a-th block, $\hat{\sigma}^2_{la}$ is the l-th component standard deviation for the a-th block, C_a is the number of coordinates in block a, $l = 1, \cdots, m$, $a = 1, \cdots, B$. Elmore (2003) points out that this estimate is biased in the downward direction but performed quite well in the simulations. We found the standard errors based on (3.91) to be seriously biased downward and did not provide them here.

3.10.1 Reaction time data

This data comes from a cognitive experiment discussed in Miller et al. (2001) (also see Cruz-Medina et al. (2004)) and is available at `http://www.blackwellpublishing.com/rss`. The experiment recruited 9 year-old children that were largely normally developing with some of the children having certain language impairments. The children were subjected to several task conditions. The task condition of interest here measures the reaction time (RT), in milliseconds, for a child to give the correct response to a visual stimuli. The visual stimuli for this task involves two two-dimensional images on a computer monitor. The left image is

the target stimulus and the right image is either identical to the target stimulus or a mirror image of the target image. The child was to press one key indicating if he/she thought the right image was identical to the target or another key if they thought it was the mirror image. There were $k = 6$ trials for this task and the reaction time (RT) in milliseconds for the child to choose their answer was recorded for each trial. Of all the children in this experiment, $n = 197$ choose the correct response for all six trials. We will consider only these children. The six trials of the reaction time task were embedded in a sequence of trials for several other tasks. The children could not anticipate when the trials for this task would appear within sequence and each trial contained a different two-dimensional shape. In this setup, the conditional independence assumption seems reasonable. For more discussion of the conditional independent assumption see Section 3.10.1.2.

3.10.1.1 Selecting the number of components for RT data

We applied the pBIC model selection method of Section 3.9 to this dataset using $m = 1, 2, 3,$ and 4 components. Table 3.30 shows the results. The goal is to choose the smallest pBIC. In this case, the model with the smallest pBIC is one with three components. It does appear however that the model with two components has pBIC close to that for three components. It may be argued that we could also use the two component model. Benaglia (2008) obtained similar results as the tilted method and Cruz-Medina et al. (2004) found three components using their method but presented an argument for two components. We will continue the analysis using three components.

Table 3.30. pBIC for RT data

Components	ℓ_P	Number of Parameters	pBIC
1	-6244.747	0	12489.49
2	-6123.371	13	12315.42
3	**-6081.632**	**26**	**12300.63**
4	-6068.395	39	12342.84

3.10.1.2 Validity of the conditional independence assumption

In this section we address the validity of the conditional independence assumption with regards to the Reaction Time data. Recall the conditional independence assumption means that conditional on the group membership, the RTs for each subject are independent. To check this condition we calculated two sets of correlations. One set is the sample correlations and the other set contains the correlations calculated assuming the conditional independence assumption is true. Next, Fisher's transformation was performed on both sets. The plot of Fisher's transformation of the correlations assuming conditional independence (log.r.ind) versus Fisher's transformation of the sample correlations (log.r) is shown in Figure 3.23. Upper and lower bounds of $2\frac{1}{\sqrt{n-3}}$ are also included on the plot.

Based on the plot, all of the transformed correlations assuming conditional independence are within the upper and lower bounds for the transformed sample correlations. Therefore, this suggests that the correlation between the coordinates is due to the mixture. This provides some support for the conditional independence assumption.

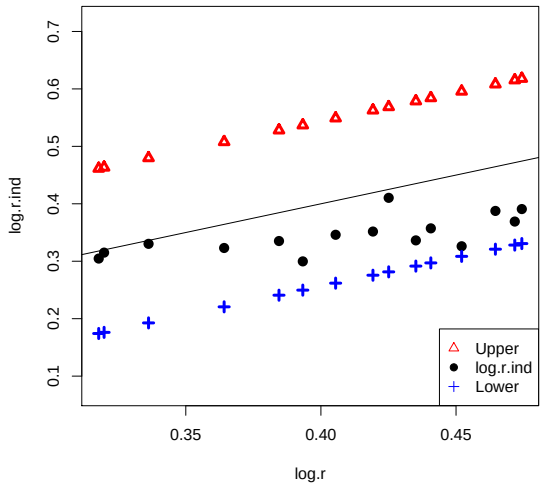


Fig. 3.23. Plot of the transformed correlations assuming conditional independence against the transformed sample correlations.

3.10.1.3 Evaluating the data

The data can be written as $x_{ij}, i = 1, 2, ..., 197; j = 1, \ldots, 6$. The model under consideration is

$$H(y_1, \ldots, y_6) = \lambda_1 \prod_{j=1}^{6} F_j(y_j) + \sum_{l=2}^{3} \lambda_l \prod_{j=1}^{6} G_{lj}(x_j). \tag{3.92}$$

with $G_{lj}(x_j)$ given by (3.44). Using our method, we can estimate the marginal CDFs, means and standard deviations of the three components of the mixture model for each coordinate by using (3.44), (3.46) and (3.47), for $l = 1, 2, 3$ and $j = 1, \ldots, 6$. Table 3.31 shows the estimated component means and standard deviations for the RT data using the semiparametric method, the normal mixture assuming conditionally independence and the nonparametric method proposed by Benaglia et al. (2008). Figure 3.24 shows the estimated component cdfs. The component means and standard deviations are similar for exponential tilt model and the normal mixture model. From the Table, we can see that the smallest component, approximately 20% of the children, have the shortest mean reaction times. The second group, consisting of about 31% of the children, have the next smallest reaction times. The largest group, about 49%, consist of the children that have the slowest reaction times. These results differ from the nonparametric method that has the shortest RT group being about 31%, the next group being 43%, and the group with the largest reaction times being only 18%.

Elmore (2003) considered the repeated measures to be conditionally i.i.d. We conduct the likelihood ratio test for all the repeated measures being conditionally i.i.d. (or one block) and three components. The log profile likelihood is -6104.615 and the likelihood ratio statistic using the log profile likelihood in Table 3.30 is 45.966. Therefore there is no evidence to suggest that the repeated measures are conditionally i.i.d. but we present the conditionally i.i.d results just for comparison. The results using the cut point method, the tilted method, the nonparametric method (both with equal bandwidth and different bandwidths), and the normal method are shown in Table 3.32. In the table, Cut Point^1 is what is presented in Elmore (2003) and Cut Point^2 was found in `mixtools`.

From Table 3.32, Cut Point^1 provide the results in Elmore (2003) and Cut Point^2 give the results with a larger log likelihood using `mixtools`. Also, we provide two results using the nonparametric method: npEM using the same bandwidth for each component and npEM* using different bandwidths. The results using the normal mixture method, the Cut Point^1 method, the tiltedEM, and the nonparametric method (npEM) are similar. The smallest group, about 20% for the tilted method, are those with the fastest reaction times. The next largest group, about 25% have the slowest reaction times. The results using Cut Point^2 and npEM* were not similar.

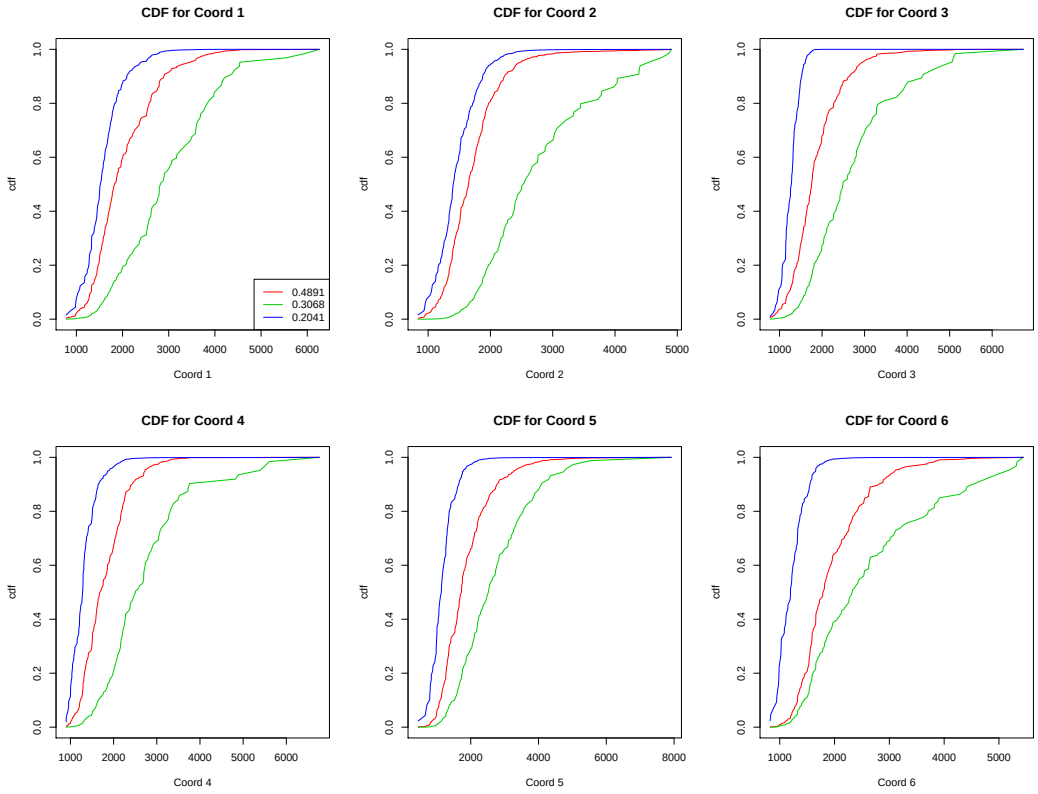


Fig. 3.24. Semiparametric estimation of the CDFs for the Reaction Time (RT) data $F_j, G_{lj},\ j = 1, \ldots, 6,\ l = 2, 3$ under the exponential tilt model.

