Semi-parametric Mixture Models with Ranked Set Samples

by

© Seyed Jalaleddin Moniri

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics

Memorial University of Newfoundland

August 2021

St. John's, Newfoundland and Labrador, Canada

Abstract

Simple random sampling (SRS) is the common method in data collection. In many applications, measuring the variable of interest is costly, but ranking the units can be done easily. In these situations, one can use rank set sampling (RSS) to get more representative samples from the population. This thesis investigates the estimation of the semi-parametric finite mixture models (FMMs) with RSS. We develop a semi-parametric version of the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood (ML) estimate of the population with RSS data. We then propose the ML estimation of FMM with RSS data in a semi-parametric framework. Our numerical studies show that the proposed EM algorithm estimates more efficiently the FMM. The proposed methods are finally applied to analyze the bone mineral data.

Keywords: Finite mixture model, Ranked set sampling, Semi-parametric estimation, Misplacement probability model, EM algorithm, Bone mineral data.

This work is dedicated to my family.

Lay Summary

Simple random sampling (SRS) is the most common sampling design in data analysis. In many surveys, such as medical research, measuring the variable of interest is difficult. This difficulty may include the situations the measurement procedure is costly and/or time-consuming and/or invasive. In Osteoporosis research, for example, the bone disorder status of patients must be determined by bone mineral density (BMD). Although BMD is the most reliable predictor of bone disorder status, BMD measurements are obtained through dual X-ray absorptiometry (DXA) images. Measuring BMD requires a costly and time-consuming procedure, including DXA imaging and manual segmentation of images by medical experts. Ranked set sampling (RSS) is cost-effective sampling technique that can be applied in situations where the precise measurement of the variable of interest is expensive or hard to achieve; however, sampling units can be ranked via extra variables or judgment ranking, without actual measurements on the variable of interest. For example, one can use RSS to get more representative data of the BMD in the underlying population of interest. To do this, we need to create artificial strata based on ranks during the sampling process. We can rank a small number of patients in comparison sets using expert information or get data on individuals based on their medical history or measurements of reasonable auxiliary variables related to BMD such as age, weight, body mass index, etc. In the standard estimation methods for FMMs, the samples are typically extracted from the population using SRS. In this thesis, we used ranked set sampling to collect more informative samples from the FMMs and developed more efficient semi-parametric estimations for the FMMs.

Acknowledgments

I would like to thank my supervisor, Dr. Armin Hatefi, for all his constant encouragement, guidance and friendship as well as unconditional support throughout these years at Memorial University of Newfoundland. I gratefully acknowledge the financial support in the form of graduate fellowships and teaching assistantships provided by Memorial University of Newfoundlands School of Graduate Studies, the Department of Mathematics and Statistics, and my supervisor.

Candidate's Contribution to the Work

My supervisor proposed the research question. The candidate was responsible for the literature review and data analysis. The candidate was also responsible for the first version of the thesis, the presentation, and the results from this research. In addition, Dr. Armin Hatefi offered advice on the research methodology and computational programming.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, we plan to use a Ranked Set Sampling (RSS) design to analyze finite mixture models (FMMs). In many surveys, ranked-based samplings, as cost-efficient sampling techniques, are more desirable than most commonly used simple random sampling (SRS). Rank-based sampling designs include rank set sampling, judgement post-stratified (JPS) sampling and their variations. On the other side, finite mixture models are convenient and flexible statistical tools that have been used in various scientific disciplines such as medical research, biology and genetics. In this thesis, we study the problem of semi-parametric finite mixture modelling with ranked set sampling.

This chapter is organized as follows. In Section 1.1, we first provide an overview of the finite mixture models. Section 1.2 gives an introduction to ranked set sampling design. Section 1.3 describes semi-parametric finite mixture modelling. Finally, Section 1.4 presents the outline of the thesis.

1.1 Finite Mixture Models

Finite mixture models are powerful statistical tools in situations that observations arise from heterogeneous subpopulations. Finite mixture models, as powerful tools to analyze complex data, have increasingly being used during the last decade. Mixture models help explain a wide variety of random phenomena because of their flexibility. Therefore in many fields of science, they are used to model complex processes and systems. Examples of applications include clustering, density estimation and classification.

Let X be a random variable representing the population of the study. Let X follow a finite mixture model with M subpopulations. The probability density function (pdf) of random variable X is given by

$$g(x, \mathbf{\Psi}) = \sum_{j=1}^{M} \pi_j f_j(x, \theta_j), \tag{1.1}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{M-1})$ represents the vector of unknown mixing parameters with $\sum_{j=1}^{M} \pi_j = 1$ and f_j, $j = 1, \ldots, M$ represents the pdf of the jth component of the model. Also $\boldsymbol{\xi}^\top = (\theta_1^\top, \theta_2^\top, \ldots, \theta_M^\top)$ represents the vector of unknown component parameters. We use $\mathbf{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi})$ to show the vector of all unknown parameters of model (1.1). It is usually assumed that the pdf's f_j belong to a common parametric family $\mathcal{F} = \{f(\cdot; \theta), \theta \in \mathbb{R}^d\}$ characterized by a vector of parameters θ so that the pdf of the finite mixture model is given by

$$g(x, \mathbf{\Psi}) = \sum_{j=1}^{M} \pi_j f(x; \theta_j), \tag{1.2}$$

where $\boldsymbol{\Psi} = \{(\pi_j, \theta_j); j = 1, \ldots, M\}$.

Because of advances in simulation and computational methods, finite mixture models have become flexible and valuable statistical tools in data analysis. Finite mixture models have been broadly employed in statistical analysis, for example, modelling unknown distributional forms, analyzing the data including group structures, model-based classification and clustering analysis.

In addition, finite mixture models have been used in various fields, including astronomy, biology, genetics, medicine, psychiatry, economics, engineering, and marketing, among many other biological, physical, and social sciences. For example, Sodium and Lithium Counter-transport (SLC) activity in red blood cells is an essential characteristic in quantitive genetics; because it associates with blood pressure. Furthermore, SLC action is more straightforward to examine than blood pressure. Assume the action of a particular gene defines SLC characteristic with alleles A and a. Chen et al. (2012) then investigated the presence of a significant gene by using FMMs for analysis of the SLC groups. Moreover, FMMs have been applied in genetics (e.g., Schork et al., 1996; Roeder, 1994 and Chen and Chen, 2003), medical studies (e.g., Schlattmann, 2009) and different engineering fields, such as in speech recognition, medical imaging, pattern recognition (e.g., El Zaart et al., 2002).

FMMs are reliable means for modelling various random events and cluster data sets. FMMs present practical principles for understanding data with complicated structures. Because of these flexibilities, FMMs have attracted many researchers over time, both in theory and applications (McLachlan and Krishnan, 2007). Pearson (1894) and Cohen (1967) use the method of moments, for estimation of finite mixture models. Harding (1949) and Cassie (1954) used graphical methods to es-

timate the finite mixture models. Among all methods, maximum likelihood (ML) estimation is considered to be the most common method to estimate the parameters of the mixture models (Furman and Lindsay, 1994). The likelihood equations for the FMM are usually nonlinear. So, it is challenging to solve the problems analytically. One has to obtain the estimates through iterative methods. We here use EM-algorithm (Dempster et al., 1977) to obtain ML estimates of the parameters of the FMMs.

The EM-algorithm has been applied to a mixture problem in genetics by Tan and Chang (1972) and by Hosmer Jr (1973) in the Monte Carlo study of ML estimation. Duda et al. (1973) studied the EM algorithm for mixtures of multivariate normal densities and explained its performance. Peters and Walker (1978) proposed a convergence investigation of the EM algorithm for mixtures of multivariate normal densities and introduced changes of the algorithm to stimulate convergence. Lange (1995) considered an acceleration of the EM algorithm based on classical quasi-Newton optimization methods. This acceleration seeks to drive the EM algorithm constantly to the Newton-Raphson algorithm, which has a quadratic convergence rate. The significant distinction between the current algorithm and a naive quasi-Newton algorithm is that the early steps of the current algorithm match the EM algorithm rather than the steepest ascent.

1.2 Ranked Set Sampling

Simple random sampling (SRS) is the most common sampling design in data analysis. In many surveys, such as medical research, measuring the variable of interest

is difficult. This difficulty may include the situations the measurement procedure is costly and/or time-consuming and/or invasive. In fishery surveys, for example, the age characteristic of fish is the most important variable in stock assessment and fishery management. While age characteristic is critical, the age determination of fish is a costly and destructive procedure. To measure the exact age, the fish should be dissected. Then the exact age of fish will be obtained from otoliths through a substantial time-consuming process (Hatefi et al., 2015). In Osteoporosis research, for example, the bone disorder status of patients must be determined by bone mineral density (BMD). Although BMD is the most reliable predictor of bone disorder status, BMD measurements are obtained through Dual X-ray Absorptiometry (DXA) images. Measuring BMD requires a costly and time-consuming procedure, including DXA imaging and manual segmentation of images by medical experts (Omidvar et al., 2018).

In these surveys, while measuring the variable of interest is costly, practitioners typically have access to easy-to-measure characteristics about individuals. Ranked set sampling employs these easy-to-measure characteristics for ranking the sampling units and then incorporates this ranking information efficiently into both the data collection. Ranked set sampling (RSS), as informative and cost-effective sampling scheme, is more desirable than simple random sampling in these situations (Hatefi and Jozani, 2013).

We construct an RSS of size nH when H is set size and n is cycle size as follows. First, we take a SRS of size H, $X_1, \ldots, X_H$, from the population. We rank the sample as $O_r(X_1, \ldots, X_H) = (X_{[1]}, \ldots, X_{[H]})$ from the smallest to the largest by using a ranking operator, $O_r(\cdot)$. Note that we rank the units in each set without

measuring their interest variable (i.e. X-variable). We rank the units based on an easy-to-measure concomitant variable (correlated with X-variable). We choose the unit with the smallest rank and measure only the X-variable of this unit, denoted by $X_{[1]1}$. Then we take another SRS of size H, independent of the first set, from the population. Ranking the set, we select the unit with the second smallest rank for full measurement, denoted by $X_{[2]1}$. We continue this process until we measure the item with the largest rank for last set, denoted by $X_{[H]1}$. We consider this whole process as a cycle. To obtain the total number of nH observations from the population, we repeat this process for n cycles.

We consider $X_{[r]i}$ as the value of the r-th ordered unit in the i-th cycle. Also the $X_{[r]i}$ represents the r-th judgment order statistic in the i-th cycle. The balanced RSS is given by $\{X_{[r]i}, r = 1, \ldots, H; i = 1, \ldots, n\}$. The RSS is called balanced when we obtain the same number of observations from each rank strata; otherwise, the RSS is called unbalanced. To show the construction of a balanced RSS, Table 1.1 represents an illustrative example when set size $H = 4$ and cycle size $n = 2$. In our example, we denote $X_{[r]i}, r = \{1, 2, 3, 4\}$ and $i = \{1, 2\}$ as the measured balanced RSS observations.

As described above, the RSS sampling is similar to the stratified sampling. RSS creates artificial strata by ranking the sampling units. In other words, RSS can be considered as a stratification of the sampling units based on their ranks in the sets. Although we need to identify nH^2 units from the population, we measure the variable of interest for only nH units. Note that the RSS statistics are independent because we select all the units for the measurement in our sample from independent sets. However, RSS data, unlike SRS data, are not identically distributed.

Table 1.1: An example of balanced RSS sample

cycle	set	ranking the units within the sets	Observation
1	1	$\{X_1, X_2, X_3, X_4\} \rightarrow \{\boldsymbol{X_{[1]}}, X_{[2]}, X_{[3]}, X_{[4]}\}$	$X_{[1]1}$
	2	$\{X_5, X_6, X_7, X_8\} \rightarrow \{X_{[1]}, \boldsymbol{X_{[2]}}, X_{[3]}, X_{[4]}\}$	$X_{[2]1}$
	3	$\{X_9, X_{10}, X_{11}, X_{12}\} \rightarrow \{X_{[1]}, X_{[2]}, \boldsymbol{X_{[3]}}, X_{[4]}\}$	$X_{[3]1}$
	4	$\{X_{13}, X_{14}, X_{15}, X_{16}\} \rightarrow \{X_{[1]}, X_{[2]}, X_{[3]}, \boldsymbol{X_{[4]}}\}$	$X_{[4]1}$
2	1	$\{X_{17}, X_{18}, X_{19}, X_{20}\} \rightarrow \{\boldsymbol{X_{[1]}}, X_{[2]}, X_{[3]}, X_{[4]}\}$	$X_{[1]2}$
	2	$\{X_{21}, X_{22}, X_{23}, X_{24}\} \rightarrow \{X_{[1]}, \boldsymbol{X_{[2]}}, X_{[3]}, X_{[4]}\}$	$X_{[2]2}$
	3	$\{X_{25}, X_{26}, X_{27}, X_{28}\} \rightarrow \{X_{[1]}, X_{[2]}, \boldsymbol{X_{[3]}}, X_{[4]}\}$	$X_{[3]2}$
	4	$\{X_{29}, X_{30}, X_{31}, X_{32}\} \rightarrow \{X_{[1]}, X_{[2]}, X_{[3]}, \boldsymbol{X_{[4]}}\}$	$X_{[4]2}$

The RSS was first introduced by McIntyre (1952) to obtain the estimation of the mean pasture yields. RSS has had applications in a wide range of fields such as medical research, industrial statistics, environmental and ecological research. For instance, analyzing the environmental dangers of hazardous waste places, including poisonous chemicals and their ecological influence, needs significant scientific processing of materials and, consequently, high expenses. Despite that, one can rank hazardous waste sites according to their pollution levels from visual examination of soil or water discolouration (Barabesi and El-Sharaawi, 2001). RSS is extensively used in medical research. For example, the biomarkers perform critical functions in evaluating lung cancer status. It would be expensive and time-consuming to obtain the results from all patients from the biomarkers by doing lab experiments. However, it is possible to rank the patients based on their smoking exposure levels. Taking advantage of the connection between smoking exposure and biomarkers, Chen and Wang (2004) employed RSS for lung cancer research.

In addition, there are more applications of RSS in natural sciences. For example, Halls and Dell (1966) show that RSS is more efficient than SRS to estimate browse

and herbage weights in a pine-hardwood forest. Muttlak and McDonald (1992) used RSS estimates to obtain mean values of forest and grassland resources with higher performance than other standard sampling techniques. Wang et al. (2009) describe how the RSS can improve sample collection efficiency and decrease the cost of data collection through two actual samples. The first case is a study of fish stocks in Australia, and another one is a fish age measurement research in Bangladesh.

Ranking plays an essential role in the efficiency of the RSS method. If there is no error in the ranking of sampling units in each set, then the method is called perfect RSS. In this case, the maximum ranking information is incorporated into data collection and estimations. Consequently, the RSS-based inference acquires the highest efficiency relative to the SRS-based inference (Hatefi et al., 2014). Although perfect RSS data result in the highest efficiency, the ranking error is undeniable in the real-life applications where we obtain ranking information using external concomitant variables. The method presumes ranking error is called imperfect RSS. In imperfect RSS, the rank assigned to RSS statistics (i.e. obtained from different sets) may differ from their true rank. When ranking error increases in RSS data collection, the efficiency of RSS-based estimators decreases. Note that when ranks are assigned randomly (i.e. the worst ranking scenario), the imperfect RSS result in simple random sampling (Dell and Clutter, 1972). Therefore, when the the rankings within each set are more accurate, we will obtain the more efficiency (Chen, 2000, Barabesi and El-Sharaawi, 2001). In this thesis, we use the square brackets to show the possibility of ranking errors in RSS data collection; then imperfect RSS data is indicated by $\{X_{[r]i}, r = 1, \ldots, H; i = 1, \ldots, n\}$.

RSS is introduced in different aspects of nonparametric inference. Bohn (1996)

studied the nonparametric methods for data from RSS and presented the similarities and dissimilarities in the characteristics of the RSS methods. Presnell and Bohn (1999) obtained the asymptotic distribution for random sample U-statistics using RSS data. The results show that the RSS method is asymptotically as efficient as the SRS scheme. Öztürk (1999) developed a new two-sample testing method to examine the equality of the two populations using RSS data. Barabesi (2001) suggested the sign test under unbalanced RSS (URSS) that develops a generalization of the standard RSS. Sinha et al. (1996) investigated the theory of RSS when the population is partially known. They discuss the estimation challenges of a normal mean and a normal variance, and an exponential mean. For all three issues, RSS results in significantly improved estimators compared to an SRS. For an overview of the theory and applications of ranked set sampling designs, see (Chen et al., 2003).

In the usual ways of modeling and inference for FMMs, we typically consider that the samples are drawn from the population using the SRS (e.g., McLachlan and Peel, 2004). However, in various applications, more informative, more economical samples are desirable for analyzing FMMs. Hatefi et al. (2014) study ML estimation of the parameters of a FMM for RSS data. They suggest two RSS designs from a FMM and describes how to estimate the unknown parameters of the model. The results show that estimators based on the RSS are more efficient than the SRS. Hatefi et al. (2015) introduce a new method to estimate the parameters of a FMM based on partially rank-ordered set (PROS) sampling. They also suggest a proper EM algorithm to estimate the parameters of the FMMs based on PROS samples. The results show that the ML estimators based on PROS samples work much better than their SRS equivalents, also with small samples. Omidvar et al. (2018) investigated ML

estimation of unknown parameters of a FMM from judgement post-stratified (JPS). The results showed that JPS estimators perform better than their SRS equivalents.

1.3 Semi-parametric Mixture Models

We can use FMMs as an essential tool to analyze complex data in many scientific fields such as Statistics, Economics, Epidemiology and Finance. Parametric FMMs are commonly used because the models can be explained easily and have substantial theoretical aspects. But, parametric FMMs are based on assumptions, like linearity and normality, which are violated in real-life applications. Therefore, semi-parametric FMMs are motivated to ease the assumptions of the parametric family for component densities of the FMMs.

Bordes et al. (2006) investigated a two-component mixture of locations model where the component density is symmetric. Their results show that the estimators are well consistent, following the mild regularity assumptions. Bordes et al. (2007a) generalize the EM algorithm to semi-parametric FMMs based on a two-component mixture of locations model where the component density is considered symmetric. Chang and Walther (2007) developed the EM algorithm to work with the flexible, nonparametric class of log-concave component distributions. Bordes et al. (2006) suggested a semi-parametric two-component FMM that one component was known. They examined data to identify statistically differentially expressed genes in bovine trophoblast between artificial insemination (AI) and *in vitro* fertilization (IVF) gestation modes. This statistical analysis helps the biologist understand the biological differences between the two gestation modes and improve IVF procedures to de-

crease the mortality rate related to this gestation mode.

Some semi-parametric FMMs have been developed theoretically and practically and shown to have a better performance. Hunter et al. (2007) found the identifiability of the location-shifted FMM for 2 and 3 components. Bordes et al. (2007b) employed a stochastic EM algorithm to estimate the unknown parameters of the location-shifted FMM. Benaglia et al. (2009) introduced a more flexible and proper algorithm to reduce the stochasticity of Bordes et al. (2007b) and it is possible to extend that model to various numbers of mixture components.