Table 3.33 shows the estimated standard errors for the conditionally independent case with three components. The table contains the results from 100 bootstrap samples using the nonparametric bootstrap (Boot.) and the weighted bootstrap (W.Boot.). The

	Tilted	Normal	NP
λ_1	0.2041	0.2109	0.1834
μ_{11}	1577.285	1561.746	3120
μ_{12}	1456.347	1491.968	2920
μ_{13}	1265.697	1285.378	2960
μ_{14}	1312.848	1262.763	3090
μ_{15}	1171.741	1197.136	2980
μ_{16}	1216.518	1239.846	3270
σ_{11}	420.5318	387.067	1120
σ_{12}	337.1962	361.436	921
σ_{13}	200.5727	218.882	1220
σ_{14}	332.9831	200.524	1190
σ_{15}	402.6174	321.639	1240
σ_{16}	261.1347	246.518	1130
λ_2	0.3068	0.2737	0.3860
μ_{21}	3024.508	3001.941	1730
μ_{22}	2776.833	2844.746	1480
μ_{23}	2761.526	2840.340	1390
μ_{24}	2771.616	2779.317	1480
μ_{25}	2729.925	2876.759	1490
μ_{26}	2661.687	2753.632	1580
σ_{21}	1074.7024	1148.530	544
σ_{22}	907.8391	976.440	338
σ_{23}	1101.4164	1163.328	421
σ_{24}	1097.1932	1182.511	451
σ_{25}	1162.0580	1280.091	801
σ_{26}	1180.5029	1227.115	651
λ_3	0.4891	0.5154	0.4306
μ_{31}	2024.910	2113.257	2320
μ_{32}	1712.228	1733.232	2040
μ_{33}	1864.909	1880.391	2170
μ_{34}	1799.368	1884.538	2000
μ_{35}	1870.053	1846.019	2020
μ_{36}	1957.992	1954.508	1890
σ_{31}	691.9057	772.155	886
σ_{32}	469.8500	429.312	700
σ_{33}	609.5184	574.428	674
σ_{34}	516.4987	525.220	614
σ_{35}	777.8147	662.308	760
σ_{36}	636.0090	617.279	553

Table 3.31. Estimated component means and standard deviations for the RT data with three components. The Tilted column represents the estimates based on the semiparametric estimation method. The Normal column represents the estimates based a mixture of normal distributions.

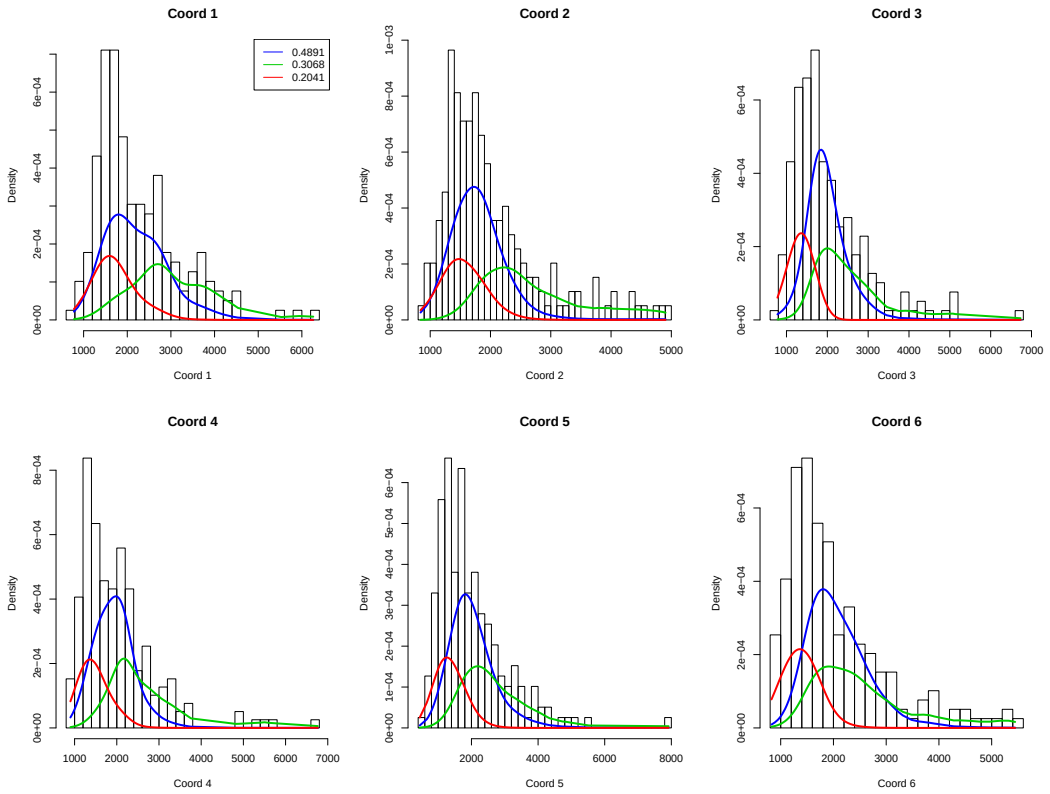


Fig. 3.25. Semiparametric estimation of the PDFs for the Reaction Time (RT) data $f_j, g_{lj},\ j = 1, \ldots, 6,\ l = 2, 3$ under the exponential tilt model.

standard errors for the bootstraps were calculated using $0.75 * IQR$, where IQR is the interquartile range. They were calculated this way as a robust estimate for the standard error. The nonparametric bootstrap means are shown in Figure 3.28. The results for the other parameters and the results from the weighted bootstrap were similar.

In conclusion, the tilted method found three components using pBIC which was also found using the nonparametric method. The results using the tilted method were similar to those of the normal method but not for the nonparametric method. It was suggested by Elmore (2003) that the repeated measures are conditionally i.i.d. however we found the conditional independent model is the better fit using the likelihood ratio test.

Parameter	Normal	Cut Point[1]	Cut Point[2]	tiltedEM	npEM	npEM*
λ_1	0.1992	0.1567	0.0352	0.1987	0.2159	0.1702
λ_2	0.2729	0.2926	0.3534	0.2555	0.2622	0.3449
λ_3	0.5279	0.5507	0.6113	0.5458	0.5220	0.4849
μ_1	1330.765	1321.774	1033.511	1334.920	1375.876	3109.284
μ_2	2851.239	2808.088	2707.462	2887.046	2872.941	1473.139
μ_3	1893.562	1841.303	1716.684	1905.168	1901.579	2073.479
σ_1	303.951	472.586	185.7194	368.065	465.435	1209.890
σ_2	1179.021	1087.774	1058.667	1127.709	1114.812	508.198
σ_3	601.029	636.493	606.8727	657.020	648.977	696.573

Table 3.32. Estimated component means and standard deviations using the cut-point model, tiltedEM, normal, and npEM (npEM* has different bandwidths). Cut Point[1] has log likelihood of -1393.387 and Cut Point[2] has log likelihood of -1391.751.

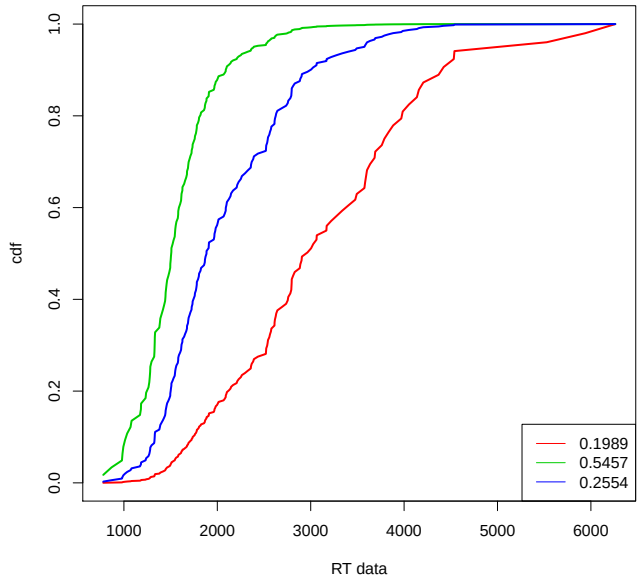


Fig. 3.26. Semiparametric estimation of the CDFs for the Reaction Time (RT) data with i.i.d. repeated measures.

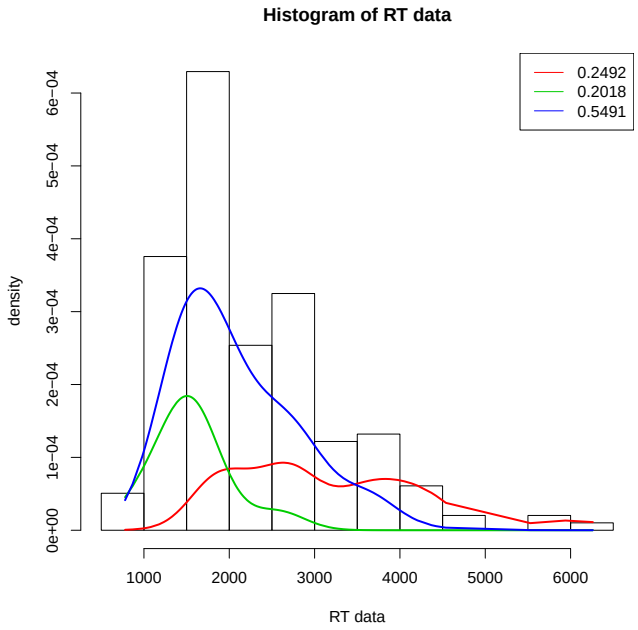


Fig. 3.27. Semiparametric estimation of the PDFs for the Reaction Time (RT) data with i.i.d. repeated measures.

Par.	Est.	Boot.	W. Boot.	Par.	Est.	Boot.	W. Boot
μ_{11}	1577.185	108.00	135.00	σ_{11}	420.532	108.975	110.057
μ_{12}	1456.347	60.00	69.00	σ_{12}	337.196	36.075	63.349
μ_{13}	1265.687	141.00	196.50	σ_{13}	200.573	233.475	162.123
μ_{14}	1312.848	116.25	129.75	σ_{14}	332.983	83.250	116.051
μ_{15}	1171.741	215.25	175.50	σ_{15}	402.617	207.975	258.515
μ_{16}	1216.518	283.50	309.75	σ_{16}	261.135	298.950	214.756
μ_{21}	3024.508	270.00	366.00	σ_{21}	1074.702	178.875	193.301
μ_{22}	2776.833	235.50	363.75	σ_{22}	907.839	173.875	195.712
μ_{23}	2761.526	219.75	283.50	σ_{23}	1101.416	129.300	323.767
μ_{24}	2771.616	219.75	300.00	σ_{24}	1097.193	172.725	321.330
μ_{25}	2719.925	250.50	289.50	σ_{25}	1162.058	194.250	346.131
μ_{36}	2661.687	146.25	222.00	σ_{26}	1180.503	170.850	194.441
μ_{31}	2024.910	249.00	219.75	σ_{31}	691.906	162.00	181.668
μ_{32}	1712.228	265.50	244.50	σ_{32}	469.850	118.875	145.548
μ_{33}	1864.909	313.50	420.75	σ_{33}	609.518	187.200	260.762
μ_{34}	1799.368	414.75	337.50	σ_{34}	516.499	270.525	299.504
μ_{35}	1870.053	322.50	411.00	σ_{35}	777.815	336.075	299.561
μ_{36}	1957.992	345.75	486.00	σ_{36}	636.009	177.00	270.498

Table 3.33. The nonparametric bootstrap (Boot.) and the weighted bootstrap (W. Boot.) standard error estimates for the Reaction Time dataset with three components.

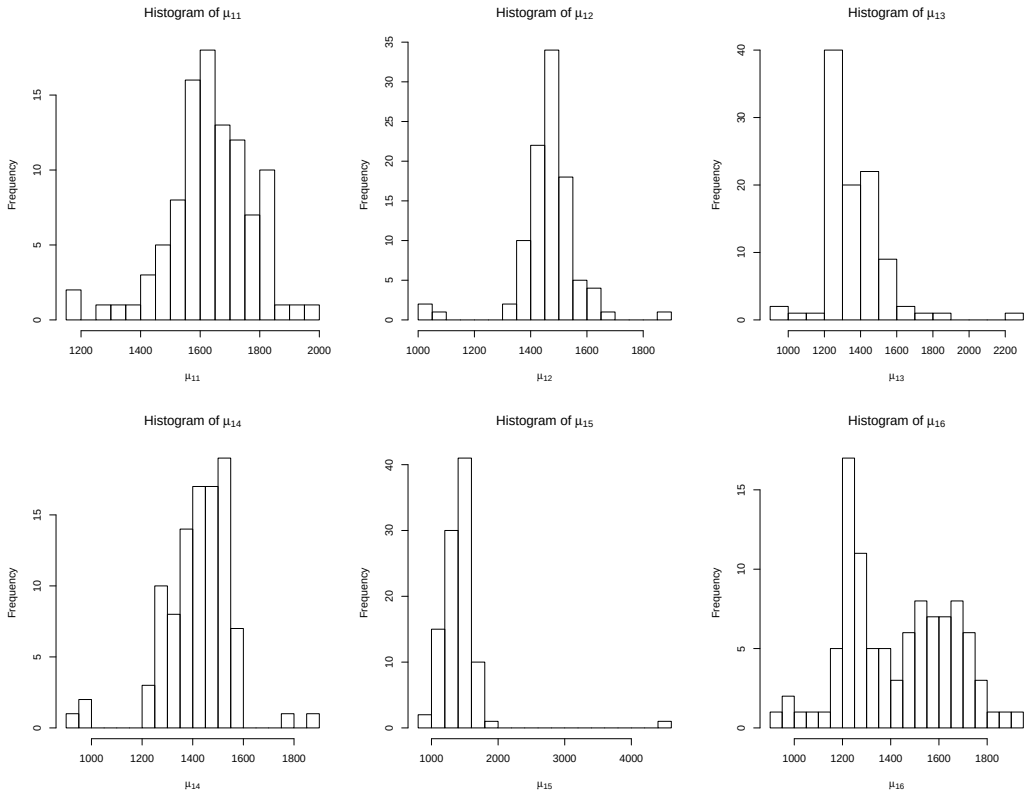


Fig. 3.28. Histogram of 100 nonparametric bootstrap means for the first component for the RT data.

3.10.2 Water level data

This data comes from a cognitive experiment discussed in Thomas and Lohaus (1993). In this experiment, 405 children ages 11 to 16 were selected for an experiment to determine spatial independence. There were eight tasks for this particular experiment. Each child was given eight sheets of paper each with a picture of a water vessel tilted at the following orientations: 1, 2, 4, 5, 7, 8, 10, and 11 o'clock. The child was to draw a line across the vessel indicating where he/she thought the water level should be. The researchers drew straight line between the two points that intersected the vessel. Then they recorded the angular error from zero (or horizontal). The measurement recorded was the angular error from horizontal multiplied by the sign of the slope of the line. Based on the clock orientations, it is reasonable to suggest the recorded measurements at orientations (1,7), (2, 8), (4,10), and (5, 11) (corresponding to coordinates (7,4),(3,8), (6,1), and (2,5)) could be i.i.d. Therefore, we will consider the model with these four blocks. We will conduct the likelihood ratio test for this structure.

In the following sections, we first examine the conditional independence assumption, then we test for a particular block structure, and finally we select the number of components. The final results are shown in Section 3.10.2.2. We compare the results to those from the normal mixture and the nonparametric method suggested by Benaglia et al. (2008).

3.10.2.1 Selecting the number of components for water-level data

We applied the pBIC model selection method (Section 3.9) to the water-level data with both the conditionally independent assumption and the assumption of blocks of conditionally independent coordinates. Table 3.34 shows the results for the conditionally independent case. The smallest pBIC is the model that will "best" fit the data. From the table, the model with the smallest pBIC is one with 4 components. Benaglia (2008) also found that four components best fits the data. In her thesis, she uses the minimum distance method to determine the number of components. For a comparison of pBIC and the minimum distance (MD) method, see Section 3.9. Table 3.35 shows the results when we assume a model with a blocking structure. The model with the smallest pBIC

is 5 components with the pBIC with four components close. We choose to continue with four components based on the two tables.

Components	ℓ_p	No.	pBIC
2	-18769.84	17	37641.75
3	-18583.09	34	37370.81
4	**-18405.68**	**51**	**37117.56**
5	-18376.57	68	37161.40

Table 3.34. pBIC results for choosing the number of components using the Water-level data. In the table, ℓ_p is the log profile likelihood, No. is the number of components.

Components	ℓ_p	No.	pBIC
2	-18775.38	9	37604.79
3	-18600.26	18	37308.59
4	-18440.13	27	37042.36
5	**-18398.78**	**36**	**37013.70**

Table 3.35. pBIC results for choosing the number of components using the Water-level data with a block structure. In the table, ℓ_p is the log profile likelihood, No. is the number of components.