1.4 An Overview of Thesis

In this thesis, we focus on the mixture of the location-scale models. We investigate how we can use properties of ranked set sampling to develop more efficient estimations of semi-parametric mixture models of Bordes et al. (2007b). In Chapter 2, we present the likelihood function of the mixture model based on simple random samples. We use a semi-parametric version of the EM algorithm to obtain the ML estimate of the parametric and non-parametric elements of the underlying mixture model. Chapter 3 investigates the ML estimation of the FMMs based on the RSS data. We develop an estimate of FMMs based on RSS data in a semi-parametric framework. To do so, we also develop an EM algorithm to obtain the RSS-based ML estimate of semi-parametric FMMs. In Chapter 4, we conduct simulation studies to investigate the performance of the RSS estimators. Then we apply the methods discussed in Chapters 2 and 3 to a real data example.

Chapter 2

Mixture Models from Simple Random Samples

In this chapter, we focus on semi-parametric finite mixture models (FMMs) from simple random samples (SRS). Here, we present the likelihood function of the mixture model based on SRS data. We use a semi-parametric version of the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood (ML) estimate of the parametric and non-parametric elements of the underlying FMM.

This chapter is organized as follows. In Section 2.1, we introduce the semi-parametric FMMs. Section 2.2 describes how one can use the missing-data mechanism to accommodate latent variables and EM algorithm to obtain the ML estimate of the FMM in a parametric setting. Finally, we discuss the ML estimation of semi-parametric FMMs using simple random sampling in Section 2.3.

2.1 Introduction

In many real-life applications, the population of interest comprises of several subpopulations. In these cases, let X be a random variable representing the population of the study. Let X follow a FMM consisting of M subpopulations. Hence, the probability density function (pdf) of random variable X is given by

$$g(x, \boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j f_j(x, \theta_j), \qquad x \in R, \tag{2.1}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_{M-1})$ represents the vector of unknown mixing parameters with $\sum_{j=1}^{M} \pi_j = 1$ and f_j, $j = 1, \ldots, M$ represents the pdf of the jth component of model. Also $\boldsymbol{\xi}^\top = (\theta_1^\top, \theta_2^\top, \ldots, \theta_M^\top)$ represent the vector of unknown component parameters. We use $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi})$ to show the vector of all unknown parameters of the model (2.1). It is usually assumed that the pdf's f_j belong to a common parametric family $\mathcal{F} = \{f(\cdot; \theta), \theta \in \mathbb{R}^d\}$ indexed by a vector of parameters θ so that the pdf of the FMM is given by

$$g(x, \boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j f(x; \theta_j), \tag{2.2}$$

where $\boldsymbol{\Psi} = \{(\pi_j, \theta_j); j = 1, \ldots, M\}$. The choice of a parametric family $\mathcal{F}$ may be difficult to find for a FMM. It is also important to note that the model (2.2) will be more flexible if M is considered to be an unknown parameter, thus one needs to estimate M. Various research studies in the literature have investigated the estimation of the number of components of a mixture model. Leroux (1992) developed a consistent estimator for the mixing parameters' distribution and proposed an estimation

procedure for the number of components of the mixture model. Lemdani et al. (1999) proposed the asymptotic distribution of the likelihood ratio test statistics to examine the number of components of a FMM. Dacunha-Castelle et al. (1999) used the stationary autoregressive moving average (ARMA) time series properties and developed a test statistic for the number of components of mixture models. However, throughout this thesis, we assume that the number of components of the underlying mixture model is known.

Another challenge with which practitioners typically deal is that we know the number of components of the FMM; however, little information is available about the distribution of the subpopulations. Non-identifiablity is one of the biggest challenges associated with the estimation of FMMs.

Definition 2.1. *A distribution* $g(x; \mathbf{\Psi})$ *is identifiable if different values of the parameter* $\mathbf{\Psi}$ *determine differnt members of the family of densities* $g(x; \mathbf{\Psi})$*; that is,* $g(x; \mathbf{\Psi}) = g(x; \mathbf{\Psi}^*)$*, if and only if* $\mathbf{\Psi} = \mathbf{\Psi}^*$*.*

An example of non-identifiability in FMMs can happen when the component densities of the model can themselves be written as a mixture of distributions. In this case, the underlying mixture model is easily non-identifiable. To deal with non-identifiability issue, we need to make additional assumption about the underlying mixture models to make inference. As one possible solution to cope with the issue, Hall (1981) and Titterington (1983) proposed to use the properties of training data and then developed non-parametric estimations of component densities of the underlying mixture model.

Hettmansperger and Thomas (2000) proposed a method to estimate the mixing

proportions and the number of components of the mixture distribution when there are no parametric assumptions about the component distributions. They modeled the mixing distributions with mixture of binomials and studied the efficiency and robustness of this method in the estimation of the multivariate normal mixtures. Cruz-Medina and Hettmansperger (2004) developed non-parametric estimates of the mixing proportions, locations and variances of components of the FMM by assuming that the components are symmetric and without making parametric model assumptions on the components. Hall et al. (2003) discussed the identifibility of p-variate mixture model with two components. Under mild regularity conditions, they show that the FMMs with two components are identifiable when $p \geq 3$.

To make the FMMs identifiable for the case $p \leq 2$, one way is to focus on the restricted parameter space. Hunter et al. (2007) proposed a new method for identifiability of FMMs named M-identifiability, where M represents the number of components of the FMM. To overview the theory and possible solutions to the non-identifiability issue in finite mixture model, see McLachlan and Peel (2004).

To deal with the non-identifiability problem in $p = 1$ in this thesis, we restrict the underlying model to the mixture of the location-scale symmetric models (Bordes et al., 2007a). Hence the semi-parametric version of the FMM (2.2) is given by

$$g(x; \boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j f(x - \mu_j), \quad x \in R, \tag{2.3}$$

where the unknown parameters of the model are given by $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi})$ and $(\boldsymbol{\Psi}, f) \in (\Theta, \mathcal{G})$ with

$$\mathcal{G} = \left\{ \text{Even pdf on } R \right\}$$

and $\Theta = \{(\pi_j, \mu_j), j = 1, \ldots, M\}$ and $\mu_j \neq \mu_l$ for $1 \leq j < l \leq M$.

2.2 Latent Variable Setup and EM Algorithm

In this section, we study the maximum likelihood (ML) estimation of a FMM in a parametric framework. When we want to obtain the ML estimation in a parametric setting, it may not be possible to estimate the parameters in closed form in mixture models. To solve this problem, there are various optimization methods developed in the literature, such as Newton-Raphson (Lindsay, 1995). Among all the methods, the Expectation-Maximization (EM) algorithm of Dempster et al. (1977) is considered as the established method to obtain the ML estimates of the FMMs.

This section presents how one can use the EM algorithm to derive the ML estimates of the parameters of the FMM (2.3). Suppose $X = (x_1, \ldots, x_n)$ be a random sample of size n having the distribution of the M-component mixture model as follows

$$g(x, \mathbf{\Psi}) = \sum_{j=1}^{M} \pi_j f(x; \theta_j) \tag{2.4}$$

From (2.4), the likelihood function of $\mathbf{\Psi}$ based on $\mathbf{x}$ data is given by

$$L_{\mathbf{x}}(\mathbf{\Psi}) = \prod_{i=1}^{n} g(x_i; \mathbf{\Psi}) = \prod_{i=1}^{n} \left\{ \sum_{j=1}^{M} \pi_j f(x; \theta_j) \right\}, \tag{2.5}$$

From (2.5), the log-likelihood function is given by

$$\begin{aligned} l_{\mathbf{x}}(\mathbf{\Psi}) &= \log L_{\mathbf{x}}(\mathbf{\Psi}) \\ &= \sum_{i=1}^{n} \log g(x_i; \mathbf{\Psi}) \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{M} \pi_j f(x_i; \theta_j) \right\} \end{aligned} \tag{2.6}$$

One can obtain the ML estimate of $\mathbf{\Psi}$, through maximizing the log-likelihood function (2.6) as follows

$$\hat{\mathbf{\Psi}}_{ML} = \underset{\mathbf{\Psi}}{\operatorname{argmax}} \;\; l_{\mathbf{x}}(\mathbf{\Psi}). \tag{2.7}$$

Finding the $\hat{\mathbf{\Psi}}_{ML}$ as the solution to the equation (2.7) is not feasible. The log-likelihood function (2.6) is practically untractable with respect to component parameters of the mixture model.

Therefore, we need to introduce latent variables to make the likelihood function more tractable to find the ML estimates. To do so, we need to view $\mathbf{X} = (X_1, \ldots, X_n)$ as incomplete data. Accordingly, from now on, the likelihood function (2.5) and the log-likelihood function (2.6) are called incomplete data likelihood function and incomplete data log-likelihood function, respectively. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the complete data such that $Y_i = (X_i, \boldsymbol{Z_i})$, where $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iM})$ represents the latent variable. We introduce the latent variable $\boldsymbol{Z_i}$ for each x_i, $i = 1, \ldots, n$ as follows

$$Z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component } j; \\ 0 & \text{otherwise.} \end{cases}$$

One can easily see that the $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iM}) \sim Multi(1, \pi_1, \ldots, \pi_M)$ such that the pdf of latent variable $\boldsymbol{Z_i}$ is given by

$$f(\boldsymbol{z_i}) = \prod_{j=1}^{M} (\pi_j)^{z_{ij}} \tag{2.8}$$

From (2.8), the joint pdf of $Y_i = (X_i, \boldsymbol{Z_i})$, $i = 1, \ldots, n$ is given by

$$\begin{aligned} f_Y(y_i; \boldsymbol{\Psi}) &= f(x_i, \boldsymbol{z_i}; \boldsymbol{\Psi}) \\ &= f(x_i | \boldsymbol{z_i}; \boldsymbol{\Psi}) f(\boldsymbol{z_i}) \\ &= \prod_{j=1}^{M} \Big\{ \pi_j f(x_i; \theta_j) \Big\}^{z_{ij}}. \end{aligned} \tag{2.9}$$

Using (2.9), the complete data likelihood function is given by

$$L_{\mathbf{y}}(\boldsymbol{\Psi}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{M} \Big\{ \pi_j f(x_i; \theta_j) \Big\}^{z_{ij}}. \tag{2.10}$$

Therefore, the complete data log-likelihood function becomes

$$\begin{aligned} l_{\mathbf{y}}(\boldsymbol{\Psi}) &= \log L_{\mathbf{y}}(\boldsymbol{\Psi}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log f(x_i; \theta_j). \end{aligned} \tag{2.11}$$

2.2.1 The EM Algorithm: A Parametric Framework

In this subsection, we present the EM algorithm method to obtain ML estimate of FMM in a parametric framework. In this parametric framework; the parametric family of the component density f is assumed to be known; however, the component

density is characterized by unknown component parameter θ. Thus, $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi}^\top)$ represents the vector of all unknown parameters of the underlying complete likelihood function (2.10) where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ and $\boldsymbol{\xi}^\top = (\theta_1^\top, \ldots, \theta_M^\top)$. Now, we propose an EM algorithm to find the ML estimates of $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi}^\top)$. EM algorithm enables us to optimize the complete data log-likelihood function (2.11) through iterating between Expectation step (E-step) and Maximization step (M-step).

Note that we have not observed the latent variable $\boldsymbol{Z_i}$, $i = 1, \ldots, n$. Hence, we need to treat them as latent variable and impute them throughout the EM algorithm. We impute the latent variable $\boldsymbol{Z_i}$, $i = 1, \ldots, n$ with the conditional expectation of $\boldsymbol{Z_i}$ given incomplete data. To do so, we first require to obtain the conditional distribution $\boldsymbol{Z_i}|X_i$ and conditional expectation $\mathbb{E}(\boldsymbol{Z_i}|X_i = xi)$, for $i = 1, \ldots, n$. We know that $\boldsymbol{Z_i}|X_i = x_i$ are independent and identically distributed. From (2.8) and (2.9); we can obtain the conditional pdf $f(\boldsymbol{z_i}|X_i = x_i)$ as follows

$$\begin{aligned} f(\boldsymbol{z_i}|x_i; \boldsymbol{\Psi}) &= \frac{f(x_i, \boldsymbol{z_i}; \boldsymbol{\Psi})}{g(x_i; \boldsymbol{\Psi})} \\ &= \frac{\prod_{j=1}^{M} \left\{ \pi_j f(x_i; \theta_j) \right\}^{z_{ij}}}{\sum_{j=1}^{M} \pi_j f(x_i; \theta_j)} \\ &= \prod_{j=1}^{M} \left\{ \frac{\pi_j f(x_i; \theta_j)}{\sum_{j=1}^{M} \pi_j f(x_i; \theta_j)} \right\}^{z_{ij}} \end{aligned} \tag{2.12}$$

From (2.12), it is easy to see that $(\boldsymbol{Z_i}|X_i = x_i) \overset{iid}{\sim} Mult\left(1, \frac{\pi_j f(x_i;\theta_j)}{\sum_{j=1}^{M} \pi_j f(x_i;\theta_j)}\right)$. Also, we can easily see that

$$\boldsymbol{Z_i}|X_i = x_i \overset{iid}{\sim} Mult\left(1, \frac{\pi_1 f(x_i; \theta_1)}{g(x; \boldsymbol{\Psi})}, \ldots, \frac{\pi_M f(x_i; \theta_M)}{g(x; \boldsymbol{\Psi})}\right)$$

Also, the conditional expectation $\mathbb{E}(Z_{ij}|X_i = x_i)$, for $j = 1, \ldots, M$ can be obtained as follows

$$\begin{aligned}\tau_{ij}(\boldsymbol{\Psi}) &= \mathbb{E}(Z_{ij}|X_i = x_i) \\ &= \mathbb{P}(Z_{ij} = 1|X_i = x_i) \\ &= \frac{\pi_j f(x_i;\theta_j)}{\sum_{j=1}^{M} \pi_j f(x_i;\theta_j)} \end{aligned} \tag{2.13}$$

In the case of the mixture of location-shifted models (2.3), we know that $\mathcal{F} = \Big\{ f(\cdot|\mu) = f(\cdot - \mu),\ \mu \in R \Big\}$. Hence, the conditional expectation of latent variables is given by

$$\tau_{ij}(\boldsymbol{\Psi}) = \frac{\pi_j f(x_i - \mu_j)}{\sum_{j=1}^{M} \pi_j f(x_i - \mu_j)} \quad j = 1, \ldots, M;\ \ i = 1, \ldots, n. \tag{2.14}$$

In the EM algorithm, we try to maximize the conditional expectation of the complete data log-likelihood to obtain the ML estimates of the FMM parameters. To do so, from equation (2.11) the conditional expectation of the complete data log-likelihood is given by

$$Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^*) = \mathbb{E}\Big(l_{\mathbf{y}}(\boldsymbol{\Psi})|\mathbf{x}, \boldsymbol{\Psi}^* \Big) \tag{2.15}$$

$$\begin{aligned} &= \sum_{i=1}^{n}\sum_{j=1}^{M} \mathbb{E}(Z_{ij}|\mathbf{x}, \boldsymbol{\Psi}^*) \Big(\log \pi_j + \log f(x_i;\theta_j) \Big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\Psi}^*) \log \pi_j + \sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\Psi}^*) \log f(x_i;\theta_j) \\ &= Q_1(\boldsymbol{\pi}, \boldsymbol{\Psi}^*) + Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^*), \end{aligned}$$

where $\tau_{ij}(\boldsymbol{\Psi}^*)$ is obtained from (2.14). EM algorithm, as an iterative algorithm, requires initialization step.

0-step: Initialization

Let $\boldsymbol{\Psi}^{(0)} = (\boldsymbol{\pi}^{(0)}, \boldsymbol{\xi}^{\top(0)})$ represents the initial values of $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi}^{\top})$.

In order to better describe the iterative steps of EM algorithm, let $\boldsymbol{\Psi}^{(p)}$ represents the estimate of $\boldsymbol{\Psi}$ updated from the pth iteration of the EM algorithm. Now we shall update $\boldsymbol{\Psi}^{(p+1)}$.

E-step:

First, we use (2.14) and $\boldsymbol{\Psi}^{(p)}$ and compute the $\tau_{ij}(\boldsymbol{\Psi}^{(p)})$ as follows

$$\tau_{ij}(\boldsymbol{\Psi}^{(p)}) = \frac{\pi_j^{(p)} f(x_i - \mu_j^{(p)})}{\sum_{j=1}^{M} \pi_j^{(p)} f(x_i - \mu_j^{(p)})} \qquad j = 1, \ldots, M; \;\; i = 1, \ldots, n. \tag{2.16}$$

We then use the $\boldsymbol{\Psi}^{(p)}$ and (2.15) and (2.16) to update the conditional expectation of the complete data log-likelihood as follows

$$\begin{aligned} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^*)|_{\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^{(p)}} &= Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}) \\ &= \mathbb{E}(l_{\mathbf{y}}(\boldsymbol{\Psi})|\mathbf{x}, \boldsymbol{\Psi}^{(p)}) \\ &= Q_1(\boldsymbol{\pi}, \boldsymbol{\Psi}^{(p)}) + Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}), \end{aligned} \tag{2.17}$$

where from (2.15) and (2.16); we have

$$Q_1(\boldsymbol{\pi}, \boldsymbol{\Psi}^{(p)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\Psi}^{(p)}) \log \pi_j \tag{2.18}$$

and

$$Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\Psi}^{(p)}) \log f(x_i; \theta_j), \tag{2.19}$$

and $\tau_{ij}(\boldsymbol{\Psi}^{(p)}$ is obtained from equation (2.16).

M-step:

Once we updated the $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)})$ from the E-step of the $(p+1)$-th iteration, the $\boldsymbol{\Psi}^{(p+1)}$ is obtained the solution to:

$$\boldsymbol{\Psi}^{(p+1)} = \underset{\boldsymbol{\Psi}}{\operatorname{argmax}} \;\; Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(p)}) \tag{2.20}$$

From the decomposition (2.17), we can maximize the mixing proportions and component parameters separately. From (2.14) and (2.15) and using the Lagrangian Multiplier, one can update the mixing proportions $\boldsymbol{\pi}^{(p+1)}$ as the solution to

$$Q_1(\boldsymbol{\pi}; \boldsymbol{\Psi}^{(p)}) - \lambda(\sum_{j=1}^{M} \pi_j - 1) = 0 \tag{2.21}$$

By differentiating (2.21) from π_j, $j = 1, \ldots, M$, with respect to the constraint $\sum_{j=1}^{M} \pi_j = 1$, it is easy to obtain the $\pi_j^{(p+1)}$ as follows

$$\pi_j^{(p+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ij}(x_i : \boldsymbol{\Psi}^{(p)}). \tag{2.22}$$

In addition, one can find the update $\boldsymbol{\xi}^{(p+1)}$ as solution to the following equation:

$$\boldsymbol{\xi}^{(p+1)} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}} \;\; Q_2(\boldsymbol{\xi}, \boldsymbol{\Psi}^{(p)})$$

Finally, we obtain the ML estimate, $\hat{\boldsymbol{\Psi}}_{ML}$, of the parameters $\boldsymbol{\Psi}$ of the undelying mixture model by alternating the E-step and M-step until the EM algorithm converges and $\|\boldsymbol{\Psi}^{(p+1)} - \boldsymbol{\Psi}^{(p)}\|_{\infty}$ becomes negligible.