3.10.2.2 Evaluating the data

The next step is to determine if the coordinates are conditionally independent or if the block structure is the correct model. Therefore, we conduct the likelihood ratio test described in Section 3.7. The log likelihood for the conditionally independent case with four components is -18405.68 and the log likelihood for the model with the four blocks is -18440.13. Thus, the likelihood ratio statistic is 68.9. When we compare with a Chi-square with 24 degrees of freedom, we fail to reject the null hypothesis and conclude that the model with the conditionally independent coordinates is the better fit. We will proceed, however, using to blocks to compare our results with those in Benaglia (2008).

To check the validity of the conditional independence between the four blocks of data, we applied the technique described in Section 3.6. For this dataset, we have 4 blocks therefore we have $\binom{4}{2} = 6$ correlations between the blocks and 4 correlations for

the coordinates within each block. Figure 3.29 shows the plot of the correlations; see Section 3.6. Since all the points do not fall within the bounds, the data may not be consistent with the assumption that the blocks of coordinates are conditionally independent. However, since most are within the bounds we proceed as if the assumption holds.

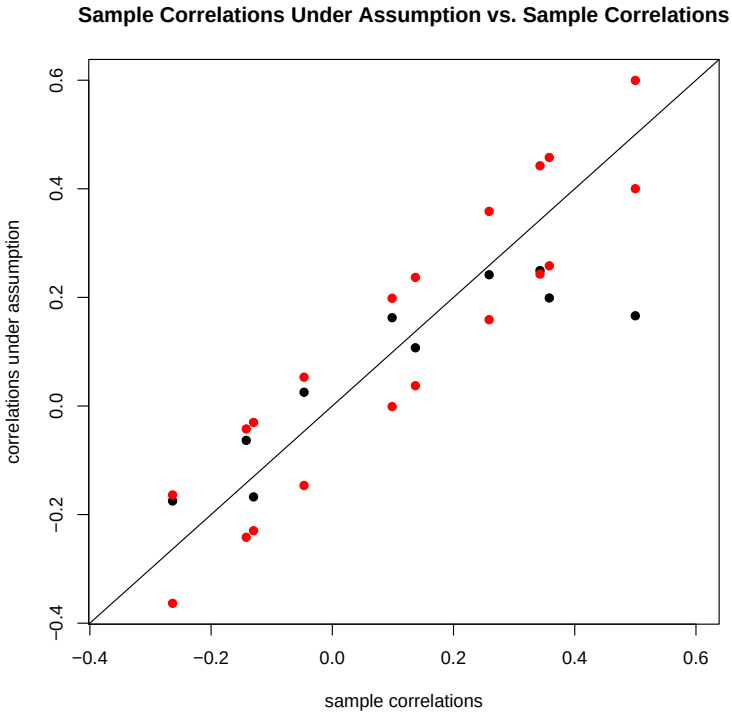


Fig. 3.29. The plot of the correlations for the Water Level with four components and four blocks. The red points represent the bounds and the block points are the correlations.

Table 3.36 shows the estimated component means and standard deviations for the four block using the tilted method (TiltedEM), the nonparametric method (NP^1) using a common bandwidth equal to 4, the nonparametric method (NP^2) using the option for different bandwidths, and the normal mixture method (Normal). The three methods produce very different results. Figure 3.31 shows the estimated component densities for the four blocks of conditionally i.i.d. coordinates.

The results for the tilted method show the largest group, approximately 35% of the children, have measurements close to 0 and have small standard deviations. This

group of children are those who understand the concept and consistently draw the line indicating the water level as horizontal. The next group, consisting of approximately 39% of the children, do not seem to understand as much as the other group. The standard deviations are larger as well as the means. Next, we find a group of 23% of the students. The children in this group have the largest standard deviation. From Figure 3.31, we can see that the estimated component CDFs looks almost like that of a Uniform distribution. Perhaps the children in this group are those that do not understand the concept and are guessing. The last group, consisting of about 3% of the data, are those that draw the water line parallel to the top of the vessel. Notice how these measurements are centered around 30, -30, 60, and -60 degrees.

The results from the nonparametric method using the same bandwidth show mixing proportions different from those of the tilted method. The largest group, about 48% of the data, consist of the students that understand the concept. The next largest group, about 36%, consist of the children that understand the concept but not as well as the largest group. Next, the group of children that seemed to guess, consists of about 12%. The smallest group, about 5% of the data, consist of the students that seem to draw the line parallel to the top of the vessel. However, the results change significantly when their are unequal bandwidths for each of the component (NP^2 in the table). Choosing a bandwidth is not an easy task. One of the advantages for the tilted method is not having to choose a bandwidth. Also, the tilted method does provide results similar to the nonparametric method. The normal method results do not provide results similar to either the nonparametric method or the tilted method. It does seem that assuming the component densities are normally distributed is not correct. It is interesting that both the nonparametric and the tilted method have similar interpretations of the data even though their mixing proportions are not the same.

The bootstrap standard errors for the component means and standard deviations using both the nonparametric bootstrap (Boot.) and the weighted bootstrap (W.Boot) are show in Table 3.37 for 100 bootstrap samples. The bootstrap standard errors were found by computing $0.75 * IQR$, where IQR is the interquartile range. We choose this calculation due to the skewness of the estimates. We plotted the histograms for

the 100 nonparametric bootstrap means for the first component in Figure 3.30. The bootstrap estimates for the other components and for the weighted bootstrap estimates were similar.

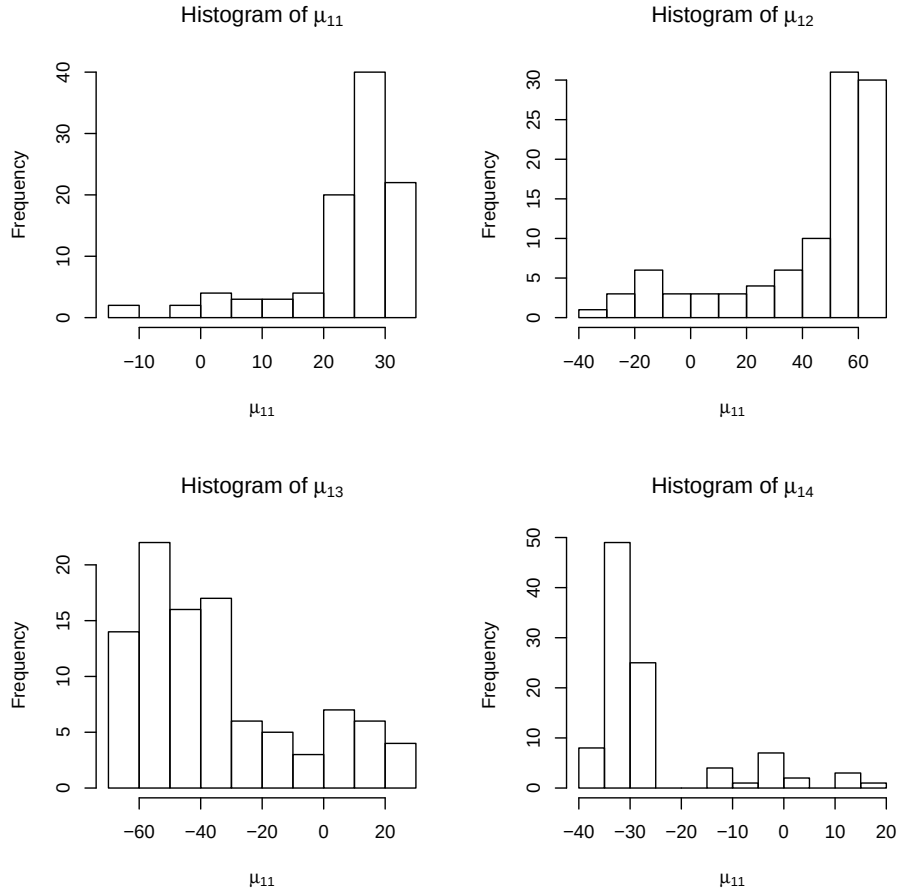


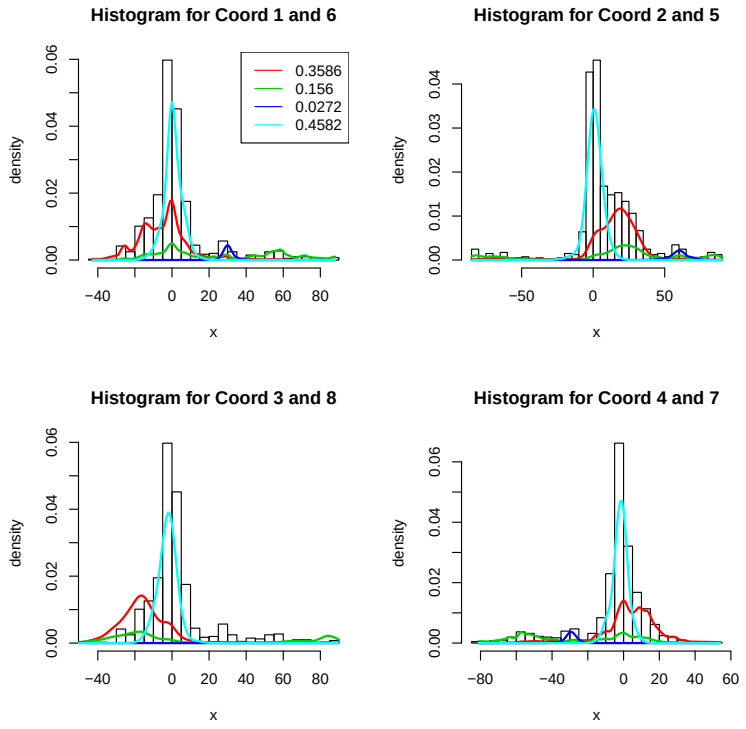
Fig. 3.30. The histograms of 100 nonparametric bootstrap means for the first components for the Waterdata.

Parameter	TiltedEM	NP[1]	NP[2]	Normal
λ_1	0.0296	0.049	0.0302	0.1035
λ_2	0.2258	0.117	0.1057	0.1566
λ_3	0.3513	0.355	0.3793	0.1603
λ_4	0.3933	0.478	0.4590	0.5795
μ_{11}	29.669	28.2	22.40	1.1832
μ_{12}	60.494	58.2	-27.80	2.1655
μ_{13}	-61.454	-48.2	-4.43	0.2457
μ_{14}	-30.621	-31.0	-4.43	0.6412
μ_{21}	15.196	18.0	29.10	-0.431
μ_{22}	15.771	-0.50	47.0	-0.523
μ_{23}	-10.410	0.30	-34.20	-1.392
μ_{24}	-18.286	-22.90	-34.2	-1.452
μ_{31}	0.482	-1.90	0.090	-0.619
μ_{32}	0.350	15.60	14.30	-0.270
μ_{33}	-1.519	-14.50	-13.20	-4.850
μ_{34}	-1.602	0.50	-3.24	-3.705
μ_{41}	-3.854	0.30	0.64	5.073
μ_{42}	8.393	0.90	0.30	14.973
μ_{43}	-10.308	-2.70	-2.47	-13.438
μ_{44}	-1.844	-1.70	-2.20	-7.105
σ_{11}	3.182	12.0	45.30	2.780
σ_{12}	8.475	16.3	37.80	2.718
σ_{13}	5.061	36.2	40.60	2.777
σ_{14}	5.630	10.2	40.60	2.768
σ_{21}	28.170	34.6	18.10	1.812
σ_{22}	40.672	49.0	40.60	2.027
σ_{23}	44.694	51.9	14.60	1.945
σ_{24}	28.824	35.2	14.60	1.912
σ_{31}	3.065	14.8	18.20	6.001
σ_{32}	2.876	16.9	22.70	4.934
σ_{33}	2.901	18.0	24.50	4.378
σ_{34}	3.182	16.4	21.20	5.083
σ_{41}	11.294	5.30	5.59	22.546
σ_{42}	15.392	5.30	6.61	30.189
σ_{43}	12.733	4.30	6.35	31.749
σ_{44}	13.462	5.10	6.35	23.906

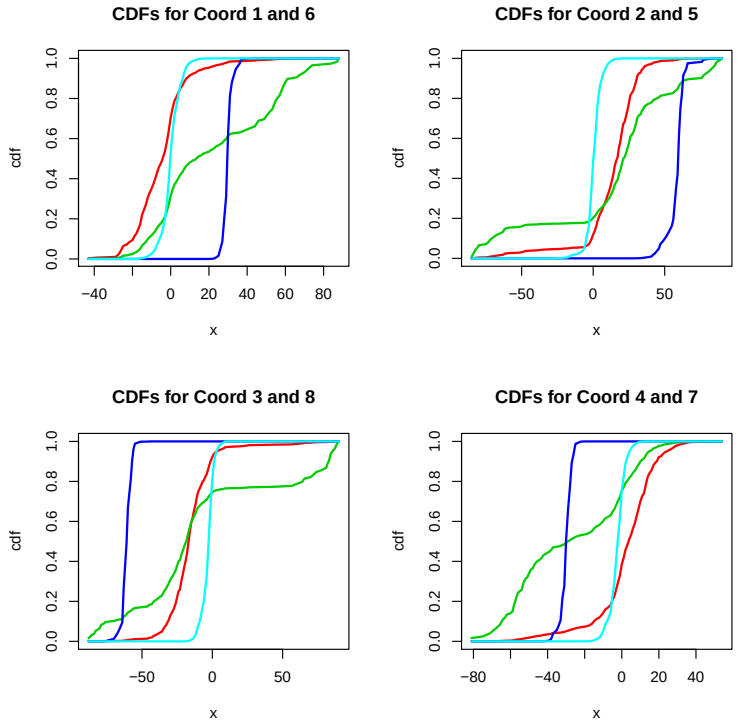
Table 3.36. The estimates for the Water Level data with four components and a block structure with four blocks. The notation for μ and σ are written as component then block.

Par.	Est.	NP Boot.	W.Boot	Par.	Est.	Boot.	W. Boot
μ_{11}	29.669	4.537	4.035	σ_{11}	3.182	3.656	2.796
μ_{12}	60.494	16.11	17.258	σ_{12}	8.475	2.820	2.617
μ_{13}	-61.454	7.988	9.263	σ_{13}	5.061	15.645	16.042
μ_{14}	-30.621	1.425	1.875	σ_{13}	5.061	3.889	2.584
μ_{21}	15.196	4.403	3.604	σ_{21}	28.170	4.553	4.275
μ_{22}	15.771	6.289	6.005	σ_{22}	40.672	5.257	8.888
μ_{23}	-10.410	6.043	6.878	σ_{23}	44.694	5.707	5.775
μ_{24}	-18.286	6.075	8.010	σ_{24}	28.824	5.798	5.940
μ_{31}	0.482	1.656	1.464	σ_{31}	3.065	5.104	4.418
μ_{32}	0.350	8.124	7.996	σ_{32}	2.876	5.500	7.026
μ_{33}	-1.519	2.723	2.085	σ_{33}	2.901	4.902	6.221
μ_{34}	-1.602	0.727	1.036	σ_{34}	3.182	5.843	6.247
μ_{41}	-3.854	1.088	0.210	σ_{41}	11.294	0.525	0.520
μ_{42}	8.393	0.293	0.266	σ_{42}	15.392	0.696	0.624
μ_{43}	-10.308	5.900	0.021	σ_{43}	12.733	0.272	0.366
μ_{44}	-1.844	0.245	0.217	σ_{44}	13.462	0.229	0.227

Table 3.37. The bootstrap standard errors for the component means and standard deviations for Water Level data with four components and a block structure with four blocks. The notation for μ and σ are written as component then block.



(a) The estimated component densities of the Water Level data with four components.



(b) The estimated component CDFs of the Water Level data with four components.

Fig. 3.31. The semiparametric estimates for the Water Level data with four components

Chapter 4

Discussion and Future Work

In this thesis, we developed semiparametric methods for fitting mixtures using an exponential tilt model for both the univariate and multivariate case. Although we did not prove the parameters are identifiable for the univariate case, the results show the method performs well. In the multivariate case, the assumption of conditional independence was needed for identifiability. In this last chapter, we will summarize the findings from these methods and also discuss future work with these methods..

4.1 Discussion

In Chapter 1, we discuss motivation for choosing the exponential tilt model to estimate the component densities of the mixture. The exponential tilt model has been shown be a flexible model and provide good estimates of the densities even when the exponential tilt assumption is not valid for the simulated model. For the multivariate model, we require k measurements on each subject. Our model is not able to handle data where there are missing measurements for a subject. We also need to assume there are m groups (or components) in the population. If theory does not suggest the number of components, we developed a tool in the multivariate case to choose the number of components based on the data.

We discussed identifiability and previous work for mixture models. The issue of identifiability for the univariate case still needs to be addressed. Previous work in the univariate case includes parametric methods and a semiparametric method where the component densities are found using a kernel density estimate. The semiparametric exponential tilted approach would be a balance between these two. For the multivariate case, the parameters in the model are identifiable as long as the assumption of conditional independence holds. Again, the semiparametric method would be a good balance

between parametric and nonparametric methods because it will have the advantage of being robust and also the advantage of having a likelihood.

In Chapter 2, we introduce the semiparametric method for the univariate case. We showed in detail the method proposed by Efron and Tibshirani (1996). We extend their method to the univariate mixture model setting by modeling the data as a mixture of Multinomials. Although identifiability is not shown for the univariate model in this thesis, the method provided satisfactory estimates for the component means and standard deviations and mixing proportions. Since the normal density is an example of an exponential tilt distribution the simulations showed that it provided estimates very similar to those from a normal mixture method. In some instances, it performs better than the method proposed by Benaglia et al. (2009a) even for a symmetric-location mixture. A sufficient condition for identifiability in the method proposed by Benaglia et al. (2009a) is a symmetric-location assumption, the function in `mixtools`, `spEMsymloc`, cannot handle scale mixtures. The tilted method is able to handle models with unequal variances and for skewed distributions. The simulations we provided have large sample sizes ($n > 200$) and we have not explored cases with smaller sample size. Since the number of bins or breaks can be large without compromising the estimates, the method should be able to handle smaller sample size especially if the groups are well separated. Overall, the univariate tilted method performed well. It also performed well when the component densities are skewed. The simulations for the gamma mixtures produced good estimates for the component means and standard deviations even when the exponential tilt assumption is not satisfied. This is not the case for parametric methods. If the incorrect parametric model is chosen, it may lead to incorrect results. It is reasonable to say that the tilted method is a flexible alternative to a parametric approach.