2.3 A Semi-parametric EM Algorithm

The previous section provides some information on the estimation of the unknown parameters of the mixture model by using the EM algorithm in a parametric framework. Based on model (2.3), the component density f is also unknown in the semi-parametric location-shifted mixture model. In this framework, we should treat f also as another parameter of the mixture model. Hence we have to designate the EM algorithm such that not only we estimate the ML estimation of mixing proportion $\boldsymbol{\pi}$, and the location parameter $\boldsymbol{\xi} = (\mu_1, \ldots, \mu_M)$, but also we have to estimate f parameter as well. In this semi-parametric framework, let $\boldsymbol{\xi}^{\top} = (\mu_1, \ldots, \mu_M)$ represents the vector of all unknown location parameters and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ represents the vector of unknown mixing proportions. We show the vector of all known parameters of semi-parametric location-shifted mixture model such that $\boldsymbol{\zeta} = (\boldsymbol{\Psi}, f)$ where $\boldsymbol{\Psi} = (\boldsymbol{\pi}, \boldsymbol{\xi}^{\top})$. Let $X_1, \ldots, X_n$ be a random sample of size n from the semi-parametric mixture of location-shifted models. Using the $\boldsymbol{\zeta}$, the vector of all unknown parameters of the model, the pdf of X_i, $i = 1, \ldots, n$, is given by

$$g(x_i; \boldsymbol{\zeta}) = \sum_{j=1}^{M} \pi_j f(x_i - \mu_j). \tag{2.23}$$

Similar to the parametric setting, we introduce latent variable $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ for each x_i, $i = 1, \ldots, n$, such that

$$Z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ belongs to component } j; \\ 0 & \text{otherwise.} \end{cases}$$

Based on collection of latent variables and incomplete data, we introduce $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ as complete data where $\mathbf{Y}_i = (X_i, \boldsymbol{Z_i})$. We can easily write that $\boldsymbol{Z_i} = (Z_{i1}, \ldots, Z_{iM}) \sim Multi(1, \pi_1, \ldots, \pi_M)$ such that the pdf of latent variable $\boldsymbol{Z_i}$ is given by

$$f(\boldsymbol{z_i}) = \prod_{j=1}^{M} (\pi_j)^{Z_{ij}} \tag{2.24}$$

Using (2.23) and (2.24), the joint pdf of $Y_i = (X_i, \boldsymbol{Z_i})$; $i = 1, \ldots, n$, is given by

$$\begin{aligned} f(y_i; \boldsymbol{\zeta}) &= f\big(x_i, \boldsymbol{z_i}; \boldsymbol{\zeta}\big) \\ &= \prod_{j=1}^{M} \Big\{ \pi_j f(x_i - \mu_j) \Big\}^{z_{ij}}. \end{aligned} \tag{2.25}$$

From (2.25), the complete data likelihood function of $\boldsymbol{\zeta}$ is given by

$$\begin{aligned} L_{\mathbf{y}}(\boldsymbol{\zeta}) &= \prod_{i=1}^{n} f(y_i; \boldsymbol{\zeta}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{M} \Big\{ \pi_j f(x_i - \mu_j) \Big\}^{z_{ij}}, \end{aligned} \tag{2.26}$$

and the complete data log-likelihood function can be written as follows

$$\begin{aligned} l_{\mathbf{y}}(\boldsymbol{\zeta}) &= \log L_{\mathbf{y}}(\boldsymbol{\zeta}) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log \pi_j + \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log f(x_i - \mu_j). \end{aligned} \tag{2.27}$$

2.3.1 The EM Algorithm: A Semi-parametric Framework

In this subsection, we use the EM algorithm to obtain the ML estimate of the parameters $\boldsymbol{\zeta}$ of the mixture of location-shifted models in a semi-parametric framework. As we observed in the previous section, we require the conditional distribution and conditional expectation of the latent variables given the incomplete data.

One can use the joint distribution od (X_i, Y_i) and the marginal distribution of X_i from (2.23) to obtain the conditional distribution of $\boldsymbol{Z_i}|X_i = x_i$. The conditional expectation of $\boldsymbol{Z_i}|X_i = x_i$ is then given by

$$\begin{aligned}\tau_{ij}(\boldsymbol{\zeta}^*) &= \mathbb{E}(Z_{ij}|\mathbf{x}, \boldsymbol{\zeta}^*) \\ &= p(Z_{ij} = 1|\mathbf{x}, \boldsymbol{\zeta}^*) \\ &= \frac{\pi_j^* f^*(x_i - \mu_j^*)}{\sum_{j=1}^M \pi_j^* f^*(x_i - \mu_j^*)}. \end{aligned} \tag{2.28}$$

It should note that in the above expectation, we not only have to use $\boldsymbol{\Psi}^*$, the mixture parameters updated from the previous iteration of the EM algorithm, but also we require f^*, the non-parametric estimate of density f from the previous iteration of the EM algorithm in this semi-parametric framework. Once we complete the conditional expectation of the latent variables, the conditional expectation of the complete data log-likelihood given incomplete data completed as follows

$$\begin{aligned}Q(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*) &= \mathbb{E}\Big(l_{\mathbf{y}}(\boldsymbol{\zeta})|\mathbf{x}, \boldsymbol{\zeta}^* \Big) \\ &= \sum_{i=1}^n \sum_{j=1}^M \tau_{ij}(\boldsymbol{\zeta}^*) \log \pi_j + \sum_{i=1}^n \sum_{j=1}^M \tau_{ij}(\boldsymbol{\zeta}^*) \log f(x_i - \mu_j) \\ &= Q_1(\boldsymbol{\pi}, \boldsymbol{\zeta}^*) + Q_2(\boldsymbol{\xi}^\top, f, \boldsymbol{\zeta}^*), \end{aligned} \tag{2.29}$$

where $\tau_{ij}(\boldsymbol{\zeta}^*)$ is computed by (2.28).

To estimate the unknown parameters $\boldsymbol{\zeta}$ by using the EM algorithm, we require to start the EM algorithm with initial values.

0-step: Initialization

Let $\boldsymbol{\zeta}^{(0)} = (\boldsymbol{\Psi}^{(0)}, f^{(0)})$ be the initial values of the EM algorithm in the mixture of location-shifted models from (2.23) in the semi-parametric setting.

Similar to the previous section, let $\boldsymbol{\zeta}^{(p)} = (\boldsymbol{\Psi}^{(p)}, f^{(p)})$ denotes the update of the parameters of the mixture model from pth iteration of the EM algorithm. We shall use $\boldsymbol{\zeta}^{(p)}$ and update $\boldsymbol{\zeta}^{(p+1)}$.

E-step:

From equation (2.28) and updated $\boldsymbol{\zeta}^{(p)}$, we can update the conditional expectation of the latent variable as follows

$$\tau_{ij}(\boldsymbol{\zeta}^{(p)}) = \frac{\pi_j^{(p)} f^{(p)}(x_i - \mu_j^{(p)})}{\sum_{j=1}^{M} \pi_j^{(p)} f^{(p)}(x_i - \mu_j^{(p)})} \tag{2.30}$$

Using above equation, the conditioanal expectation of the complete data log-likelihood is given by

$$\begin{aligned} Q(\boldsymbol{\zeta}, \boldsymbol{\zeta}^*)|_{\boldsymbol{\zeta}^*=\boldsymbol{\zeta}^{(p)}} &= Q(\boldsymbol{\zeta}, \boldsymbol{\zeta}^{(p)}) \\ &= \mathbb{E}(l_{\mathbf{y}}(\boldsymbol{\zeta})|\mathbf{x}, \boldsymbol{\zeta}^{(p)}) \\ &= Q_1(\boldsymbol{\pi}, \boldsymbol{\zeta}^{(p)}) + Q_2(\boldsymbol{\xi}, f, \boldsymbol{\zeta}^{(p)}), \end{aligned} \tag{2.31}$$

where

$$Q_1(\boldsymbol{\pi}, \boldsymbol{\zeta}^{(p)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\zeta}^{(p)}) \log \pi_j, \tag{2.32}$$

and

$$Q_2(\boldsymbol{\xi}, f, \boldsymbol{\zeta}^{(p)}) = \sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\zeta}^{(p)}) \log f(x_i - \mu_j), \tag{2.33}$$

and $\tau_{ij}(\boldsymbol{\zeta}^{(p)})$ is obtained from equation (2.30).

M-step: Based on the decomposition (2.31), we can maximize the mixing proportions based on $Q_1(\boldsymbol{\pi}, \boldsymbol{\zeta}^{(p)})$ separate from other parameters of the model. Hence, the $\pi_j^{(p+1)}$, $j = 1, \ldots, M$, can be obtained as solution to:

$$\sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\zeta}^{(p)}) \log \pi_j - \lambda(\sum_{j=1}^{M} \pi_j - 1) = 0 \tag{2.34}$$

From the above lagrangian multipliers, one can easily show that

$$\pi_j^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}(\boldsymbol{\zeta}^{(p)}), \tag{2.35}$$

where $\tau_{ij}(\boldsymbol{\zeta}^{(p)})$ is obtained from equation (2.30).

To update $\mu_j^{(p+1)}$; one requires to maximize $Q_2(\boldsymbol{\xi}, f, \boldsymbol{\zeta}^{(p)})$ from (2.33) with respect to μ_j. According to the fact that f belongs to the location-shifted family of distribution, hence we can easily see that μ_j is the mean, median, and mode of f given $Z_{ij} = 1$; i.e., $\mathbb{E}(X_i|Z_{ij} = 1) = \mu_j$. Therefore, the ML estimate of μ_j is the same with the method of moments estimate of μ_j. Accordingly, one can obtain the $\mu_j^{(p+1)}$ as

$$\mu_j^{(p+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}(\boldsymbol{\zeta}^{(p)}) x_i}{\sum_{i=1}^{n} \tau_{ij}(\boldsymbol{\zeta}^{(p)})} \qquad j = 1, \ldots, M. \tag{2.36}$$

S-step: Now, we need to update the pdf $f^{(p+1)}$. Bordes et al. (2007b) suggested to estimate the pdf f by using a nonparametric density estimate based on $\mathbf{x} = (x_1, \ldots, x_n)$. To be able to estimate f nonparametrically, we require to have data generated from f distribution. The challenge here is that the incomplete data $(x_1, \ldots, x_n)$ are from the mixture of the location-shifted models and they are not observations from f distribution. Bordes et al. (2007b) work with observations centered back to f instead of using x_i observations directly. We denote the vector of observations "centered back" by $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$. If we assume that we have the complete-data $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ and that $\boldsymbol{\Psi}$ is known, then we can estimate the pdf f by the two following steps;

1. We need to compute $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_n)$, where $\tilde{x}_i = x_i - \mu_{z_i}, (i = 1, \ldots, n)$.
2. Estimate the pdf f by using a Kernel density as follows

$$\hat{f}_{\tilde{\mathbf{x}}}(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_i}{h_n}\right)$$

where K is a Kernel function and h_n denotes the bandwith. In this case, if we consider that $\boldsymbol{z_i}$ is missing and the true parameter $\boldsymbol{\zeta}$ is known, then it is difficult to recover a sample from f. Hence we recover a sample from f by the following steps;

S1-step: We can simulate the stochastic version of the $\tau_{ij}(\boldsymbol{\zeta}^{(p)})$ as follows $Z(x_i, \boldsymbol{\zeta}) \sim Mult\Big(1, \tau_{i1}(\boldsymbol{\zeta}), \ldots, \tau_{iM}(\boldsymbol{\zeta})\Big)$, for $i = 1, \ldots, n$.

S2-step: We can obtain observations centered back through $\tilde{x}_i = x_i - \mu_{Z(x_i;\boldsymbol{\zeta})}$, for $i = 1, \ldots, n$.

In the following lemma, we can show that the centered-back observations $(\tilde{x}_1, \ldots, \tilde{x}_n)$

obtained from the above algorithm are observations from f distribution. The proof can be found in Bordes et al. (2007b). For the sake of completeness, we provide the lemma and its proof in the following lemma.

Lemma 2.1. *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a sample from the pdf* $g(x; \boldsymbol{\zeta})$ (2.23). *The centered back observations* $\tilde{\mathbf{X}} = (\tilde{X}_1, \ldots, \tilde{X}_n)$ *is a sample from* f.

Proof. Since $\mathbf{X} = (X_1, \ldots, X_n)$ is an i.i.d. sample from mixture model $g(x; \boldsymbol{\zeta})$, it is enough to check this property for one observation. Let $X \sim g(x; \boldsymbol{\zeta})$ and $\tilde{X} = X - \mu_{Z(x,\boldsymbol{\zeta})}$ as described in S1 and S2 steps.

$$
\begin{aligned}
\mathbb{P}_{\boldsymbol{\zeta}}(\tilde{X} < y) &= \int \mathbb{P}(\tilde{X} < y, X = x)dx \\
&= \int \mathbb{P}(\tilde{X} < y | X = x) g(x; \boldsymbol{\zeta})dx \\
&= \int \mathbb{P}\Big(x - \mu_{Z(x,\boldsymbol{\zeta})} < y\Big) g(x; \boldsymbol{\zeta})dx \\
&= \int \sum_{j=1}^{M} \mathbb{P}\Big(x - \mu_j < y | Z(x; \boldsymbol{\zeta}) = j\Big) \mathbb{P}(Z(x; \boldsymbol{\zeta}) = j) g(x; \boldsymbol{\zeta})dx \\
&= \int \sum_{j=1}^{M} \mathbb{P}\Big(x - \mu_j < y | Z(x; \boldsymbol{\zeta}) = j\Big) \frac{\pi_j f(x; \mu_j)}{g(x; \boldsymbol{\zeta})} g(x; \boldsymbol{\zeta})dx \\
&= \sum_{j=1}^{M} \pi_j \int \mathbb{I}_{(x-\mu_j<y)} f(x - \mu_j)dx \\
&= \Big(\sum_{j=1}^{M} \pi_j\Big) \mathbb{P}(X < y) \\
&= \mathbb{P}(X < y) = F_X(y)
\end{aligned}
$$

where F is the cdf of X. $\square$

As we can see, in Lemma 2.1, we can assume that $\boldsymbol{\zeta}$ is known. Similarly, the $\boldsymbol{\zeta}$ parameters are replaced by $\boldsymbol{\zeta}^{(p)}$ updated from the previous iteration and they are treated known as well in each iteration of the EM algorithm.

We show the complete step $\boldsymbol{\zeta}^{(p)} \rightarrow \boldsymbol{\zeta}^{(p+1)}$ of the semi-parametric EM algorithm (SEM) as follows;

E – Step : From (2.30) we compute $\tau_{ij}(\boldsymbol{\zeta}^{(p)})$ for $i = 1, \ldots, n$, $j = 1, \ldots, M$.

S – Step :

(**S1**) We simulate stochastic centered back observations from f distribution as follows

$$Z^{(p+1)}(x_i, \boldsymbol{\zeta}^{(p)}) \sim Mult\Big(1, \tau_{i1}(x_i; \boldsymbol{\zeta}^{(p)}), \ldots, \tau_{iM}(x_i; \boldsymbol{\zeta}^{(p)})\Big)$$

(**S2**) We simulate

$$\tilde{x}_i^{(p+1)} = x_i - \mu^{(p)}_{Z^{(p+1)}(x_i, \boldsymbol{\zeta}^{(p)})}$$

Using the centered data $\tilde{\mathbf{x}}^{(p+1)}$, we obtain a kernel density estimate $f^{(p+1)}$ of f by using a symmetric assumption in the model. Therefore, we have

(**S3**) Kernel density estimate:

$$\hat{f}_{\tilde{\mathbf{x}}^{(p+1)}}(u) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{u - \tilde{x}_i^{(p+1)}}{h_n}\right) \tag{2.37}$$

(**S4**) Symmetrization:

$$f^{(p+1)}(u) = \frac{\hat{f}_{\tilde{\mathbf{x}}^{(p+1)}}(u) + \hat{f}_{\tilde{\mathbf{x}}^{(p+1)}}(-u)}{2} \tag{2.38}$$

M − Step : We update the parameter $\boldsymbol{\Psi}^{(p+1)}$ as follows

$$\pi_j^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n} \tau_{ij}(x_i : \boldsymbol{\zeta}^{(p)}),$$

and

$$\mu_j^{(p+1)} = \frac{\sum_{i=1}^{n} \tau_{ij}(x_i : \boldsymbol{\zeta}^{(p)})x_i}{\sum_{i=1}^{n} \tau_{ij}(x_i : \boldsymbol{\zeta}^{(p)})}$$

We obtain ML estimate of the parameters $\boldsymbol{\zeta} = (\boldsymbol{\Psi}, f)$ by alternating the E-step, S-step and M-step until $\|\boldsymbol{\Psi}^{(p+1)} - \boldsymbol{\Psi}^{(p)}\|_\infty$ becomes negligible.

Chapter 3

Semi-parametric Mixture Models with RSS Data

In the standard methods of modeling and inference for finite mixture model (FMM), we typically assume that we have samples drawn from the underlying mixture model from simple random sampling (e.g., McLachlan and Peel, 2004; Titterington et al., 1985). In many applications, for instance in fishery studies, using ranked set sampling (RSS) can be more cost-effective and more advantageous. We can also get the results in better and more informative sampling from the underlying mixture models (Hatefi et al., 2014).

In Chapter 2, we discussed semi-parametric modeling of the FMMs based on the SRS data. This chapter investigates the maximum likelihood (ML) estimation of the FMMs based on the RSS data. We also develop an estimate of FMMs based on RSS data in a semi-parametric framework. To do so, we also develop an EM algorithm to obtain the RSS-based ML estimate of FMMs. This chapter is organized as follows. In Section 3.1, we discuss the ML estimation of unknown parameters of

FMMs with RSS data in a parametric framework. We develop the ML estimation of semi-parametric FMMs based on RSS data in Section 3.2.