There are advantages and disadvantages to using the univariate method described in Chapter 2. When using a nonparametric method, such as described in Bordes et al. (2007) and Benaglia et al. (2009a), a bandwidth needs to be chosen. It is not always easy to chose a bandwidth and different choices may lead to different results. In the tilted method, we have shown that choosing a large bandwidth does not negatively effect

the density estimate nor the estimates of the component means and standard deviations. The disadvantage of using our method is that identifiably is not shown.

A function to calculate the estimates using the exponential tilt model is written using R code and we hope to include it in `mixtools` in the future. Our function is capable of handling univariate mixtures with an arbitrary number of components ($m > 1$) and an arbitrary order in the exponent.

In Chapter 3, we introduced the multivariate method. We presented an EM algorithm to fit the mixture and also developed some tools to help with analyzing the data. The method has the advantage over earlier work in that we do not require identically distributed measures. This is an advantage over the method proposed in Elmore et al. (2004) in both analyzing and interpreting datasets such as the Water level data for example (see Section 3.10.2). Although they found results not unlike the ones found using the tilted method, it is unlikely that the measurements in that dataset are conditionally i.i.d. and their interpretation of the data does not seem valid under that assumption. The interpretation is clearer under the conditional independence case.

Since the tilted method has the advantage of having a likelihood, we developed a likelihood ratio test for testing the presents of blocks of conditionally i.i.d. coordinates. This is an important aspect because it may be hypothesized that certain measurements are conditionally i.i.d. while others are conditionally independent. Analyzing the data using the i.i.d coordinates to estimate the parameters produces better estimates then treating them as independent. Benaglia et al. (2009a) also have the capability of treating data as blocks of conditionally i.i.d. measurements, but they do not have the capability of testing to see if this is appropriate given the data.

We presented a model selection procedure based on BIC, called pBIC, to help determine the number of components if theory does not suggest it. We provided simulations that demonstrated the performance of our method. For the simulations from normal mixtures at different sample sizes and various number of coordinates pBIC satisfactory results selecting the number of components. In some cases, it performed better that the minimum distance method suggested by Benaglia (2008) for the same models.

We also provided simulations for pBIC for gamma mixtures that did not satisfy our exponential tilt assumption and pBIC performed satisfactory at choosing the true number of components.

We developed a tool to check the conditional independence assumption based on the sample correlation and the sample correlation under the conditional independence assumption. The tool will help show if there is strong dependence among the repeated measures. If it does not show dependence, it is advisable to check independence using methods that utilize higher order moments such as those suggested in Benaglia (2008).

In Section 3.8, we presented simulation results for various models. For the normal mixtures, the tilted method produced results similar to those from the normal mixture method and the nonparametric method suggested by Benaglia et al. (2009a). When the component densities were skewed, the tilted method perform much better than the normal mixture method but the nonparametric method performed only slightly better. This was expected since the gamma mixtures shown in that section did not satisfy the exponential tilt assumption. There was one set of simulations, the model with different component densities, where the tilted method performed better than the nonparametric method. There are cases where the measurements may come from very different distributions where some might be symmetric while others may be skewed and the tilted method does appear to be a better choice.

For the simulations, we chose various sample sizes ranging from $n = 50$ to $n = 500$. It appears that the method performs well at larger samples sizes but still performed satisfactory when $n = 100$. When considering if the method will perform well with a particular sample size it is helpful to the think about the number of parameters in the model. For the exponential tilt model, there are $(2k + 1)(m - 1)$ parameters. If theory suggests the number of components in the mixture, a conjecture of how many observations will be used to estimate each of the parameters could be helpful.

Overall, as with the univariate case, the multivariate tilted method is a flexible model. It performs nearly as well, and in some cases better, than the nonparametric method even if the exponential tilt is not satisfied. Being robust is a great advantage over parametric methods. Its greatest advantage is over the nonparametric method

is having a likelihood. This advantage allows for various additional methods for model selection that the nonparametric method is not capable of performing and yet it provides results similar to those from the nonparametric method.

4.2 Future work

In this section, we will discuss some issues that need to be addressed and other future research. First, identifiability needs to be shown for the univariate case. For the nonparametric method to be identifiable, the components must be symmetric and location shifted. This is not the case in this model.

Although the tilted EM algorithm is currently written as a function for R, it is not included in `mixtools`. There are plans to include the function in `mixtools` in the future. Although the function can handle an arbitrary number of components and coordinates, the function sometimes runs slowly. A more efficient version of the function would be beneficial.

In many real life situations, the conditional independence assumption does not seem valid. It is of interest to develop a model that relieves this assumption and will still be identifiable. We are looking at ways to introduce dependence into the model. For example, consider a mixture model where the underlying component densities are conditionally independent but the dependence is introduced into the exponential tilt. Also, it might be of interest where one of the coordinates has a known parametric form but the others are related by an exponential tilt.

Since we are capable to using the likelihood for likelihood ratio tests, it may be of interest to investigate different uses of the test. For example, we could test if particular exponent tilt parameters are equal to zero. Also, we are looking into a test for conditionally independent coordinates versus not conditionally independent coordinates since that assumption is essential for identifiability.

Appendix A

Additional Simulations from Chapter 2

	μ_1	σ_1	μ_2	σ_2	λ
True Values	0	1	3	1	0.30
Bandwidth=.5					
Breaks=30	0.0236	1.1320	2.722	1.2290	0.2836
Breaks=100	0.0919	1.1698	2.8228	1.2018	0.3013
Bandwidth=1					
Breaks=30	-0.0116	1.0102	2.9586	1.0410	0.3227
Breaks=100	0.0088	1.0214	2.9705	1.0328	0.3277
Bandwidth=2					
Breaks=30	-0.0138	0.9903	2.9867	1.0080	0.3289
Breaks=100	-0.0011	0.9978	2.9933	1.0039	0.3317
Bandwidth=3					
Breaks=30	-0.0084	0.9949	2.9883	1.0073	0.3298
Breaks=100	0.0046	1.0027	2.9949	1.0033	0.3327

Table A.1. The semiparametric estimates for the mixing proportion and the component means and standard deviations using different bandwidths and breaks for normal scores.

Parameter	True	sp.density	sp.density*	normal	np
λ	0.3	0.2836(0.0486)	0.3056(0.1016)	0.2867(0.0458)	0.2688(0.0430)
μ_1	-1	-1.0664(0.1939)	-0.973(0.4878)	-1.0958(0.1680)	-0.9225(0.3683)
μ_2	2	1.9307(0.1293)	1.932(0.1268)	1.9579(0.1262)	1.8303(0.1865)
σ_1	1	0.9844(0.1506)	1.0274(0.2794)	0.9340(0.1413)	1.1512(0.1554)
σ_2	1	1.0575(0.0998)	1.0349(0.1274)	1.0190(0.0955)	1.1632(0.1394)

Table A.2. The semiparametric estimates of the component means (standard errors) and standard deviations for Model (2.70).

Parameter	True	sp.density	sp.density*	normal
λ	0.3	0.2733(0.0582)	0.3220(0.0778)	0.3072(0.0479)
μ_1	0	-0.0285(0.1721)	0.0056(0.1619)	0.0125(0.1630)
μ_2	5	4.8074(0.2772)	4.9487(0.2366)	5.0119(0.2090)
σ_1	1	0.9693(0.1549)	1.0045(0.1242)	1.0013(0.1173)
σ_2	2	2.1630(0.2150)	2.0366(0.1860)	1.9838(0.1601)

Table A.3. The semiparametric estimates of the component means (standard errors) and standard deviations for Model (2.71).

Appendix B

Additional Simulations from Chapter 3

Significance Level (α)	Rejection Rate
0.01	0.0200
0.05	0.0600
0.10	0.1000
0.25	0.2700

Table B.1. The rejection rates for 300 simulations of sample size $n = 500$ from Model (3.71) at various significance levels with $\lambda = 0.3$.

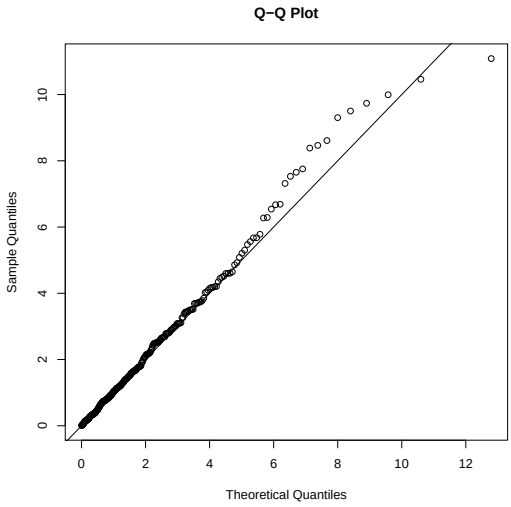


Fig. B.1. The Q-Q plot for the test statistics from the simulations of Model (3.71) with $\lambda = 0.3$. The theoretical distribution is a Chi-square with degrees of freedom equal to 2.

Parameter	True	TiltedEM	Normal	npEM
λ	0.5	0.4987(0.0318)	0.4996(0.0298)	0.4987(0.0307)
μ_{11}	0	-0.0085(0.0760)	-0.0053(0.0739)	-0.0092(0.0742)
μ_{12}	0	-0.0033(0.0775)	-0.0021(0.0730)	-0.0032(0.0724)
μ_{13}	0	0.0084(0.0889)	-0.0004(0.0820)	0.0381(0.0846)
μ_{21}	1	0.9980(0.0737)	0.9967(0.0718)	0.9986(0.0731)
μ_{22}	1.5	1.5060(0.0750)	1.5034(0.0714)	1.5059(0.0721)
μ_{23}	2.5	2.4841(0.0932)	2.4971(0.0831)	2.4545(0.0858)
σ_{11}	1	0.9976(0.0505)	0.9984(0.0496)	0.9973(0.0490)
σ_{12}	1	0.9912(0.0566)	0.9953(0.0523)	0.9922(0.0512)
σ_{13}	1	1.0092(0.0777)	1.0005(0.0570)	1.0460(0.0655)
σ_{21}	1	0.9961(0.0533)	0.9976(0.0522)	0.9963(0.0507)
σ_{22}	1	0.9905(0.0548)	0.9932(0.0513)	0.9911(0.0494)
σ_{23}	1	1.0119(0.0767)	0.9981(0.0576)	1.0481(0.0679)

Table B.2. Two component normal mixture with conditionally independent cpprdinates with $\lambda = 0.5$, $\boldsymbol{\mu}_1 = (0, 0, 0)$, $\boldsymbol{\mu}_2 = (1, 1.5, 2.5)$, and $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2 = (1, 1, 1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$.

Parameter	True	TiltedEM	Normal	npEM
λ	0.8	0.7975(0.0301)	0.8006(0.0264)	0.7984(0.0270)
μ_{11}	0	-0.0094(0.0542)	-0.0058(0.0529)	-0.0097(0.0530)
μ_{12}	0	-0.0033(0.0578)	0.0015(0.0548)	-0.0035(0.0543)
μ_{13}	0	0.0008(0.0617)	-0.0021(0.0582)	0.0163(0.0584)
μ_{21}	1	0.9992(0.1302)	0.9990(0.1241)	1.0044(0.1267)
μ_{22}	1.5	1.5000(0.1470)	1.5026(0.1365)	1.5071(0.1378)
μ_{23}	2.5	2.4508(0.2085)	2.4939(0.1658)	2.3956(0.1781)
σ_{11}	1	0.9986(0.0391)	1.0003(0.0386)	0.9969(0.0389)
σ_{12}	1	0.9931(0.0441)	0.9970(0.0416)	0.9912(0.0420)
σ_{13}	1	0.9995(0.0509)	0.9994(0.0403)	1.0168(0.0447)
σ_{21}	1	0.9884(0.0861)	0.9894(0.0812)	0.9934(0.0825)
σ_{22}	1	0.9864(0.0928)	0.9852(0.0867)	0.9897(0.0892)
σ_{23}	1	1.0370(0.1667)	0.9907(0.1031)	1.1123(0.1489)

Table B.3. Two component normal mixture with conditionally independent cpprdinates with $\lambda = 0.8$, $\boldsymbol{\mu}_1 = (0, 0, 0)$, $\boldsymbol{\mu}_2 = (1, 1.5, 2.5)$, and $\boldsymbol{\sigma}_1^2 = \boldsymbol{\sigma}_2^2 = (1, 1, 1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$..

Parameter	True	TiltedEM	Normal	npEM
λ	0.5	0.5012(0.0239)	0.4997(0.0236)	0.5051(0.0238)
μ_{11}	0	0.0013(0.0687)	0.0000(0.0682)	0.0081(0.0682)
μ_{12}	0	-0.0002(0.0706)	-0.0018(0.0699)	0.0068(0.0709)
μ_{13}	0	0.0118(0.0735)	0.0024(0.0682)	0.0358(0.0726)
μ_{21}	2	2.0047(0.0764)	1.9998(0.0755)	2.0134(0.0753)
μ_{22}	2.5	2.5118(0.0964)	2.5055(0.0936)	2.5242(0.0945)
μ_{23}	3	2.9980(0.0695)	2.9983(0.0689)	2.9969(0.0686)
σ_{11}	1	0.9996(0.0458)	0.9983(0.0451)	1.0052(0.0457)
σ_{12}	1	0.9966(0.0468)	0.9956(0.0458)	1.0047(0.0481)
σ_{13}	1	1.0116(0.0642)	0.9972(0.0509)	1.0424(0.0637)
σ_{21}	1.2247	1.2200(0.0566)	1.2234(0.0553)	1.2155(0.0557)
σ_{22}	1.4142	1.4056(0.0687)	1.4056(0.0682)	1.3937(0.0679)
σ_{23}	1	0.9936(0.0494)	0.9948(0.0481)	0.9991(0.0490)

Table B.4. Two component normal mixture with conditionally independent coordinates with $\lambda = 0.5$, $\boldsymbol{\mu}_1 = (0,0,0)$, $\boldsymbol{\mu}_2 = (2,2.5,3)$, $\boldsymbol{\sigma}_1^2 = (1,1,1)$, and $\sigma_2^2 = (1.5,2,1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$.

Parameter	True	TiltedEM	Normal	npEM
λ	0.8	0.8003(0.0195)	0.7994(0.0194)	0.8027(0.0191)
μ_{11}	0	-0.0042(0.0525)	-0.0048(0.0527)	-0.0016(0.0528)
μ_{12}	0	-0.0014(0.0521)	-0.0018(0.0518)	0.0021(0.0522)
μ_{13}	0	0.0038(0.0521)	0.0002(0.0509)	0.0133(0.0523)
μ_{21}	2	2.0075(0.1335)	2.0004(0.1307)	2.0210(0.1318)
μ_{22}	2.5	2.5070(0.1626)	2.4971(0.1596)	2.5227(0.1607)
μ_{23}	3	2.9863(0.1098)	2.9890(0.1064)	2.9834(0.1107)
σ_{11}	1	0.9954(0.0352)	0.9952(0.0347)	0.9968(0.0351)
σ_{12}	1	0.9981(0.0372)	0.9979(0.0364)	1.0017(0.0376)
σ_{13}	1	1.0038(0.0424)	0.9978(0.0397)	1.0157(0.0442)
σ_{21}	1.2247	1.2035(0.0903)	1.2073(0.0901)	1.1986(0.0918)
σ_{22}	1.4142	1.4045(0.1092)	1.4110(0.1085)	1.3972(0.1085)
σ_{23}	1	1.0000(0.0873)	0.9980(0.0771)	1.0121(0.0922)

Table B.5. Two component normal mixture with conditionally independent cpprdinates with $\lambda = 0.8$, $\boldsymbol{\mu}_1 = (0,0,0)$, $\boldsymbol{\mu}_2 = (2,2.5,3)$, $\boldsymbol{\sigma}_1^2 = (1,1,1)$, and $\sigma_2^2 = (1.5,2,1)$. Displayed are the means (standard errors) of the estimates based on 1000 simulations from model (3.83) with sample size $n = 500$.

Parameter	True	$n = 50$	$n = 300$
λ	0.3	0.332(0.084)	0.304(0.030)
μ_{11}	0	0.917(1.614)	0.038(0.166)
μ_{12}	4	3.238(1.605)	3.943(0.319)
μ_{13}	0	-0.048(1.480)	0.001(0.118)
μ_{21}	4	3.692(0.980)	4.001(0.203)
μ_{22}	0	0.239(0.862)	0.004(0.108)
μ_{23}	0	-0.005(1.044)	-0.001(0.381)
σ_{11}	1.000	1.611(1.070)	1.150(0.411)
σ_{12}	2.828	2.459(0.913)	2.786(0.331)
σ_{13}	1.000	1.853(1.859)	0.995(0.115)
σ_{21}	2.828	2.616(0.627)	2.793(0.233)
σ_{22}	1.414	1.468(0.599)	1.422(0.152)
σ_{23}	5.657	5.103(1.373)	5.648(0.448)

Table B.6. Two component mixture of different distributions with conditionally independent coordinates with $\lambda = 0.3$. Displayed are the means (standard errors) of the estimates based on the second set of 1000 simulations from model (3.85) with sample sizes $n = 50$ and $n = 300$.