3.1 Parametric Estimation of FMMs with RSS

We consider X as a random variable that follows the FMM (2.2). Let $\left\{X_{[r]i}; r = 1,\ldots,H, i = 1,\ldots,n\right\}$ denote an RSS data of size nH with set size H and the number of cycles n. We construct the RSS data as described in Section 1.2 in Chapter 1. As described in Chapter 1, in imperfect RSS, we use a concomitant variable to rank the sampling units in each set. Hence the assigned ranks to sampling units may be different from their true ranks of the units. We employ the method of Hatefi et al. (2015) to model the ranking errors in RSS. It is possible to model the selection of $X_{[r]i}$ using a latent variable method that accommodates the possibility of ranking error between the judgmental ranks and the true ranks in each set. Let $\boldsymbol{\alpha}$ denote the misplacement probability matrix,

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \ldots & \alpha_{1,H} \\ \alpha_{2,1} & \alpha_{2,2} & \ldots & \alpha_{2,H} \\ \vdots & \vdots & \ldots & \vdots \\ \alpha_{H,1} & \alpha_{H,2} & \ldots & \alpha_{H,H} \end{bmatrix}_{H\times H},$$

where we define $\alpha_{r,h}$ as the probability that the h-th order statistic (true rank) received the h-th judgmental rank (i.e. the assigned rank); hence

$$\alpha_{r,h} = \mathbb{P}\Big(X_{[r]} = X_{(r)}\Big).$$

It is at once apparent that $\boldsymbol{\alpha}$ matrix should be doubly stochastic matrix such that $\sum_{h=1}^{H} \alpha_{r,h} = \sum_{r=1}^{H} \alpha_{h,r} = 1$. For each $X_{[r]i}$, we define an H dimensional random vector, $\boldsymbol{\Delta}_i^{[r]}$, which follows a multinomial distribution with parameters 1 and $\boldsymbol{\alpha}_{[r]}$, where $\boldsymbol{\alpha}_{[r]} = (\alpha_{r,1}, \ldots, \alpha_{r,H})$. Then we can write

$$\boldsymbol{\Delta}_i^{[r]} = \left(\Delta_i^{[r,1]}, \ldots, \Delta_i^{[r,H]} \right),$$

which has only one nonzero entry and all the other entries are zero. Since $\boldsymbol{\Delta}_i^{[r]}$ is a random vector, we can write

$$\mathbb{P}(\boldsymbol{\Delta}_i^{[r]} = \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\alpha}) = \prod_{h=1}^{H} \left\{ \alpha_{r,h} \right\}^{\delta_i^{[r,h]}}. \tag{3.1}$$

Also we define the $g^{(h:H)}(x; \boldsymbol{\Psi})$ as the pdf of the h-th order statistic in a set of size H from FMM (2.2) as follows

$$g^{(h:H)}(x; \boldsymbol{\Psi}) = H \binom{H-1}{h-1} g(x; \boldsymbol{\Psi}) [G(x; \boldsymbol{\Psi})]^{h-1} [\bar{G}(x; \boldsymbol{\Psi})]^{H-h}, \tag{3.2}$$

where $\bar{G}(\cdot; \boldsymbol{\Psi}) = 1 - G(\cdot; \boldsymbol{\Psi})$ and

$$g(x; \boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j f(x, \theta_j),$$

and

$$G(x; \boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j F(x, \theta_j),$$

represent the pdf and cdf of the FMM 2.2. Therefore, we can write the conditional distribution of $X_{[r]i}$ given $\boldsymbol{\Delta}_i^{[r]}$ as

$$f(x_{[r]i}|\boldsymbol{\delta}^{[r]};\boldsymbol{\Psi}) = \prod_{h=1}^{H}\left\{g^{(h:H)}(x_{[r]i};\boldsymbol{\Psi})\right\}^{\delta_i^{[r,h]}}. \tag{3.3}$$

In addition, the joint distribution of $(X_{[r]i}, \boldsymbol{\Delta}_i^{[r]})$ follows from (3.1) and (3.3) and is given by

$$\begin{aligned} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]};\boldsymbol{\Omega}) &= \prod_{h=1}^{H}\left\{\alpha_{r,h}g^{(h:H)}(x_{[r]i};\boldsymbol{\Psi})\right\}^{\delta_i^{[r,h]}} \\ &= \prod_{h=1}^{H}\left\{\alpha_{r,h}H\binom{H-1}{h-1}g(x;\boldsymbol{\Psi})[G(x;\boldsymbol{\Psi})]^{h-1}[\bar{G}(x;\boldsymbol{\Psi})]^{H-h}\right\}^{\delta_i^{[r,h]}}. \end{aligned} \tag{3.4}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Psi}, \boldsymbol{\alpha})$. Then we obtain the marginal distribution of $X_{[r]i}$ as follows

$$f(x_{[r]i};\boldsymbol{\Omega}) = \sum_{\boldsymbol{\delta}_i^{[r]}} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]};\boldsymbol{\Omega}) = \sum_{h=1}^{H}\alpha_{r,h}g^{(h:H)}(x_{[r]i};\boldsymbol{\Psi}). \tag{3.5}$$

Form (3.4) and (3.5), we can easily write

$$f(\boldsymbol{\delta}_i^{[r]}\big|x_{[r]i};\boldsymbol{\Omega}) = \prod_{h=1}^{H}\left\{\frac{\alpha_{r,h}B_{h,H+1-h}(G(x_{[r]i};\boldsymbol{\Psi}))}{\sum_{h=1}^{H}\alpha_{r,h}B_{h,H+1-h}(G(x_{[r]i};\boldsymbol{\Psi}))}\right\}^{\delta_i^{[r,h]}}, \tag{3.6}$$

where $B_{a,b}(\cdot)$ represents a beta density function with parameters a and b.

3.1.1 Likelihood Functions based on RSS Data

We estimate the unknown parameter $\boldsymbol{\Psi}$ of the FMM (2.2) based on RSS data through ML method in this section. According to the fact that RSS statistics $\left\{X_{[r]i}; r = 1, \ldots, H, \quad i = 1, \ldots, n\right\}$ are independent, we can obtain the likelihood function for the RSS data from equation (3.5) as follows

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^{n} \prod_{r=1}^{H} f(x_{[r]i}; \boldsymbol{\Omega}). \tag{3.7}$$

The log-likelihood function is also given by

$$l(\boldsymbol{\Omega}) = \log L(\boldsymbol{\Omega}).$$

We define the ML estimator of $\boldsymbol{\Psi}$, denoted by $\widehat{\boldsymbol{\Psi}}$, as an appropriate solution to

$$\widehat{\boldsymbol{\Psi}}_{ML} = \underset{\boldsymbol{\Psi}}{\operatorname{argmax}} \; l(\boldsymbol{\Omega}).$$

This optimization is practically untractable with respect to the parameters of the components $\boldsymbol{\Psi}$ and sampling parameters $\boldsymbol{\alpha}$. To solve this problem we have to rewrite the likelihood function $L(\boldsymbol{\Omega})$ through a latent variable model (Hatefi et al., 2015). Therefore, we write the likelihood function of $\boldsymbol{\Omega}$ by using the latent multinomial random variables $\boldsymbol{\Delta}_i^{[r]}$ from equation (3.4) as follows

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^{n} \prod_{r=1}^{H} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}). \tag{3.8}$$

We can easily obtain the ML estimator of $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\alpha}}_{ML}$, by using the EM algorithm in the conditional expected log-likelihood function given the RSS data. Although

the likelihood function (3.8) is tractable with respect to sampling parameters; the likelihood function is still untractable with respect to $\mathbf{\Psi}$, due to the presence of $\sum_{i=1}^{n}\sum_{r=1}^{H}\log\Big(g(x,\mathbf{\Psi})\Big)$ and $\sum_{i=1}^{n}\sum_{r=1}^{H}\log\Big(G(x,\mathbf{\Psi})\Big)$ terms in the logarithm of the likelihood function (3.8). To overcome the problem, we have to view the $\Big\{X_{[r]i}; r=1,\ldots,H, \quad i=1,\ldots,n\Big\}$ as incomplete RSS data. Given $\mathbf{\Delta}_i^{[r]}=\boldsymbol{\delta}_i^{[r]}$, we have to introduce three additional latent variables $\mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}, \mathbf{V}_i^{[r]}$ for each $x_{[r]i}$. We consider $\{\Delta_i^{[r,h]}=1\}$ as the event that the rth entry of the vector $\mathbf{\Delta}_i^{[r]}$ is only one and other elements are zero. We define an M-dimensional latent vector $\mathbf{Z}_i^{[r]}=\{Z_{i1}^{[r]},\ldots,Z_{iM}^{[r]}\}$, where $Z_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}$ is one if $x_{[r]i}$ belongs to the j-th component of the mixture model $j=1,\ldots,M$. That is,

$$Z_{ij}^{[r]}\big|\{\Delta_i^{[r,h]}=1\}=\begin{cases}1 & \text{if } x_{[r]i} \text{ belongs to component } j;\\ 0 & \text{otherwise,}\end{cases}$$

where

$$\sum_{j=1}^{M}\Big(Z_{ij}^{[r]}\big|\{\Delta_i^{[r,h]}=1\}\Big)=1.$$

Thus $\mathbf{Z}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}$, $i=1,\ldots,n; r=1,\ldots,H$, has a multinomial distribution including one draw on M classes with probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_M)$, we can write

$$f(\mathbf{z}_i^{[r]}\big|\boldsymbol{\delta}_i^{[r]};\boldsymbol{\pi})=\prod_{h=1}^{H}\left\{\binom{1}{z_{i1}^{[r]},\ldots,z_{iM}^{[r]}}\prod_{j=1}^{M}\pi_j^{z_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}. \tag{3.9}$$

Given $\mathbf{\Delta}_i^{[r]}=\boldsymbol{\delta}_i^{[r]}$, we also consider $\mathbf{W}_i^{[r]}=\{W_{i1}^{[r]},\ldots,W_{iM}^{[r]}\}$ as an M-dimensional vector, where $W_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}$ denotes the number of observations less than $x_{[r]i}$ in

the set which come from the component j of FMM (2.2) with

$$\sum_{j=1}^{M}\left(W_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}\right)=h-1.$$

Accordingly, the latent vectors $\mathbf{W}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}, i=1,\ldots,n; r=1,\ldots,H$, follows a multinomial distribution including $h-1$ draws on M classes with probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_M)$ and we can write

$$f(\mathbf{w}_i^{[r]}|\boldsymbol{\delta}_i^{[r]};\boldsymbol{\pi})=\prod_{h=1}^{H}\left\{\binom{h-1}{w_{i1}^{[r]},\ldots,w_{iM}^{[r]}}\prod_{j=1}^{M}\pi_j^{w_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}. \tag{3.10}$$

Finally, given $\boldsymbol{\Delta}_i^{[r]}=\boldsymbol{\delta}_i^{[r]}$ we define $\mathbf{V}_i^{[r]}=\{V_{i1}^{[r]},\ldots,V_{iM}^{[r]}\}$ as an M-dimensional vector, where $V_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}$ shows the number of observations bigger than $x_{[r]i}$ in the set that come from the component j of the FMM (2.2) with

$$\sum_{j=1}^{M}\left(V_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}\right)=H-h.$$

The vectors $\mathbf{V}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}, i=1,\ldots,n; r=1,\ldots,H$, also has a multinomial distribution including $H-h$ draws on M classes with probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_M)$, and we have

$$f(\mathbf{v}_i^{[r]}|\boldsymbol{\delta}_i^{[r]};\boldsymbol{\pi})=\prod_{h=1}^{H}\left\{\binom{H-h}{v_{i1}^{[r]},\ldots,v_{iM}^{[r]}}\prod_{j=1}^{M}\pi_j^{v_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}. \tag{3.11}$$

It is easy to see that the latent variables $\mathbf{Z}_i^{[r]}$, $\mathbf{W}_i^{[r]}$ and $\mathbf{V}_i^{[r]}$ given $\boldsymbol{\Delta}_i^{[r]}$ are conditionally independent. We present the joint distribution of $X_{[r]i}$ and its latent variables

$\boldsymbol{\Delta}_i^{[r]}, \mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}$ and $\mathbf{V}_i^{[r]}$ in the following lemma, where proof can be found in (Hatefi et al., 2015).

Lemma 3.1. *For fixed values i and r, $i = 1, \ldots, n$, $r = 1, \ldots, H$, we have*

$$f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega}) \propto \prod_{h=1}^{H}\prod_{j=1}^{M}\left\{\alpha_{r,h} \quad \pi_j^{\{z_{ij}^{[r]}+w_{ij}^{[r]}+v_{ij}^{[r]}\}}\right\}^{\delta_i^{[r,h]}}$$

$$\times \left\{[f(x_{[r]i}, \theta_j)]^{z_{ij}^{[r]}}[F(x_{[r]i}, \theta_j)]^{w_{ij}^{[r]}}[\bar{F}(x_{[r]i}, \theta_j)]^{v_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}.$$

Using all the latent variables, we introduce the complete RSS data as

$$\mathbf{Y}_{RSS} = \{X_{[r]i}, \boldsymbol{\Delta}_i^{[r]}, \mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}, \mathbf{V}_i^{[r]}, i = 1, \ldots, n; r = 1, \ldots, H\}.$$

From Lemma 3.1, likelihood function based on complete RSS data is given by

$$L(\boldsymbol{\Omega}|\mathbf{y_{RSS}}) = \prod_{i=1}^{n}\prod_{r=1}^{H} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega}).$$

The complete data log-likelihood function of $\boldsymbol{\Omega}$ based on RSS data is given by

$$l(\boldsymbol{\Omega}|\mathbf{y_{RSS}}) \propto l_1(\boldsymbol{\alpha}|\mathbf{y_{RSS}}) + l_2(\boldsymbol{\pi}|\mathbf{y_{RSS}}) + l_3(\boldsymbol{\xi}|\mathbf{y_{RSS}}), \tag{3.12}$$

where

$$l_1(\boldsymbol{\alpha}|\mathbf{y_{RSS}}) = \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\log\alpha_{r,h},$$

$$l_2(\boldsymbol{\pi}|\mathbf{y_{RSS}}) = \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\sum_{j=1}^{M}\left[\left\{z_{ij}^{[r]} + w_{ij}^{[r]} + v_{ij}^{[r]}\right\}\log\pi_j\right],$$

$$l_3(\boldsymbol{\xi}|\mathbf{y_{RSS}}) = \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\times\sum_{j=1}^{M}\left[z_{ij}^{[r]}\log f_j(x_{[r]i};\theta_j) + w_{ij}^{[r]}\log F_j(x_{[r]i};\theta_j) + v_{ij}^{[r]}\log\bar{F}_j(x_{[r]i};\theta_j)\right].$$

3.1.2 EM Algorithm for RSS Data

In this section, we present an EM algorithm to maximize the log-likelihood function (3.12). The EM algorithm starts with an initial value of the population parameter $\boldsymbol{\Omega}^{(0)}$. To compute the EM algorithm, we require to fit the conditional expectation of the latent variables given incomplete RSS data. From equation (3.6), the conditional expectation of $\Delta_i^{[r,h]}|\{X_{[r]i} = x_{[r]i}\}$ is given by

$$\phi_i^{[r,h]}(\boldsymbol{\Omega}) = \frac{\alpha_{r,h} B_{h,H-h+1}(G(x_{[r]i}; \boldsymbol{\Psi}))}{\sum_{h=1}^{H} \alpha_{r,h} B_{h,H-h+1}(G(x_{[r]i}; \boldsymbol{\Psi}))}, \tag{3.13}$$

where $B_{a,b}(\cdot)$ is a beta density function with parameters a and b. In addition, from Lemma 3.1, equation (3.4) and independence conditional of latent variables $\boldsymbol{Z}_{\boldsymbol{i}}^{[\boldsymbol{r}]}, \boldsymbol{W}_{\boldsymbol{i}}^{[\boldsymbol{r}]}, \boldsymbol{V}_{\boldsymbol{i}}^{[\boldsymbol{r}]}$, we derived the conditional distribution of latent variables $\boldsymbol{Z}_{\boldsymbol{i}}^{[\boldsymbol{r}]}, \boldsymbol{W}_{\boldsymbol{i}}^{[\boldsymbol{r}]}, \boldsymbol{V}_{\boldsymbol{i}}^{[\boldsymbol{r}]}$ given $\Delta_i^{[r]}$ and $X_{[r]i}$ as follows

$$Z_{ij}^{[r]}|\{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(1, \frac{\pi_j f(x_{[r]i}; \theta_j)}{g(x_{[r]i}; \boldsymbol{\Psi})}\right), \tag{3.14}$$

$$W_{ij}^{[r]}|\{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(h-1, \frac{\pi_j F(x_{[r]i}; \theta_j)}{G(x_{[r]i}; \boldsymbol{\Psi})}\right), \tag{3.15}$$

$$V_{ij}^{[r]}|\{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(H-h, \frac{\pi_j \bar{F}(x_{[r]i}; \theta_j)}{\bar{G}(x_{[r]i}; \boldsymbol{\Psi})}\right), \tag{3.16}$$

all for $i = 1, \ldots, n$ and $j = 1, \ldots, M$ and $r = 1, \ldots, H$. From (3.14), (3.15) and (3.16), we compute $\tau_{i,j}^{[r]}(\boldsymbol{\Omega})$, $\beta_{i,j}^{[r]}(\boldsymbol{\Omega})$ and $\gamma_{i,j}^{[r]}(\boldsymbol{\Omega})$, the conditional expectation of the

latent variables $\boldsymbol{Z_i^{[r]}}, \boldsymbol{W_i^{[r]}}, \boldsymbol{V_i^{[r]}}$ respectively. They are given as follows

$$\tau_{i,j}^{[r]}(\boldsymbol{\Omega}) = \frac{\pi_j f(x_{[r]i}; \theta_j)}{g(x_{[r]i}; \boldsymbol{\Psi})} \tag{3.17}$$

$$\beta_{i,j}^{[r]}(\boldsymbol{\Omega}) = (h-1) \times \frac{\pi_j F(x_{[r]i}; \theta_j)}{G(x_{[r]i}; \boldsymbol{\Psi})} \tag{3.18}$$

$$\gamma_{i,j}^{[r]}(\boldsymbol{\Omega}) = (H-h) \times \frac{\pi_j \bar{F}(x_{[r]i}; \theta_j)}{\bar{G}(x_{[r]i}; \boldsymbol{\Psi})} \tag{3.19}$$

From decomposition (3.12) and from (3.17), (3.18), (3.19), the conditional expectation of the RSS-complete log-likelihood is given by

$$Q(\boldsymbol{\Omega}, \boldsymbol{\Omega}^*) = \mathbb{E}\Big(\ell(\boldsymbol{\Omega}) | \mathbf{y_{RSS}}, \boldsymbol{\Omega}^* \Big) = \mathbf{Q_1}(\boldsymbol{\alpha}, \boldsymbol{\Omega}^*) + \mathbf{Q_2}(\boldsymbol{\pi}, \boldsymbol{\Omega}^*) + \mathbf{Q_3}(\boldsymbol{\xi}, \boldsymbol{\Omega}^*), \tag{3.20}$$

where

$$Q_1(\boldsymbol{\alpha}, \boldsymbol{\Omega}^*) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^*) \log \alpha_{r,h}, \tag{3.21}$$

$$\begin{aligned} Q_2(\boldsymbol{\pi}, \boldsymbol{\Omega}^*) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \sum_{j=1}^{M} \phi_i^{[r,h]}(\boldsymbol{\Omega}^*) \ \log \pi_j \\ \times \Big\{ \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^*) + \beta_{i,j}^{[r]}(\boldsymbol{\Omega}^*) + \gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^*) \Big\}, \end{aligned} \tag{3.22}$$

$$\begin{aligned} Q_3(\boldsymbol{\xi}, \boldsymbol{\Omega}^*) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \Bigg\{ & \log f_j(x_{[r]i}; \theta_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^*) \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^*) \\ & + \log F_j(x_{[r]i}; \theta_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^*) \beta_{i,j}^{[r]}(\boldsymbol{\Omega}^*) \\ & + \log \bar{F}_j(x_{[r]i}; \theta_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^*) \gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^*) \Bigg\}. \end{aligned} \tag{3.23}$$

To better describe the iterative nature of the EM algorithm, we suppose that we have $\mathbf{\Omega^{(p)}}$ the estimate of $\mathbf{\Omega}$ updated from the p-th iteration, we shall update $\mathbf{\Omega^{(p+1)}}$ from the following E- and M-steps.

E-Step: In the E-step, using the $\mathbf{\Omega^{(p)}}$ from the p-th iteration and from (3.20), we compute the conditional expectation of the log-likelihood in the $(p+1)$ iteration as follows

$$Q(\mathbf{\Omega}, \mathbf{\Omega^{(p)}}) = \mathbb{E}_{\mathbf{\Omega^{(p)}}}\left(\ell(\mathbf{\Omega})|\mathbf{y_{RSS}}, \mathbf{\Omega^{(p)}}\right) = \mathbf{Q_1}(\boldsymbol{\alpha}, \mathbf{\Omega^{(p)}}) + \mathbf{Q_2}(\boldsymbol{\pi}, \mathbf{\Omega^{(p)}}) + \mathbf{Q_3}(\boldsymbol{\xi}, \mathbf{\Omega^{(p)}}), \tag{3.24}$$

where Q_1, Q_2, Q_3 are computed from (3.21), (3.22) and (3.23).

M-Step: In this step, we obtain the updated estimate of $\mathbf{\Omega}^{(p+1)} = (\mathbf{\Psi}^{(p+1)}, \boldsymbol{\alpha}^{(p+1)})$ by maximizing the conditional log-likelihood function $Q(\mathbf{\Omega}, \mathbf{\Omega}^{(p)})$. Based on the decomposition (3.20), we can maximize the Q_1, Q_2 and Q_3 separately. First, we maximize $Q_1(\boldsymbol{\alpha}, \mathbf{\Omega}^{(p)})$. Because of the constraint that $\boldsymbol{\alpha}$ is a doubly stochastic matrix, we use the Lagrangian multipliers to maximize Q_1 subject to the constraint $\sum_{h=1}^{H} \alpha_{r,h} = \sum_{r=1}^{H} \alpha_{r,h} = 1$. To do so, we assume $\boldsymbol{\alpha}$ is symmetric matrix where $\alpha_{h,h'} = \alpha_{h',h}$ and using method of Arslan and Ozturk (2013), the Lagrangian multiplier is given by

$$Q_1(\boldsymbol{\alpha}, \mathbf{\Omega}^{(p)}; \lambda) = \sum_{h=1}^{H} \left\{ \sum_{h'=1}^{h-1} \phi_{h,h'}(\mathbf{\Omega}^{(p)}) \log \alpha_{h,h'} + \sum_{h'=h}^{H} \phi_{h,h'}(\mathbf{\Omega}^{(p)}) \log \alpha_{h',h} \right\}$$
$$+ \sum_{h=1}^{H} \lambda_h \left\{ \sum_{h'=1}^{h-1} \alpha_{h,h'} + \sum_{h'=h}^{H} \alpha_{h',h} - 1 \right\}.$$

To update $\boldsymbol{\pi}^{(p+1)}$ we need to use (3.22) to maximize $Q_2(\boldsymbol{\pi}, \boldsymbol{\Omega}^{(p)})$ with respect to $\boldsymbol{\pi}$ subject to the constraint that $\sum_{j=1}^{M} \pi_j = 1$. It is easy to show that the update $\boldsymbol{\pi}^{(p+1)}$ is given by

$$\hat{\pi}_j^{(p+1)} = \frac{1}{n(H)^3} \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \left\{ \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) + \beta_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) + \gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) \right\}. \quad (3.25)$$

Finally, to obtain the update $\boldsymbol{\xi}^{(p+1)}$, we need to maximize the $Q_3(\boldsymbol{\xi}, \boldsymbol{\Omega}^{(p)})$ with respect to $\boldsymbol{\xi}$. Hence the $\boldsymbol{\xi}^{(p+1)}$ can be obtained as a solution to

$$\sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \frac{\frac{\partial}{\partial \xi} f_j(x_{[r]i}; \theta_j)}{f_j(x_{[r]i}; \theta_j)} \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})$$

$$+ \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \frac{\partial}{\partial \xi} F_j(x_{[r]i}; \theta_j) \left(\sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \left\{ \frac{\beta_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})}{F_j(x_{[r]i}; \theta_j)} - \frac{\gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})}{\bar{F}_j(x_{[r]i}; \theta_j)} \right\} \right) = 0. \quad (3.26)$$

One can obtain the ML estimates of $\boldsymbol{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\Psi})$ by alternating the E- and M-steps until $\|\boldsymbol{\Omega}^{(p+1)} - \boldsymbol{\Omega}^{(p)}\|_\infty$ becomes negligible. More details about the proposed EM algorithm and RSS-ML estimation of FMMs in parametric setting can be found in Hatefi et al. (2015).

3.2 Semi-parametric FMMs with RSS

This section develops a new semi-parametric EM algorithm to estimate the unknown parameters in FMMs. As we discussed in Chapter 2, in a semi-parametric setting, the probability density function (pdf) f is unknown, and we need to estimate it by

using an appropriate Kernel density estimation method. Similar to the population configuration of Chapter 2, we consider the underlying population comes from family of the location-shifted mixture models. Accordingly, the probability density function (pdf) and cumulative density function (CDF) in mixture models are given by

$$g(x;\boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j f(x-\mu_j), \tag{3.27}$$

$$G(x;\boldsymbol{\Psi}) = \sum_{j=1}^{M} \pi_j F(x-\mu_j). \tag{3.28}$$

The unknown parameters of the model are given by $\boldsymbol{\Psi} = (\boldsymbol{\pi}, f, \boldsymbol{\xi}^\top)$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ and $\boldsymbol{\xi}^\top = (\mu_1, \ldots, \mu_M)$. We assume that the random variable X follows FMM (3.27). Let $\{X_{[r]i}; r = 1, \ldots, H; i = 1, \ldots, n\}$ denote an RSS of size nH with the set size n and the number of cycles n. In a similar view to parametric setting of Section 3.1, we introduce a missing data mechanism to model the ranking error involved in RSS data. Let $\boldsymbol{\alpha}$ show the misplacement probability in the model. The misplacement probability matrix is given by

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \ldots & \alpha_{1,H} \\ \alpha_{2,1} & \alpha_{2,2} & \ldots & \alpha_{2,H} \\ \vdots & \vdots & \ldots & \vdots \\ \alpha_{H,1} & \alpha_{H,2} & \ldots & \alpha_{H,H} \end{bmatrix}_{H\times H},$$

where $\alpha_{r,h}$ is the probability that h-th order statistic is assigned to the r-th judgmental rank stratum in the set. Similarly, $\boldsymbol{\alpha}$ should be a doubly stochastic such that $\sum_{h=1}^{H} \alpha_{r,h} = \sum_{r=1}^{H} \alpha_{r,h} = 1$. To accommodate the ranking errors in our estimation

procedure, for each $X_{[r]i}$, we define a H dimensional multinomial random vector, $\boldsymbol{\Delta}_i^{[r]}$, with parameters 1 and $\boldsymbol{\alpha}_{[r]} = (\alpha_{r,1}, \ldots, \alpha_{r,H})$ such that

$$\boldsymbol{\Delta}_i^{[r]} = \left(\Delta_i^{[r,1]}, \ldots, \Delta_i^{[r,H]}\right).$$

The pdf of $\boldsymbol{\Delta}_i^{[r]}$ is given by

$$\mathbb{P}(\boldsymbol{\Delta}_i^{[r]} = \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\alpha}) = \prod_{h=1}^{H} (\alpha_{r,h})^{\delta_i^{[r,h]}}. \tag{3.29}$$

We can write the conditional distribution of $X_{[r]i}$ given $\Delta_i^{[r]}$ as follows

$$f(x_{[r]i}|\boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}) = \prod_{\mathbf{h=1}}^{\mathbf{H}} \left\{ \mathbf{g^{(h:H)}}(\mathbf{x_{[r]i}}; \boldsymbol{\Psi}) \right\}^{\delta_{\mathbf{i}}^{\mathbf{[r,h]}}}, \tag{3.30}$$

where

$$g^{(h:H)}(x; \boldsymbol{\Psi}) = H \binom{H-1}{h-1} g(x; \boldsymbol{\Psi}) \{G(x; \boldsymbol{\Psi})\}^{h-1} \{\bar{G}(x; \boldsymbol{\Psi})\}^{H-h}, \tag{3.31}$$

and $\bar{G}(x; \boldsymbol{\Psi}) = 1 - G(x; \boldsymbol{\Psi})$. From (3.29) and (3.30), the joint distribution of $(X_{[r]i}, \boldsymbol{\Delta}_i^{[r]})$ is given by

$$f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}) = \prod_{h=1}^{H} \left\{ \alpha_{r,h} g^{(h:H)}(x_{[r]i}; \boldsymbol{\Psi}) \right\}^{\delta_i^{[r,h]}}, \tag{3.32}$$

where $\boldsymbol{\Omega} = (\boldsymbol{\Psi}, \boldsymbol{\alpha})$ and $\boldsymbol{\Psi} = (\boldsymbol{\pi}, f, \boldsymbol{\xi}^\top)$. Therefore, the marginal distribution of $X_{[r]i}$ can be computed as

$$f(x_{[r]i}; \boldsymbol{\Omega}) = \sum_{\boldsymbol{\delta}_i^{[r]}} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}) = \sum_{h=1}^{H} \alpha_{r,h} g^{(h:H)}(x_{[r]i}; \boldsymbol{\Psi}). \tag{3.33}$$

We can write the conditional distribution of $\Delta_i^{[r]}$ given $X_{[r]i}$ from (3.32) and (3.33) as follows

$$f(\boldsymbol{\delta}_i^{[r]}|x_{[r]i};\boldsymbol{\Omega}) = \prod_{h=1}^{H}\left\{\frac{\alpha_{r,h}B_{h,H+1-h}(G(x_{[r]i};\boldsymbol{\Psi}))}{\sum_{h=1}^{H}\alpha_{r,h}B_{h,H+1-h}(G(x_{[r]i};\boldsymbol{\Psi}))}\right\}^{\delta_i^{[r,h]}}, \tag{3.34}$$

where $B_{a,b}(\cdot)$ is a beta density function with parameters a and b.

3.2.1 Likelihood Functions with RSS

In this section, we estimate the unknown parameter $\boldsymbol{\Omega}$ of the location-shifted FMM (3.27) by using ML method. Using the equation (3.33), we write the likelihood function as

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^{n}\prod_{r=1}^{H} f(x_{[r]i};\boldsymbol{\Omega}). \tag{3.35}$$

Hence, the ML estimate of $\boldsymbol{\Omega}$ is the solution to

$$\widehat{\boldsymbol{\Omega}}_{ML} = \underset{\boldsymbol{\Omega}}{\operatorname{argmax}}\ L(\boldsymbol{\Omega}). \tag{3.36}$$

As saw in previous chapter, it is not tractable to obtain the ML estimate of the parameters via (3.35). Thus, we use the latent multinomial random vectors $\boldsymbol{\Delta}_i^{[r]}$ and we derive the likelihood function of $\boldsymbol{\Omega}$ as follows

$$L(\boldsymbol{\Omega}) = \prod_{i=1}^{n}\prod_{r=1}^{H} f(x_{[r]i},\delta_i^{[r]};\boldsymbol{\Omega}). \tag{3.37}$$

The likelihood function (3.37) is tractable with respect to $\boldsymbol{\alpha}$; hence the ML estimator of $\boldsymbol{\alpha}$, can be easily obtained by using the EM algorithm. Based on the equations

(3.27), (3.28) and (3.31), the likelihood function (3.37) is stil not tractable to obtain the ML estimation of $\boldsymbol{\Psi}$. In a similar view to previous chapter, we introduce three additional latent variables $\mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}, \mathbf{V}_i^{[r]}$ for each $x_{[r]i}$ given $\boldsymbol{\Delta}_i^{[r]} = \boldsymbol{\delta}_i^{[r]}$. We consider $\{\Delta_i^{[r,h]} = 1\}$ as the event that the h-th entry of the vector $\boldsymbol{\Delta}_i^{[r]}$ is one and other elements are zero. We define an M-dimensional latent vector $\mathbf{Z}_i^{[r]}|\{\Delta_i^{[r,h]} = 1\}$, where $Z_{ij}^{[r]}|\{\Delta_i^{[r,h]} = 1\}$ is one if $x_{[r]i}$ belongs to the j-th component of the mixture model; that is,

$$Z_{ij}^{[r]}\big|\{\Delta_i^{[r,h]} = 1\} = \begin{cases} 1 & \text{if } x_{[r]i} \text{ belongs to component } j; \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\sum_{j=1}^{M} \left(Z_{ij}^{[r]}|\Delta_i^{[r,h]} = 1 \right) = 1.$$

Since $\mathbf{Z}_i^{[r]}|\{\Delta_i^{[r,h]} = 1\}$, $i = 1, \ldots, n; r = 1, \ldots, H$, has a multinomial distribution including one draw on M classes with probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$, we can write

$$f(\mathbf{z}_i^{[r]}\big|\boldsymbol{\delta}_i^{[r]}; \boldsymbol{\pi}) = \prod_{h=1}^{H} \left\{ \binom{1}{z_{i1}^{[r]}, \ldots, z_{iM}^{[r]}} \prod_{j=1}^{M} \pi_j^{z_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}. \tag{3.38}$$

We also introduce $\mathbf{W}_i^{[r]}|\{\Delta_i^{[r,h]} = 1\}$ as an M-dimensional vector, where $W_{ij}^{[r]}|\{\Delta_i^{[r,h]} = 1\}$ denotes the number of observations less than $x_{[r]i}$ which are selected from the component j of the FMM (3.27) with

$$\sum_{j=1}^{M} \left(W_{ij}^{[r]}|\Delta_i^{[r,h]} = 1 \right) = h - 1.$$

The latent vectors $\mathbf{W}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}, i=1,\ldots,n; r=1,\ldots,H$, follows a multinomial distribution including $h-1$ draws on M classes with probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_M)$ and we can write

$$f(\mathbf{w}_i^{[r]}|\boldsymbol{\delta}_i^{[r]};\boldsymbol{\pi})=\prod_{h=1}^{H}\left\{\binom{h-1}{w_{i1}^{[r]},\ldots,w_{iM}^{[r]}}\prod_{j=1}^{M}\pi_j^{w_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}. \tag{3.39}$$

Finally, we introduce $\mathbf{V}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}$ as an M-dimensional vector, where $V_{ij}^{[r]}|\{\Delta_i^{[r,h]}=1\}$ shows the number of observations bigger than $x_{[r]i}$ that are selected from the component j of the FMM (3.27) with

$$\sum_{j=1}^{M}\left(V_{ij}^{[r]}|\Delta_i^{[r,h]}=1\right)=H-h.$$

Accordingly, the vectors $\mathbf{V}_i^{[r]}|\{\Delta_i^{[r,h]}=1\}, i=1,\ldots,n; r=1,\ldots,H$, also has a multinomial distribution including $H-h$ draws on M classes with probabilities $\boldsymbol{\pi}=(\pi_1,\ldots,\pi_M)$, and we have

$$f(\mathbf{v}_i^{[r]}|\boldsymbol{\delta}_i^{[r]};\boldsymbol{\pi})=\prod_{h=1}^{H}\left\{\binom{H-h}{v_{i1}^{[r]},\ldots,v_{iM}^{[r]}}\prod_{j=1}^{M}\pi_j^{v_{ij}^{[r]}}\right\}^{\delta_i^{[r,h]}}. \tag{3.40}$$

The latent variables $\mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}$ and $\mathbf{V}_i^{[r]}$ given $\boldsymbol{\Delta}_i^{[r]}$ are conditionally independent. We derive the joint distribution of the judgment order statistic $X_{[r]i}$ and the latent variables $\mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}$ and $\mathbf{V}_i^{[r]}$ in the following lemma.

Lemma 3.2. *For $i = 1, \ldots, n$, $r = 1, \ldots, H$, we have*

$$f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega}) \propto \prod_{h=1}^{H} \prod_{j=1}^{M} \left\{ \alpha_{r,h} \pi_j^{\{z_{ij}^{[r]} + w_{ij}^{[r]} + v_{ij}^{[r]}\}} \right\}^{\delta_i^{[r,h]}}$$

$$\times \left\{ [f(x_{[r]i} - \mu_j)]^{z_{ij}^{[r]}} [F(x_{[r]i} - \mu_j)]^{w_{ij}^{[r]}} [\bar{F}(x_{[r]i} - \mu_j)]^{v_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}.$$

Proof. Let $c_1 = H\binom{H-1}{h-1}$, $c_2 = \binom{1}{z_{i1}^{[r]}, \ldots, z_{iM}^{[r]}}$, $c_3 = \binom{h-1}{w_{i1}^{[r]}, \ldots, w_{iM}^{[r]}}$ and $c_4 = \binom{H-h}{v_{i1}^{[r]}, \ldots, v_{iM}^{[r]}}$.

$$f(x_{[r]i} | \{\mathbf{z}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ c_1 \prod_{j=1}^{M} [f(x_{[r]i} - \mu_j)]^{z_{ij}^{[r]}} [G(x_{[r]i}; \boldsymbol{\Psi})]^{h-1} [\bar{G}(x_{[r]i}; \boldsymbol{\Psi})]^{H-h} \right\}^{\delta_i^{[r,h]}}. \tag{3.41}$$

Using (3.29), (3.38) and (3.41) we can derive the joint distribution of $(X_{[r]i}, \boldsymbol{Z}_{\boldsymbol{i}}^{[\boldsymbol{r}]}, \Delta_i^{[r]})$ as

$$f(x_{[r]i}, \mathbf{z}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ \alpha_{r,h}\, c_1\, c_2 \prod_{j=1}^{M} [\pi_j f(x_{[r]i} - \mu_j)]^{z_{ij}^{[r]}} [G(x_{[r]i}; \boldsymbol{\Psi})]^{h-1} [\bar{G}(x_{[r]i}; \boldsymbol{\Psi})]^{H-h} \right\}^{\delta_i^{[r,h]}}. \tag{3.42}$$

Furthermore, using (3.32) and (3.42), it is easy to see that

$$f(\mathbf{z}_i^{[r]} | \{x_{[r]i}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega}) = \prod_{h=1}^{H} \left\{ c_2 \prod_{j=1}^{M} \left(\frac{\pi_j f(x_{[r]i} - \mu_j)}{g(x_{[r]i}; \boldsymbol{\Psi})} \right)^{z_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}. \tag{3.43}$$

On the other hand, the conditional distribution of $X_{[r]i}$ given $\mathbf{W}_i^{[r]}$ and $\boldsymbol{\Delta}_i^{[r]}$ can be

written as

$$f(x_{[r]i}|\{\mathbf{w}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ c_1 \, g(x_{[r]i}; \boldsymbol{\Psi}) \prod_{j=1}^{M} [F(x_{[r]i} - \mu_j)]^{w_{ij}^{[r]}} [\bar{G}(x_{[r]i}; \boldsymbol{\Psi})]^{H-h} \right\}^{\delta_i^{[r,h]}}.$$

From (3.29) and (3.39), it is easy to obtain

$$f(x_{[r]i}, \mathbf{w}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ \alpha_{r,h} \, c_1 \, c_3 g(x_{[r]i}; \boldsymbol{\Psi}) \prod_{j=1}^{M} [\pi_j F(x_{[r]i} - \mu_j)]^{w_{ij}^{[r]}} [\bar{G}(x_{[r]i}; \boldsymbol{\Psi})]^{H-h} \right\}^{\delta_i^{[r,h]}}.$$

Now, using (3.32), we can derive

$$f(\mathbf{w}_i^{[r]}|\{x_{[r]i}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega}) = \prod_{h=1}^{H} \left\{ c_3 \prod_{j=1}^{M} \left(\frac{\pi_j F(x_{[r]i} - \mu_j)}{G(x_{[r]i}; \boldsymbol{\Psi})} \right)^{w_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}. \tag{3.44}$$

Similarly, we can derive

$$f(x_{[r]i}|\{\mathbf{v}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ c_1 g(x_{[r]i}; \boldsymbol{\Psi}) \prod_{j=1}^{M} [G(x_{[r]i}; \boldsymbol{\Psi})]^{h-1} [\bar{F}(x_{[r]i} - \mu_j)]^{v_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}.$$

Again, from (3.29) and (3.40) the joint distribution is given by

$$f(x_{[r]i}, \mathbf{v}_i^{[r]}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \left\{ \alpha_{r,h} \, c_1 \, c_4 g(x_{[r]i}; \boldsymbol{\Psi}) \prod_{j=1}^{M} [G(x_{[r]i}; \boldsymbol{\Psi})]^{h-1} [\pi_j \bar{F}(x_{[r]i} - \mu_j)]^{v_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}}.$$

Once again, using (3.32), it is easy to show

$$f(\mathbf{v}_i^{[r]}|\{x_{[r]i}, \boldsymbol{\delta}_i^{[r]}\}; \boldsymbol{\Omega}) = \prod_{h=1}^{H} \left\{ c_4 \prod_{j=1}^{M} \left(\frac{\pi_j \bar{F}(x_{[r]i} - \mu_j)}{\bar{G}(x_{[r]i}; \boldsymbol{\Psi})} \right)^{v_{ij}^{[r]}} \right\}^{\delta_i^{[r,h]}} . \tag{3.45}$$

Now based on the conditional independence of the latent variables, we have

$$f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}) = f(\mathbf{z}_i^{[r]}|x_{[r]i}, \boldsymbol{\delta}_i^{[r]}).f(\mathbf{w}_i^{[r]}|x_{[r]i}, \boldsymbol{\delta}_i^{[r]}).f(\mathbf{v}_i^{[r]}|x_{[r]i}, \boldsymbol{\delta}_i^{[r]}).f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega})$$

Now from (3.32), (3.43), (3.44) and (3.45), the proof of lemma is completed. □

In the following Lemma, we show the marginalization of RSS-complete likelihood function.