Parameter	True	Normal	npEM
λ	0.3	0.345(0.106)	0.361(0.091)
μ_{11}	0	0.273(0.231)	0.261(0.202)
μ_{12}	4	3.458(0.411)	3.792(0.332)
μ_{13}	0	-0.008(0.155)	0.009(0.222)
μ_{21}	4	4.131(0.233)	4.027(0.209)
μ_{22}	0	0.015(0.116)	-0.054(0.102)
μ_{23}	0	0.005(0.403)	-0.005(0.397)
σ_{11}	1.000	1.122(0.185)	1.537(0.465)
σ_{12}	2.828	3.038(0.342)	2.819(0.329)
σ_{13}	1.000	1.076(0.562)	1.792(0.862)
σ_{21}	2.828	2.872(0.239)	2.790(0.232)
σ_{22}	1.414	1.333(0.129)	1.339(0.118)
σ_{23}	5.657	5.798(0.525)	5.611(0.445)

Table B.7. Results from the normal and nonparametric method based on the second set of 1000 simulations of sample size $n = 300$ from model (3.85)

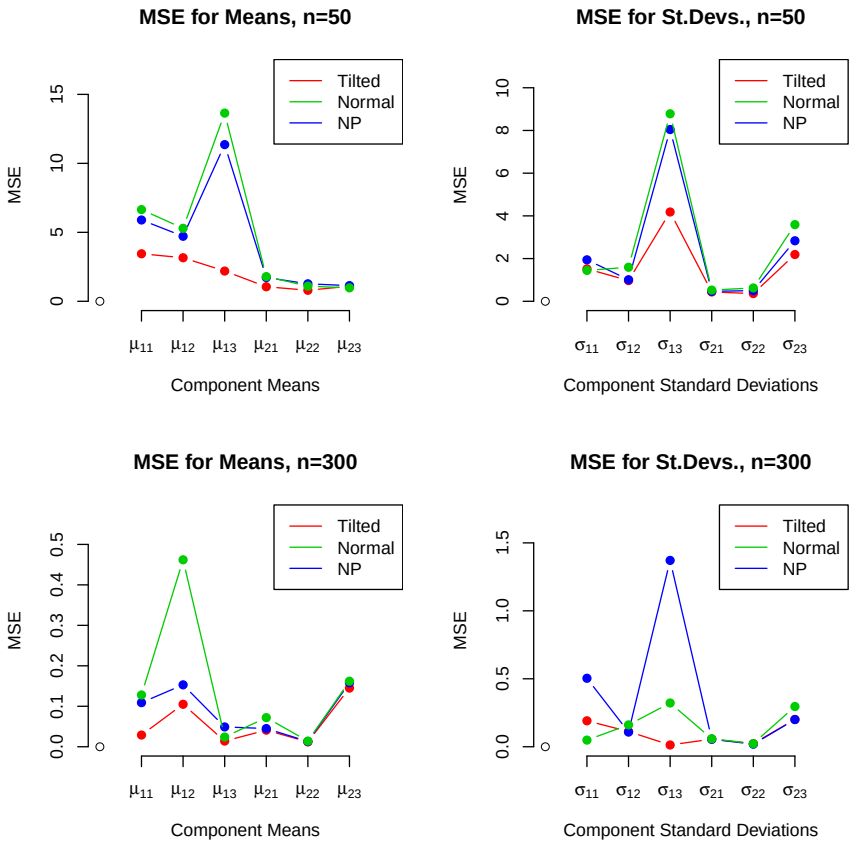


Fig. B.2. Plots of the MSEs for the tiltedEM, npEM, and the normal methods for the second set from the mixture with different distributions.

Bibliography

E. S. Allman, C. Matias, and J. A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37:3099–3132, 2009.

J. A. Anderson. Multivariate logistic compounds. *Biometrika*, 66:17–26, 1979.

A. Azzalini and A. W. Bowman. A look at some data on the old faithful geyset. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 39:357–365, 1990.

T. Benaglia. *Semi- and non-parametric methods in multivariate mixtures*. PhD thesis, Pennsylvania State University, 2008.

T. Benaglia, D. Chauveau, and D. R. Hunter. An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. Technical report, CNRS, 2008. URL `http://hal.archives-ouvertes.fr/hal-00193730/fr/`.

T. Benaglia, D. Chauveau, and D. R. Hunter. An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18:505–526, 2009a.

T. Benaglia, Chauveau D., Hunter D. R., and Young D.S. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32, 2009b.

L. Bordes, D. Chauveau, and P. Vandekerkhove. An EM algorithm for a semiparametric mixture model. 2006a.

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

L. Bordes, D. Chauveau, and P. Vandekerkhove. A stochastic EM algorithm for a semi-parametric mixture model. *Computational Statistics and Data Analysis*, 51(11):5429–5443, 2007.

J Chen and J. D. Kalbfleisch. Penalized minimum-distance estimation in finite mixture models. *The Canadian Journal of Statistics/ La Revue Canadienne de Statistique*, 24: 167–175, 1996.

I. R. Cruz-Medina and T. P. Hettmansperger. Nonparametric estimation in semiparametric univariate mixture models. *J. Stat. Comput. Simul.*, 74(7):513–524, 2004. ISSN 0094-9655.

I. R. Cruz-Medina, T. P. Hettmansperger, and H. Thomas. Semiparametric mixture models and repeated measures: the multinomial cut point model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(3):463–474, 2004. ISSN 0035-9254.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

B. Efron and R. Tibshirani. Using specially designed exponential families for density estimation. *The Annals of Statistics*, 24:2431–2461, 1996.

R. Elmore. *Semiparametric Analysis of Finite Mixture Models with Repeated Measures.* PhD thesis, The Pennsylvania State University, 2003.

R. T. Elmore, T. P. Hettmansperger, and H. Thomas. Estimating component cumulative distribution functions in finite mixture models. *Comm. Statist. Theory Methods*, 33 (9):2075–2086, 2004. ISSN 0361-0926.

R. T Elmore, P. Hall, and A. Neeman. An application of classical invariant theory to identiability in nonparametric mixtures. *Annales de l'Institut Fourier*, 55:1–28, 2005.

C. Field and E. Ronchetti. *Small Sample Asymptotics.* IMS Monograph Series v13, 1990.

P. Hall and X. H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224, 2003.

P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92:667–678, 2005.

T. P. Hettmansperger and H. Thomas. Almost nonparametric inference for repeated measures in mixture models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 62(4):811–825, 2000. ISSN 1369-7412.

D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Ann. Statist.*, 35(1):224–251, 2007. ISSN 0090-5364.

R. Kay and S. Little. Assessing the fit of the logistic model: A case study of children with the haemolytic uraemic syndrome. *Applied Statistics*, 35:16–30, 1986.

D. Leung and J. Qin. Semi-parametric inference in a bivariate (multivariate) mixture model. *Statistica Sinica*, 16:153–163, 2006.

B. G. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Ims, 1995.

J. K. Lindsey. Construction and comparison of statistical models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 36:418–425, 1974.

G. McLachlan and D Peel. *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York, 2000. ISBN 0-471-00626-2.

C. A. Miller, R. Kail, B. L. Laurence, and J. B. Tomblin. Speed of processing in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, 44:416–433, April 2001.

S. R. Murphy and A. W. van der Vaart. On profile likelihood. *Journal of American Statistical Association*, 95:449–465, 2000.

A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–49, 1988.

J. Qin. Empirical likelihood ratio based confidence intervals for mixture proportions. *Annals of Statistics*, 27:1368–1384, 1999.

J. Qin, M. Berwick, R. Ashbolt, and T. Dwyer. Quantifying the change of melanoma incidence by breslow thickness. *Biometrics*, 58:665–670, 2002.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

G. Schwartz. Estimating the dimension of a model. *The Annals of Statistics*, 5:461–464, 1978.

B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1.

H. Thomas and A. Lohaus. Modeling growth and individual differences in spatial tasks. In *Monograph of the Society of Research on Child Development*, number 237. 1993.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1985. ISBN 0-471-90763-4.

D. S. Young, T. Benaglia, D. Chauveau, R. T. Elmore, T. P. Hettmansperger, D. R. Hunter, H. Thomas, and F. Xuan. mixtools: Tools for mixture models. R package version 0.3.2, 2008.

Vita

TRACEY ANN-WROBEL HAMMEL

EDUCATION

- Ph.D. Statistics, Pennsylvania State University. December 2010
- M.S. Statistics, Western Michigan University. May 2004
- B.S. Statistics, Western Michigan University. May 2002

PROFESSIONAL EXPERIENCE

- 2006 - 2008: Research Assistant, Pennsylvania State University
- 2005: Consultant, Consulting Center, Department of Statistics, Pennsylvania State University
- 2004 - 2008: Teaching Assistant, Pennsylvania State University
- 2002 - 2004: Teaching Assistant, Western Michigan University

AWARDS

- 2002: Colonel Charles R. Bayliss Scholarship
- 2002: James H. Powell Award in Statistics

Publication

- Fienberg, S.E, Fulp, W.J., Slavkovic, A.B. and Wrobel, T. (2006). Se- cure Log-Linear and Logistic Regression Analysis of Distributed Databases. Privacy in Statistical Databases - PSD 2006. Springer Lecture Notes in Computer Science No.4302. pages 277-290, Berlin, 2006. Springer-Verlag.