Lemma 3.3. *For each* $x_{[r]i}$, $i = 1, \ldots, n; r = 1, \ldots, H$, *we have*

$$f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}) = \sum_{\mathbf{Z}|\boldsymbol{\delta}} \sum_{\mathbf{W}|\boldsymbol{\delta}} \sum_{\mathbf{V}|\boldsymbol{\delta}} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega}).$$

Proof. Let $c_1 = H\binom{H-1}{h-1}$, $c_2 = \binom{1}{z_{i1}^{[r]}, \ldots, z_{iM}^{[r]}}$, $c_3 = \binom{h-1}{w_{i1}^{[r]}, \ldots, w_{iM}^{[r]}}$ and $c_4 = \binom{H-h}{v_{i1}^{[r]}, \ldots, v_{iM}^{[r]}}$.

$$\sum_{\mathbf{Z}|\boldsymbol{\delta}}\sum_{\mathbf{W}|\boldsymbol{\delta}}\sum_{\mathbf{V}|\boldsymbol{\delta}} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega})$$

$$= \prod_{h=1}^{H} \Bigg\{ c_1 \alpha_{r,h} \left[\sum_{\mathbf{Z}_i^{[r]}|\boldsymbol{\delta}_i^{[r]}} c_2 \prod_{j=1}^{M} \{\pi_j f(x_{[r]i} - \mu_j)\}^{z_{ij}^{[r]}} \right]$$

$$\times \left[\sum_{\mathbf{W}_i^{[r]}|\boldsymbol{\delta}_i^{[r]}} c_3 \prod_{j=1}^{M} \{\pi_j F(x_{[r]i} - \mu_j)\}^{w_{ij}^{[r]}} \right] \left[\sum_{\mathbf{V}_i^{[r]}|\boldsymbol{\delta}_i^{[r]}} c_4 \prod_{j=1}^{M} \{\pi_j \bar{F}(x_{[r]i} - \mu_j)\}^{v_{ij}^{[r]}} \right] \Bigg\}^{\delta_i^{[r,h]}}$$

$$= \prod_{h=1}^{H} \Bigg\{ c_1 \alpha_{r,h} \left[\sum_{j=1}^{M} \pi_j f(x_{[r]i} - \mu_j) \right] \left[\sum_{j=1}^{M} \pi_j F(x_{[r]i} - \mu_j) \right]^{h-1}$$

$$\times \left[\sum_{j=1}^{M} \pi_j \bar{F}(x_{[r]i} - \mu_j) \right]^{H-h} \Bigg\}^{\delta_i^{[r,h]}}$$

$$= \prod_{h=1}^{H} \left\{ c_1 \alpha_{r,h} g(x_{[r]i}; \boldsymbol{\Psi}) [G(x_{[r]i}; \boldsymbol{\Psi})]^{h-1} [\bar{G}(x_{[r]i}; \boldsymbol{\Psi})]^{H-h} \right\}^{\delta_i^{[r,h]}}$$

$$= f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}; \boldsymbol{\Omega}).$$

□

Let $\mathbf{y_{RSS}} = \{(X_{[r]i}, \boldsymbol{\Delta}_i^{[r]}, \mathbf{Z}_i^{[r]}, \mathbf{W}_i^{[r]}, \mathbf{V}_i^{[r]}), i = 1, \ldots, n; r = 1, \ldots, H\}$ denote the complete RSS data; then the RSS-complte likelihood function is given by

$$L(\boldsymbol{\Omega}|\mathbf{y_{RSS}}) = \prod_{i=1}^{n} \prod_{r=1}^{H} f(x_{[r]i}, \boldsymbol{\delta}_i^{[r]}, \mathbf{z}_i^{[r]}, \mathbf{w}_i^{[r]}, \mathbf{v}_i^{[r]}; \boldsymbol{\Omega}), \tag{3.46}$$

The complete-data log-likelihood function of $\mathbf{\Omega}$ is also given by

$$
\begin{aligned}
& l(\mathbf{\Omega}|\mathbf{y_{RSS}}) \\
\propto & \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\log\alpha_{r,h} \\
& + \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\sum_{j=1}^{M}\left[\left\{z_{ij}^{[r]}+w_{ij}^{[r]}+v_{ij}^{[r]}\right\}\log\pi_j\right] \\
& + \sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{h=1}^{H}\delta_i^{[r,h]}\times\sum_{j=1}^{M}\left[z_{ij}^{[r]}\log f(x_{[r]i}-\mu_j)\right. \\
& \qquad\qquad \left. + w_{ij}^{[r]}\log F(x_{[r]i}-\mu_j) + v_{ij}^{[r]}\log\bar{F}(x_{[r]i}-\mu_j)\right] \\
= & l_1(\boldsymbol{\alpha}|\mathbf{y_{RSS}}) + l_2(\boldsymbol{\pi}|\mathbf{y_{RSS}}) + l_3(\boldsymbol{\mu}, f|\mathbf{y_{RSS}}). \qquad (3.47)
\end{aligned}
$$

3.2.2 Semi-parametric EM Algorithm

In this section, we develop a semi-parametric EM algorithm to obtain the ML estimates of FMMs based on RSS data. As discussed before, the EM algorithm requires an initial value.

Initialization Step: Let $\mathbf{\Omega}^{(0)} = (\boldsymbol{\alpha}^{(0)}, f^{(0)}, \boldsymbol{\xi}^{\top(0)})$ represent the starting value of the EM algorithm. In the E-step of the EM algorithm, we need to compute the conditional expectation of the latent variables given incomplete data. For the first layer of the latent variables, from (3.34), the conditional distribution of $\Delta_i^{[r,h]}|X_{[r]i}$ is given by

$$
\Delta_i^{[r,h]}|\{X_{[r]i}=x_{[r]i}\} \sim \text{Bin}\left(1, \phi_i^{[r,h]}(\mathbf{\Omega})\right),
$$

where $\text{Bin}(a, b)$ is a binomial distribution with parameters a and b, and the probability of success is given by

$$\phi_i^{[r,h]}(\mathbf{\Omega}) = \frac{\alpha_{r,h} B_{h,H-h+1}(G(x_{[r]i}; \mathbf{\Psi}))}{\sum_{h=1}^{H} \alpha_{r,h} B_{h,H-h+1}(G(x_{[r]i}; \mathbf{\Psi}))},$$

and $B_{a,b}(\cdot)$ is a Beta density function with parameters a and b. For the conditional expectation of the latent variables $\mathbf{Z}_i$, $\mathbf{W}_i$, and $\mathbf{V}_i$, we calculate their conditional distributions given $X_{[r]i}$ and $\Delta_i^{[r,h]}$. Using (3.43), (3.44) and (3.45), we can find the conditional distribution of the latent variables as follows

$$Z_{ij}^{[r]} | \{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(1, \frac{\pi_j f(x_{[r]i} - \mu_j)}{g(x_{[r]i}; \mathbf{\Psi})}\right),$$

$$W_{ij}^{[r]} | \{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(h - 1, \frac{\pi_j F(x_{[r]i} - \mu_j)}{G(x_{[r]i}; \mathbf{\Psi})}\right),$$

$$V_{ij}^{[r]} | \{x_{[r]i}, \Delta_i^{[r,h]} = 1\} \sim \text{Bin}\left(H - h, \frac{\pi_j \bar{F}(x_{[r]i} - \mu_j)}{\bar{G}(x_{[r]i}; \mathbf{\Psi})}\right),$$

Let $\tau_{i,j}^{[r]}(\mathbf{\Omega})$, $\beta_{i,j}^{[r]}(\mathbf{\Omega})$ and $\gamma_{i,j}^{[r]}(\mathbf{\Omega})$ denote the conditional expectations of $Z_{ij}^{[r]}$, $W_{ij}^{[r]}$ and $V_{ij}^{[r]}$, given $x_{[r]i}$ and $\Delta_i^{[r,h]} = 1$, respectively. From conditional independence of the latent variables, we have

$$\mathbb{E}_{\mathbf{\Omega}^{(p)}}\left[\Delta_i^{[r,h]} Z_{ij}^{[r]} \middle| x_{[r]i}\right] = \phi_i^{[r,h]}(\mathbf{\Omega})\, \tau_{i,j}^{[r]}(\mathbf{\Omega}),$$

$$\mathbb{E}_{\mathbf{\Omega}^{(p)}}\left[\Delta_i^{[r,h]} W_{ij}^{[r]} \middle| x_{[r]i}\right] = \phi_i^{[r,h]}(\mathbf{\Omega})\, \beta_{i,j}^{[r]}(\mathbf{\Omega}),$$

$$\mathbb{E}_{\mathbf{\Omega}^{(p)}}\left[\Delta_i^{[r,h]} V_{ij}^{[r]} \middle| x_{[r]i}\right] = \phi_i^{[r,h]}(\mathbf{\Omega})\, \gamma_{i,j}^{[r]}(\mathbf{\Omega}).$$

E-Step: In this step, we consider $\mathbf{\Omega}^{(p)}$ is the update of $\mathbf{\Omega}$ from the p-th iteration. Using $\mathbf{\Omega}^{(p)}$, the conditional expectation of log-likelihood based on complete RSS data is given by

$$\begin{aligned} Q(\mathbf{\Omega}, \mathbf{\Omega}^{(p)}) &= \mathbb{E}_{\mathbf{\Omega}^{(p)}}\left(l(\mathbf{\Omega}) | \boldsymbol{y_{RSS}} \right) \\ &= Q_1(\boldsymbol{\alpha}, \mathbf{\Omega}^{(p)}) + Q_2(\boldsymbol{\pi}, \mathbf{\Omega}^{(p)}) + Q_3(\boldsymbol{\xi}, f, \mathbf{\Omega}^{(p)}), \end{aligned} \tag{3.48}$$

where

$$Q_1(\boldsymbol{\alpha}, \mathbf{\Omega}^{(p)}) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \phi_i^{[r,h]}(\mathbf{\Omega}^{(p)}) \log(\alpha_{r,h}),$$

$$\begin{aligned} Q_2(\boldsymbol{\pi}, \mathbf{\Omega}^{(p)}) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \sum_{j=1}^{M} \phi_i^{[r,h]}(\mathbf{\Omega}^{(p)})\ \log \pi_j^{(p)} \\ \times \left\{ \tau_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) + \beta_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) + \gamma_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) \right\}, \end{aligned}$$

and

$$\begin{aligned} Q_3(\boldsymbol{\mu}, f, \mathbf{\Omega}^{(p)}) = \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \Bigg\{ & \log f^{(p)}(x_{[r]i} - \mu_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\mathbf{\Omega}^{(p)}) \tau_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) \\ & + \log F^{(p)}(x_{[r]i} - \mu_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\mathbf{\Omega}^{(p)}) \beta_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) \\ & + \log \bar{F}^{(p)}(x_{[r]i} - \mu_j) \sum_{h=1}^{H} \phi_i^{[r,h]}(\mathbf{\Omega}^{(p)}) \gamma_{i,j}^{[r]}(\mathbf{\Omega}^{(p)}) \Bigg\}. \end{aligned}$$

S-Step: We need to update the pdf $f^{(p+1)}$. In this step, in a similar view to Section

2.3, the estimate of f based on RSS data can be updated in a semi-parametric framework as follows:

S1-Step: We can simulate the stochastic version of the $\tau_{i,j}^{[r]}(\boldsymbol{\Psi})$ through

$$\boldsymbol{Z}^{(p+1)}(x_{[r]i};\boldsymbol{\Psi}^{(p)}) \sim Multi\Big\{1, \tau_{i1}(x_{[r]i};\boldsymbol{\Psi}^{(p)}), \ldots, \tau_{iM}(x_{[r]i};\boldsymbol{\Psi}^{(p)})\Big\}, r = 1, \ldots, H.$$

S2-Step: We can obtain the centered back RSS observations from f through

$$\tilde{x}_{[r]i}^{(p+1)} = x_{[r]i} - \mu_{Z^{(p+1)}(x_{[r]i};\boldsymbol{\Psi}^{(p)})}^{(p)},$$

for $i = 1, \ldots, n$. Due to the fact the f is assumed to be symmetric, we use centered back RSS data $\left\{\tilde{x}_{[1]1}^{(p+1)}, \ldots, \tilde{x}_{[H]n}^{(p+1)}\right\}$, and obtain the Kernel density estimate $f^{(p+1)}$ of f as follows:

S3-Step: Kernel density estimate of f

$$\hat{f}_{\tilde{x}_{[r]i}^{(p+1)}}(u) = \frac{1}{Hh_n}\sum_{r=1}^{H} K\left(\frac{u - \tilde{x}_{[r]i}^{(p+1)}}{h_n}\right); \tag{3.49}$$

S4-Step: Symmetrization

$$f^{(p+1)}(u) = \frac{\hat{f}_{\tilde{x}_{[r]i}^{(p+1)}}(u) + \hat{f}_{\tilde{x}_{[r]i}^{(p+1)}}(-u)}{2}. \tag{3.50}$$

Once we obtained $f^{(p+1)}$, we use the empirical cumulative density based on $f^{(p+1)}$ to update the estimate of $F^{(p+1)}$.

M-Step: In this step, we focus on the maximization of (3.48) to update the estimates of $\boldsymbol{\Omega}^{(p+1)} = (\boldsymbol{\Psi}^{(p+1)}, \boldsymbol{\alpha}^{(p+1)})$. First, we maximize $Q_1(\boldsymbol{\alpha}, \boldsymbol{\Omega}^{(p)})$ under the

constraint that $\boldsymbol{\alpha}$ is a double stochastic matrix. We can use double multipliers to take into account the constraint in the maximization. Under the assumption that the $\boldsymbol{\alpha}$ is symmetric (Arslan and Ozturk, 2013), the double Lagrangian multiplier is given by

$$Q_1(\boldsymbol{\alpha}, \boldsymbol{\Omega}^{(p)}; \lambda) = \sum_{h=1}^{H} \left\{ \sum_{h'=1}^{h-1} \phi_{h,h'}(\boldsymbol{\Omega}^{(p)}) \log \alpha_{h,h'} + \sum_{h'=h}^{H} \phi_{h,h'}(\boldsymbol{\Omega}^{(p)}) \log \alpha_{h',h} \right\}$$

$$+ \sum_{h=1}^{H} \lambda_h \left\{ \sum_{h'=1}^{h-1} \alpha_{h,h'} + \sum_{h'=h}^{H} \alpha_{h',h} - 1 \right\}.$$

We can update the estimation of the π_j by maximizing $Q_2(\boldsymbol{\pi}, \boldsymbol{\Omega}^{(p)})$. It is easy to show that $\pi_j^{(p+1)}$, $j = 1, \ldots, M$ can be obtained by

$$\hat{\pi}_j^{(p+1)} = \frac{1}{n(H)^3} \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \left\{ \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) + \beta_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) + \gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)}) \right\}. \quad (3.51)$$

Finally, to obtain the updated estimate $\boldsymbol{\xi}^{\top(p+1)} = (\mu_1^{(p+1)}, \ldots, \mu_M^{(p+1)})$, we maximize the $Q_3(\boldsymbol{\xi}, f, \boldsymbol{\Omega}^{(p)})$ with respect to $\boldsymbol{\xi}$. Consequently, the $\boldsymbol{\xi}^{(p+1)}$ is solution to the following equation,

$$\sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \frac{\frac{\partial}{\partial \mu} f^{(p+1)}(x_{[r]i} - \mu_j^{(p)})}{f^{(p+1)}(x_{[r]i} - \mu_j^{(p)})} \sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \tau_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})$$

$$+ \sum_{i=1}^{n} \sum_{r=1}^{H} \sum_{j=1}^{M} \frac{\partial}{\partial \mu} F^{(p+1)}(x_{[r]i} - \mu_j^{(p)})$$

$$\times \left(\sum_{h=1}^{H} \phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)}) \left\{ \frac{\beta_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})}{F^{(p+1)}(x_{[r]i} - \mu_j^{(p)})} - \frac{\gamma_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})}{\bar{F}^{(p+1)}(x_{[r]i} - \mu_j^{(p)})} \right\} \right) = 0. \quad (3.52)$$

We obtain the ML estimates of $\boldsymbol{\Omega}$ by alternating the E-step, S-step, and M-steps until $\|\boldsymbol{\Omega}^{(p+1)} - \boldsymbol{\Omega}^{(p)}\|_\infty$ becomes negligible.

3.2.3 Modified EM Algorithm

Based on the proposed EM algorithm, the $(p+1)$-th step estimator of $\boldsymbol{\xi}$ requires a solution to the estimating equation (3.52) with respect to μ_j, $j = 1, \ldots, M$. As we discussed in the S-step of the proposed semi-parametric EM- algorithm, there will be no closed form for the update of pdf f and CDF in the numerator and denominator of (3.52), respectively. Hence, the optimization of (3.52) is computationally infeasible. To overcome the problem, we propose a modified EM algorithm using the technique of Johnson et al. (1972) and Mehrotra and Nanda (1974) that replace the hazard rate functions in the log-likelihood function $\frac{\partial}{\partial \xi} \log F(x-\mu_j)$ and $\frac{\partial}{\partial \xi} \log(1-F(x-\mu_j))$ with their expected values. Although the following Lemma and its proof in a parametric framework can be found in Hatefi et al. (2015); we present the Lemma and its proof in the case of location-shifted mixture models for the sake of completeness.

Lemma 3.4. *Let $(X_{[1]i}, \ldots, X_{[H]i})$ be an RSS data of size H (from the cycle i) from FMM (3.33). Suppose $W_{ij}^{[r]}$ and $V_{ij}^{[r]}$ be, respectively, the j-th elements of the latent variables $W_i^{[r]}$ and $V_i^{[r]}$ associated with $X_{[r]i}$. Then, for any function $S(\cdot)$ (subject to the finiteness of the expectations) we have*

$$a) \quad \sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M} \mathbb{E}\left(\Delta_i^{[r,h]} W_{ij}^{[r]} S(X_{[r]i})\right) = c\sum_{j=1}^{M} \pi_j \mathbb{E}\left[G(X)F(X-\mu_j)\right],$$

$$b)\quad \sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\mathbb{E}\left(\Delta_i^{[r,h]}V_{ij}^{[r]}S(X_{[r]i})\right)=c\sum_{j=1}^{M}\pi_j\mathbb{E}\left[S(X)\bar{F}(X-\mu_j)\right],$$

where $c=H(H-1)$ *and the expectations on the right sides are with respect to the FMM* (3.27).

Proof. Here we present the proof of (a), the proof of (b) can be completed in a similar form.

$$\sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\mathbb{E}\left(\Delta_i^{[r,h]}W_{ij}^{[r]}S(X_{[r]i})\right)$$

$$=\sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\mathbb{E}\left(S(X_{[r]i})\frac{(h-1)\pi_jF(X_{[r]i}-\mu_j)}{G(X_{[r]i};\mathbf{\Psi})}\frac{\alpha_{r,h}g^{(h:H)}(X_{[r]i};\mathbf{\Psi})}{f(X_{[r]i};\mathbf{\Psi})}\right)$$

$$=\sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\int S(x)\frac{(h-1)\pi_jF(x-\mu_j)}{G(x;\mathbf{\Psi})}\alpha_{r,h}g^{(h:H)}(x;\mathbf{\Psi})dx$$

$$=\sum_{r^*=1}^{H}\sum_{h^*=1}^{H}\sum_{j=1}^{M}\int S(x)\frac{(h^*-1)\pi_jF(x-\mu_j)}{G(x;\mathbf{\Psi})}\alpha_{r^*,h^*}g^{(h^*:H)}(x;\mathbf{\Psi})dx$$

$$=\sum_{h^*=1}^{H}\sum_{j=1}^{M}\int S(x)\frac{(h^*-1)\pi_jF(x-\mu_j)}{G(x;\mathbf{\Psi})}g^{(h^*:H)}(x;\mathbf{\Psi})dx$$

$$=H\sum_{j=1}^{M}\pi_j\int S(x)F(x-\mu_j)g(x;\mathbf{\Psi})\left(\sum_{h^*=1}^{H}(h^*-1)\binom{H-1}{h^*-1}[G(x;\mathbf{\Psi})]^{h^*-2}[1-G(x;\mathbf{\Psi})]^{H-h^*}\right)dx$$

$$=H(H-1)\sum_{j=1}^{M}\pi_j\mathbb{E}\left[S(X)F_j(X-\mu_j)\right],$$

where

$$\sum_{h^*=1}^{H}(h^*-1)\binom{H-1}{h^*-1}[G(x;\boldsymbol{\Psi})]^{h^*-2}[1-G(x;\boldsymbol{\Psi})]^{H-h^*}=H-1.$$

□

Using Lemma 3.4 and considering $S_1(x_{[r]i})=\frac{\partial}{\partial\mu_{\mathbf{j}}}F(x_{[r]i}-\mu_j)/F(x_{[r]i}-\mu_j)$, we can easily write

$$\sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\mathbb{E}\left(\Delta_i^{[r,h]}W_{ij}^{[r]}S_1(X_{[r]i})\right)=H(H-1)\sum_{j=1}^{M}\pi_j\mathbb{E}\left[\frac{\partial}{\partial\mu_{\mathbf{j}}}F(X-\mu_j)\right].$$

We also consider $S_2(x_{[r]i})=\frac{\boldsymbol{\partial}}{\boldsymbol{\partial}\mu_j}F(x_{[r]i}-\mu_j)/\bar{F}(x_{[r]i}-\mu_j)$ in Lemma 3.4, then we have

$$\sum_{r=1}^{H}\sum_{h=1}^{H}\sum_{j=1}^{M}\mathbb{E}\left(\Delta_i^{[r,h]}V_{ij}^{[r]}S_2(X_{[r]i})\right)=H(H-1)\sum_{j=1}^{M}\pi_j\mathbb{E}\left[\frac{\partial}{\partial\mu_{\mathbf{j}}}F(X-\mu_j)\right].$$

Now, using (3.52) and Lemma 3.4, we develop the following modified estimating equation to update $\boldsymbol{\xi}^{\top(p+1)}=(\mu_1^{(p+1)},\ldots,\mu_M^{(p+1)})$ in the M-step of the EM algorithm which leads to approximate ML estimates of $\boldsymbol{\xi}$;

$$\sum_{i=1}^{n}\sum_{r=1}^{H}\sum_{j=1}^{M}\frac{\frac{\partial}{\partial\mu}f^{(p+1)}(x_{[r]i}-\mu_j^{(p)})}{f^{(p+1)}(x_{[r]i}-\mu_j^{(p)})}\left\{\sum_{h=1}^{H}\phi_i^{[r,h]}(\boldsymbol{\Omega}^{(p)})\tau_{i,j}^{[r]}(\boldsymbol{\Omega}^{(p)})\right\}=0. \tag{3.53}$$

We finally obtain the updated estimate $\boldsymbol{\xi}^{\top(p+1)}=(\mu_1^{(p+1)},\ldots,\mu_M^{(p+1)})$ by maximizing (3.53) which is similar to the updating equation for parameters of the component densities under the SRS design. Therefore, the modified version of the proposed EM algorithm for the imperfect RSS design requires the same computational efforts as

the EM algorithm based on SRS to update μ_j. Similar to the SRS case, one feature of this modified version of the EM algorithm is that the solutions to (3.53) often exist in closed form. According to the fact that f belongs to the location-shifted family of distribution, hence we can easily see that μ_j is the mean, median, and mode of f given $Z_{ij} = 1$; i.e., $\mathbb{E}(X_i|Z_{ij} = 1) = \mu_j$. Therefore, the ML estimate of μ_j is the same with the Method of Moments estimate of μ_j. Accordingly, one can obtain the $\mu_j^{(p+1)}$ as

$$\mu_j^{(p+1)} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{H}\tau_{ij}^{[r]}(\mathbf{\Omega}^{(p)})x_{[r]i}}{\sum_{i=1}^{n}\sum_{r=1}^{H}\tau_{ij}^{[r]}(\mathbf{\Omega}^{(p)})} \qquad j = 1, \ldots, M. \tag{3.54}$$

Chapter 4

Numerical Studies

In this chapter, we investigate the performance of the proposed estimation methods for semi-parametric finite mixture models (FMMs) through simulation study and a real data analysis. These estimation methods, as discussed in Chapters 2 and 3, include the development of parametric and semi-parametric version of the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood (ML) estimate of the non-parametric elements of FMMs under simple random sampling and ranked set sampling.

This chapter is organized as follows. In Section 4.1, through simulation studies, we investigate the effect of ranking errors and sampling parameters on the RSS estimators. We also compare the performance of the semi-parametric techniques in the estimation of FMMs. In Section 4.2, we apply the estimation methods to estimate more efficiently the bone mineral density (BMD) of women aged 50 and older.

4.1 Simulation Study

In this section, we use different simulation studies to investigate the performance of ML estimators of the unknown parameters of the semi-parametric finite mixture of normal distributions based on SRS and RSS data. In the simulation study, we estimate the misplacement error in the model based on the RSS data. In our simulation study, we compare the ML estimation of the parameters of the semi-parametric mixture model based on RSS data with their counterparts based on SRS data. For the sake of fair comparison, throughout the simulation studies, we treated the two-component mixture model given by

$$g(x;\mathbf{\Psi}) = \pi\phi(x;\mu_1,\sigma) + (1-\pi)\phi(x;\mu_2,\sigma), \tag{4.1}$$

as our underlying location-shifted mixture population with $(\pi,\mu_1,\mu_2,\sigma) = (0.4, 0, 3, 0.5)$ and $\phi(x;\mu,\sigma)$ is the pdf of the normal distribution with mean μ and standard deviation σ. We first generated ranked set samples and simple random samples of the same sizes from this population. We then simulated the semi-parametric estimators of population (4.1) where we assumed the component density f and parameters $\mathbf{\Psi} = (\pi,\mu_{\mathbf{1}},\mu_{\mathbf{2}})$ of population (4.1) are unknown throughout the estimation processes.

We evaluate the performance of the ML estimates under RSS and compare it with their competitors under SRS. In particular, the simulation study investigates two main characteristics of the estimators: robustness against possible ranking errors and the efficiency of the estimators. As discussed in Chapter 3, the ranking error involved in the RSS estimation method is modelled by matrix misplacement probabilities $\boldsymbol{\alpha}$. In the first simulation study, we investigate the performance of the

estimator of $\boldsymbol{\alpha}$ as the matrix of misplacement probabilities. Since the matrix $\boldsymbol{\alpha}$ is stochastic, many parameters of the matrix (out of H^2 elements) become redundant and can be obtained from the independent set of parameters. For example, when the set size is $H = 3$, there are only four independent parameters and should be estimated. These parameters include $\alpha_{1,1}, \alpha_{1,2}, \alpha_{2,1}, \alpha_{2,2}$. The other parameters of matrix $\boldsymbol{\alpha}$ can be computed from constraints $\sum_{h=1}^{H} \alpha_{h',h} = \sum_{h'=1}^{H} \alpha_{h',h} = 1$.

To obtain the ML estimation of $\boldsymbol{\alpha}$, we generated 3000 RSS samples of sizes 90 and 150 with set size $H = 3$ and cycle sizes $n \in \{30, 50\}$. To estimate the performance of ML estimate of $\boldsymbol{\alpha}$, we consider cases for imperfect ranking in RSS. These cases include $\rho = (1, 0.9, 0.75, 0.5)$, where ρ indicates the correlation between variable of interest X and ranking contaminant variable U. It is clear that when $\rho = 1$, there is no ranking error involved in RSS data. As ρ decreases, more ranking errors arise in RSS data.

We use the method of Dell and Clutter (1972) to generate the ranking variable U from the variable of interest X with $Cor(X, U) = \rho$ and estimate the matrix of misplacement probabilities. Let $\mathbf{C}$ is a matrix of $H \times H$ with all elements are 0. We first generate a set of H units from mixture models (4.1), $x_1, \ldots, x_H$, and standardize them such that $Var(X_i) = 1$. We then generate a set of H errors from a normal distribution with mean 0 and variance $\sigma_e^2 = \frac{1-\rho^2}{\rho^2}$. We now use $u_i = x_i + e_i$ to obtain a set of H observations $u_1, \ldots, u_H$ from ranking variable such that $Cor(X, U) = \rho$. To estimate the matrix misplacement probabilities based on these H observations, we first rank the H observations according to their X-variables. These ranks are called the true ranks of the units. We then rank these

H units based on their U-variable, and these ranks are called the judgmental ranks of the units. For this set, an integer (say r) is selected randomly from (1 to H). We then identify the true rank (say h) of the unit with judgemental rank r. We next update the $\mathbf{C}$ matrix as $\mathbf{C}_{h,r} = \mathbf{C}_{h,r} + 1$. This process is repeated 5000 times such that we estimate the misplacement matrix $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha} = \frac{\mathbf{C}}{5000}$ correspondingly when the ranking ability is $\rho = Cor(X, U)$. These estimations of $\boldsymbol{\alpha}$ matrices are then considered as the true parameters of misplacement matrices for the second simulation study for semi-parametric estimation of mixture model (4.1) with RSS data.

Here, we investigate the performance of the ML estimates of the FMM based RSS data and SRS data in semi-parametric setting. The ML estimates under RSS data are computed applying the modified EM algorithms developed in Chapter 3. We generate RSS and SRS data of the same size from the underlying population (4.1). In this simulation study, we shall apply the semi-parametric EM algorithms developed in Chapters 2 and 3 to obtain the non-parametric estimation of the component density of population (4.1) as well as the ML estimates of the FMM parameters $\boldsymbol{\Psi} = (\pi, \mu_{\mathbf{1}}, \mu_{\mathbf{2}})$ when σ is considered known ($\sigma = 0.5$).

To simulate the estimation methods, we generated 3000 SRS and RSS samples of the same sizes $N = \{90, 150\}$ and computed the MSE and bias as two criteria to evaluate the performance of the proposed estimators. To better assess the effect of set size on the performance of RSS-based estimates, we select set size $H = \{3, 5\}$. We select cycle sizes $n = \{30, 50\}$ (when $H = 3$) and $n = \{18, 30\}$ (when $H = 5$) to generate the RSS samples of sizes $n \times H = \{90, 150\}$ (i.e. the same size as SRS samples). We considered four ranking cases $\rho = \{1, 0.9, 0.75, 0.5\}$ in the collection

of RSS data to investigate the effect of ranking errors on the performance of the RSS-based estimates.

In each replication of the simulation, the initial values of $\boldsymbol{\Psi} = (\pi, \mu_1, \mu_2)$ in the EM-algorithms are computed following the method of Furman and Lindsay (1994) by treating the RSS sample as a simple random sample.

Table 4.1: $\hat{\alpha}_{MLE}$ based on RSS design when $H = 3$.

n	ρ	α_{11}	α_{12}	α_{21}	α_{22}
30	1	0.99924	0.00070	0.00070	0.99662
	0.9	0.99901	0.00086	0.00086	0.99650
	0.75	0.99878	0.00098	0.00098	0.99617
	0.5	0.99864	0.00111	0.00111	0.99594
50	1	0.99828	0.00138	0.00138	0.99599
	0.9	0.99895	0.00093	0.00093	0.99681
	0.75	0.99866	0.00110	0.00110	0.99658
	0.5	0.99839	0.00130	0.00130	0.99627

Table 4.1 shows the estimation of misplacement probabilities when set size H is 3 for different cycle sizes $n \in \{30, 50\}$ and different correlation coefficients $\rho \in \{1, 0.9, 0.75, 0.5\}$. Tables 4.2 and 4.3 present the biases and MSE for the parameter estimates for ranking errors $\boldsymbol{\alpha}$.

From Tables 4.2 and 4.3, it is clear that the estimators of misplacement probabilities perform very well when ranking ability (i.e. the correlation between the ranking variable and variable of interest) is strong. For example, when the ranking ability is high, $\rho \in \{1, 0.9\}$, the bias of the estimates is very small such that the estimators can be considered practically unbiased. As ρ decreases, more ranking errors are introduced in the RSS data collection; hence we observe that the MSE and bias of the estimates of misplacement probabilities increase. Table 4.4 and 4.5 present

Table 4.2: The biases and MSEs for ranking errors $\boldsymbol{\alpha}_i$, $i = 1, \ldots, 4$ when set size $H = 3$ and cycle size $n = 30$ for different correlation coefficients $\rho \in \{1, 0.9, 0.75, 0.5\}$ based on RSS design.

ρ		α_{11}	α_{12}	α_{21}	α_{22}
1	Bias	-0.0008	0.0007	0.0007	-0.0034
	MSE	0.0131	0.0099	0.0099	0.0219
0.9	Bias	0.0920	-0.0751	-0.0751	0.1745
	MSE	0.0165	0.0121	0.0121	0.0261
0.75	Bias	0.1758	-0.1380	-0.1380	0.2952
	MSE	0.0195	0.0139	0.0139	0.0313
0.5	Bias	0.2526	-0.1804	-0.1804	0.4089
	MSE	0.0209	0.0147	0.0147	0.0331

Table 4.3: The biases and MSEs for ranking errors $\boldsymbol{\alpha}_i$, $i = 1, \ldots, 4$ when set size $H = 3$ and cycle size $n = 50$ for different correlation coefficients $\rho \in \{1, 0.9, 0.75, 0.5\}$ based on RSS design.

ρ		α_{11}	α_{12}	α_{21}	α_{22}
1	Bias	-0.0017	0.0014	0.0014	-0.0040
	MSE	0.0249	0.0187	0.0187	0.0372
0.9	Bias	0.0909	-0.0881	-0.0881	0.1958
	MSE	0.0175	0.0138	0.0138	0.0293
0.75	Bias	0.1467	-0.1424	-0.1424	0.3046
	MSE	0.0213	0.0159	0.0159	0.0342
0.5	Bias	0.2074	-0.1672	-0.1672	0.4093
	MSE	0.0249	0.0178	0.0178	0.0366

the performance of ML estimators of the unknown parameters $\boldsymbol{\Psi} = (\pi, \mu_1, \mu_2)$ of FMM (4.1) based on SRS and RSS data.

From Tables 4.2 and 4.3, we observe that the bias values of all estimators are very small (even under RSS data with low ranking ability) such that we can consider these estimators are almost unbiassed in the estimation of the parameters. Comparing the MSEs of RSS-based estimators and their SRS counterparts, we see that the RSS estimators almost outperform their SRS competitors in the semi-parametric

Table 4.4: Biases and MSEs of (π, μ_1, μ_2) when set size $H \in \{3, 5\}$ and sample size $N = 90$ for different correlation coefficients $\rho \in \{1, 0.9, 0.75, 0.5\}$ based on SRS and RSS designs.

	H	n	ρ		π	μ_1	μ_2
SRS	—	—	—	Bias	-0.0017	0.0073	-0.0086
				MSE	0.0509	0.0853	0.0707
RSS	3	30	1	Bias	-0.0112	-0.0044	-0.0103
				MSE	0.0416	0.0853	0.0649
			0.9	Bias	-0.0089	-0.0045	-0.0092
				MSE	0.0421	0.0893	0.0672
			0.75	Bias	-0.0132	-0.0038	-0.0104
				MSE	0.0404	0.0865	0.0674
			0.5	Bias	-0.0164	-0.0081	-0.0084
				MSE	0.0418	0.0927	0.0731
	5	18	1	Bias	-0.0026	0.0020	-0.0095
				MSE	0.0369	0.0817	0.0611
			0.9	Bias	-0.0008	0.0052	-0.0104
				MSE	0.0372	0.0848	0.0659
			0.75	Bias	-0.0022	0.0057	-0.0096
				MSE	0.0367	0.0842	0.0661
			0.5	Bias	-0.0026	0.0043	-0.0098
				MSE	0.0382	0.0874	0.0721

estimation of the FMMs. In addition, we observe that when the ranking ability is strong, the performance of the RSS estimators increases as the set size increases from $H = 3$ to $H = 5$ (while the sample size remains the same). However, we should note that when the ranking ability is low, the performance of the RSS estimates decreases as the set size increases from $H = 3$ to $H = 5$.

We present the kernel density estimates of the component density f, the mixture distribution g and the histogram based on SRS and RSS samples of size nH for one sample data in Figures 4.1 to 4.6. Following Bordes et al. (2006), we selected the bandwidth, for both SRS and RSS data, close to the minimum of the mean inte-

Table 4.5: Biases and MSE of (π, μ_1, μ_2) when set size $H \in \{3, 5\}$ and sample size $N = 150$ for different correlation coefficients $\rho \in \{1, 0.9, 0.75, 0.5\}$ based on SRS and RSS designs.

	H	n	ρ		π	μ_1	μ_2
SRS	—	—	—	Bias	-0.0001	0.0039	-0.0075
				MSE	0.0414	0.0655	0.0543
RSS	3	50	1	Bias	-0.0067	-0.0025	-0.0078
				MSE	0.0322	0.0688	0.0545
			0.9	Bias	-0.0069	-0.0010	-0.0097
				MSE	0.0320	0.0757	0.0579
			0.75	Bias	-0.0085	-0.0043	-0.0056
				MSE	0.0312	0.0710	0.0556
			0.5	Bias	-0.0123	-0.0020	-0.0086
				MSE	0.0312	0.0797	0.0602
	5	30	1	Bias	-0.0004	0.0065	-0.0063
				MSE	0.0290	0.0626	0.0478
			0.9	Bias	-0.0007	0.0053	-0.0079
				MSE	0.0287	0.0641	0.0484
			0.75	Bias	-0.0009	0.0063	-0.0086
				MSE	0.0285	0.0648	0.0501
			0.5	Bias	-0.0013	0.0057	-0.0092
				MSE	0.0301	0.0680	0.0539

grated square error $h_n = (\frac{4}{3n})^{\frac{1}{5}}$ to obtain these non-parametric density estimates. Comparing Figures 4.2 and 4.5, we see that the non-parametric estimation of mixture model based on commonly used simple random sample assigns the same weight to different components. It thus misses the clear peaks of the mixture model. Unlike the SRS estimates, the non-parametric estimates of the mixture model based on RSS data clearly detect the peaks of the two components and appropriately estimate non-parametrically the weight of each component. This superiority of RSS-based kernel density estimator rises because ranked set sampling efficiently utilizes the ranking information of RSS statistics into the estimation process.

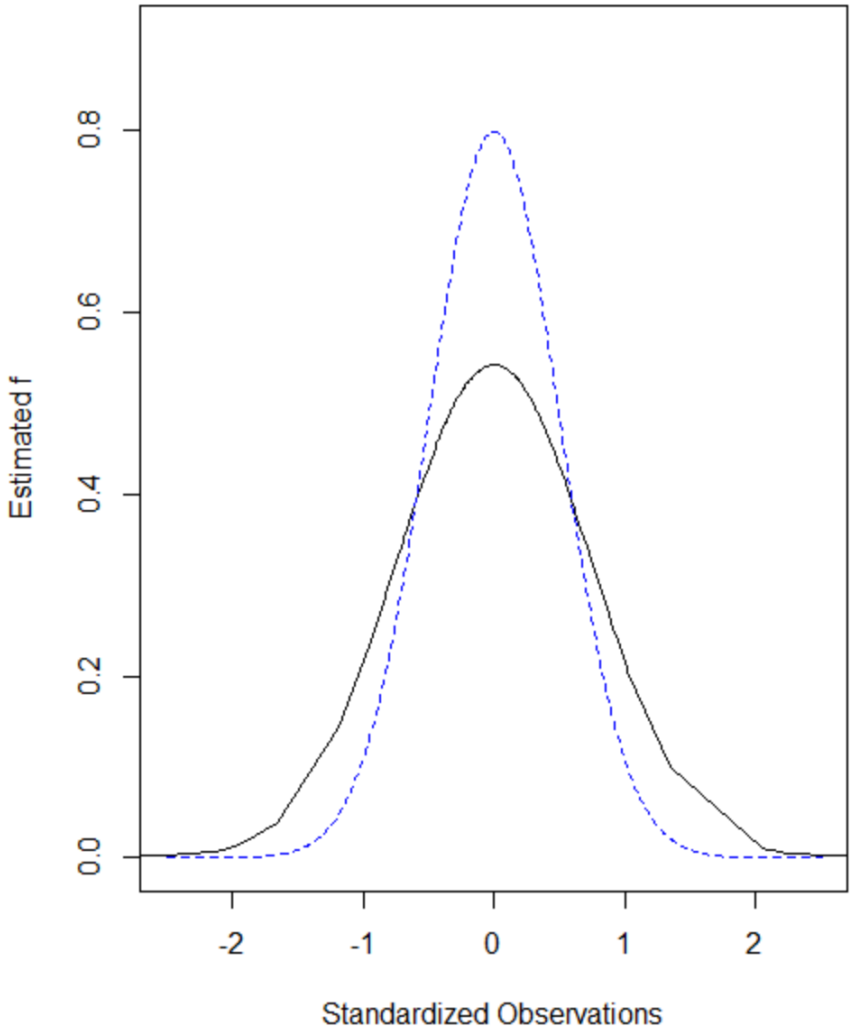

Figure 4.1: Estimated f distribution based on the SRS data and true f density (dashed blue line).

4.2 Real Data Analysis

Osteoporosis is a bone metabolic disease identified by diminished bone mineral density (BMD) that leads to increased skeletal delicacy and risk of breakage (Kanis, 1997). Osteoporosis and osteoporosis-related fractures are critical common health

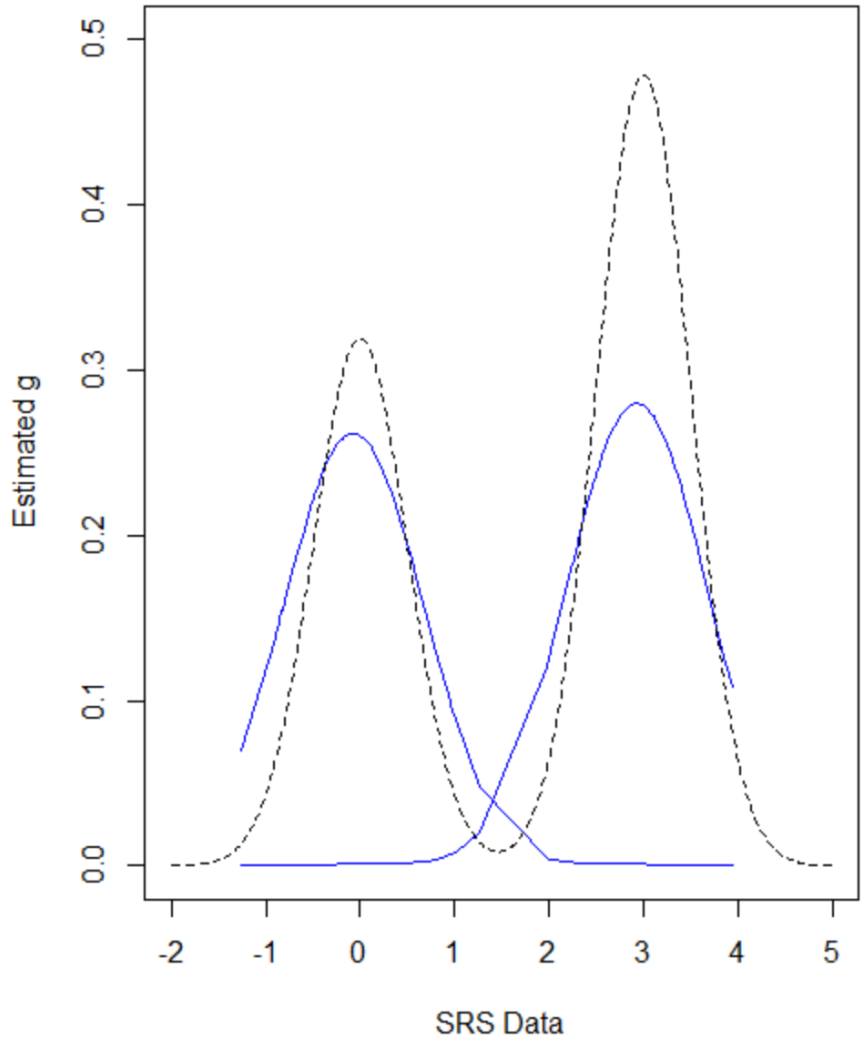

Figure 4.2: Estimated g distribution based on the SRS data (blue line) and the mixture is shown by dashed line.

difficulties. The risk of osteoporosis grows by age. In healthy young adults, bones grow until ages around 20 to 30, gradually becoming weaker as someone gets older. As the old population is rising in many nations, osteoporotic fractures are supposed to increase. Because of osteoporosis's financial, medical, and social results, controlling the prevention of this disorder is considered necessary to the care of the

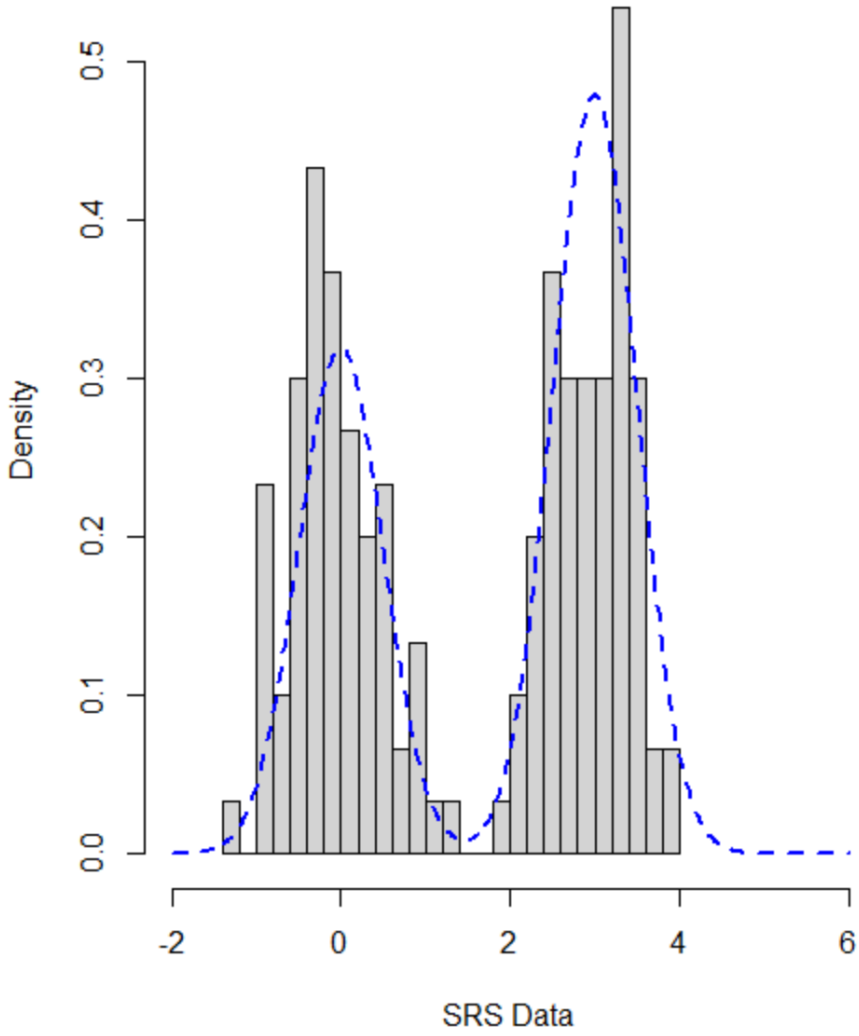

Figure 4.3: Histogram of the SRS data and the estimated mixture (dashed blue line).

well-being, quality of life, and self-confidence of older people. Furthermore, BMD measurements are collected via X-ray absorptiometry (DXA). Once we obtain the X-ray images, medical experts should examine these images manually to find final BMD measurements. We see that the procedure of obtaining BMD is expensive

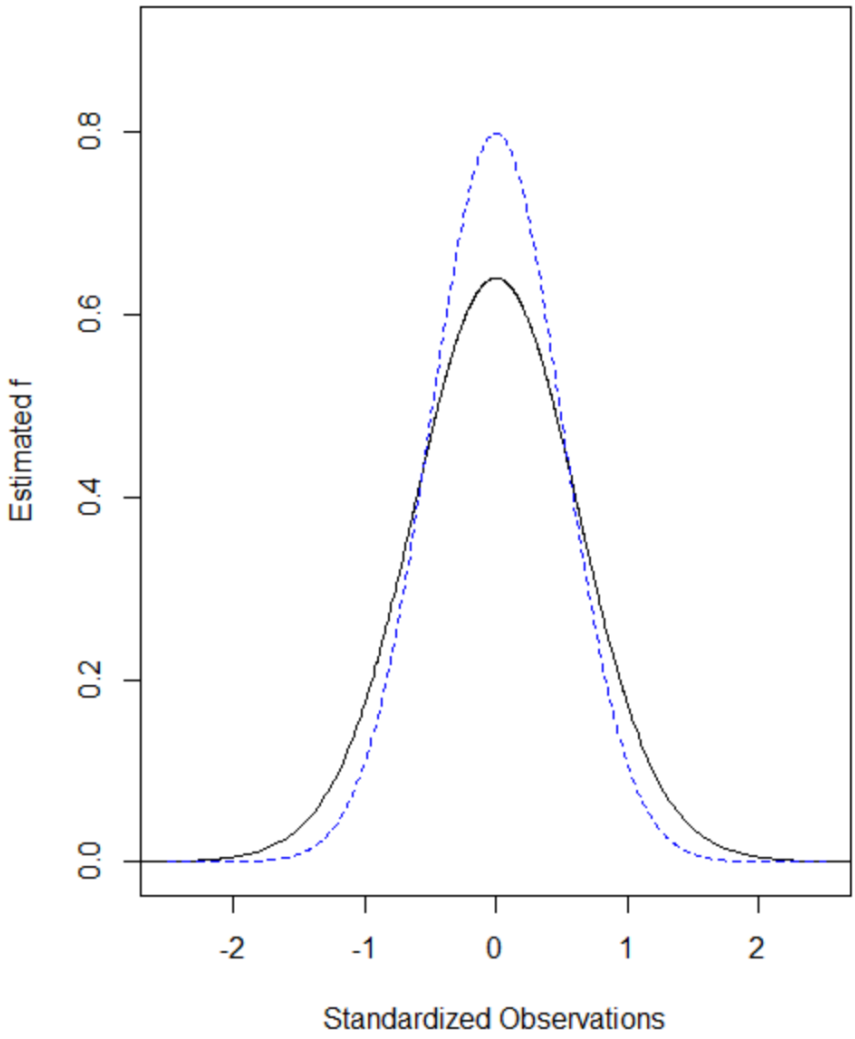

Figure 4.4: Estimated f distribution based on the RSS data and true f density (dashed blue line).

and time-consuming. In 1994, the World Health Organization (WHO) established BMD using DXA as one of the most reliable predictors for measuring osteoporosis. BMD measurements are typically converted into T-scores for diagnostic measures, showing the number of standard deviations over or under the mean in healthy adults. Particularly, the bone status of a patient is determined osteoporosis when

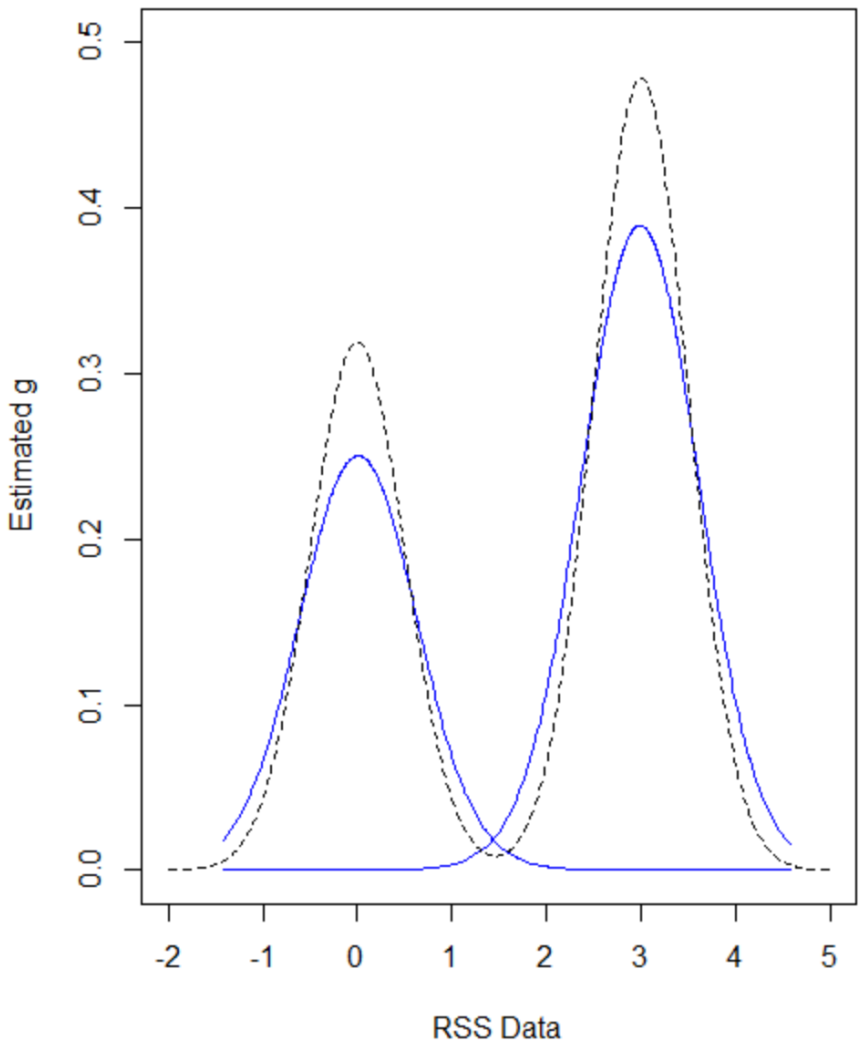

Figure 4.5: Estimated g distribution based on the RSS data (blue line) and the mixture is shown by dashed line.

the T-score is less than 2.5 SD from the BMD norm of the population.

In this section, we focus on data set from the National Health and Nutrition Examination Survey (NHANES) III conducted by Centers for Cancer Diseases and Prevention (CDC) to obtain nutrition and health information for American adults between 1988 to 1994. The data set is available online on the website of NHANES

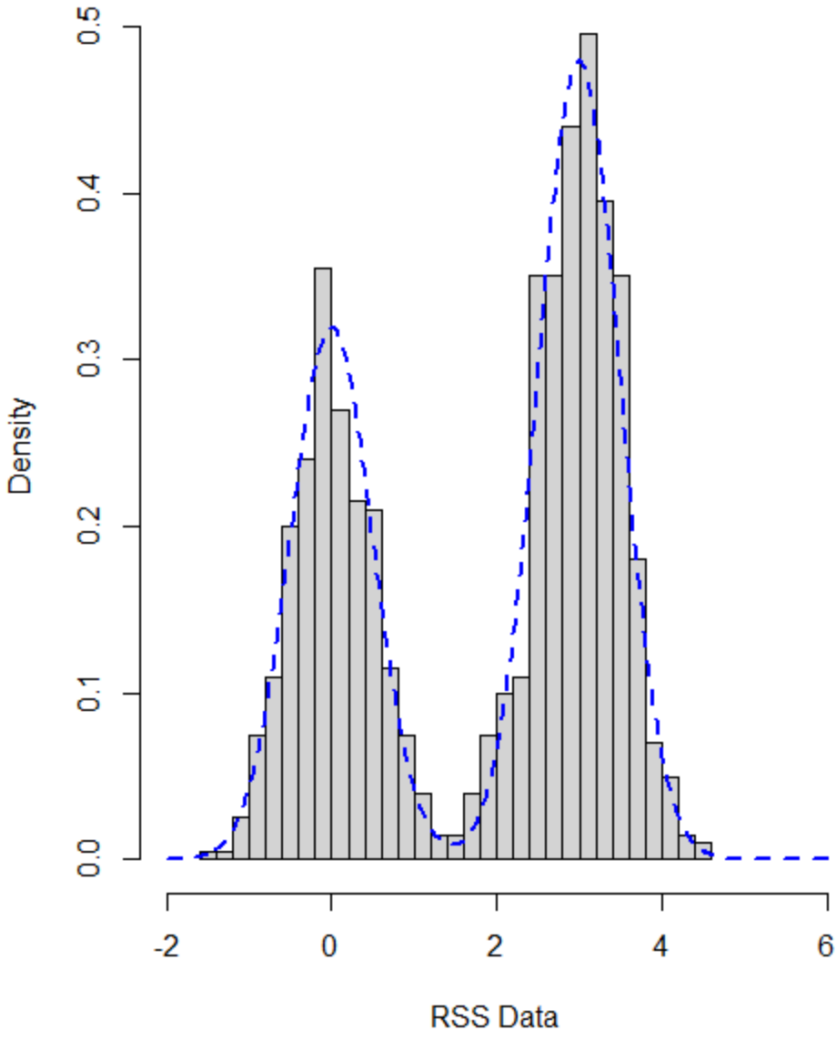

Figure 4.6: Histogram of the RSS data and the estimated mixture (dashed blue line).

III. The survey contains the health and nutritional characteristics of 33994 individuals. The osteoporosis is more common in women than men. Worldwide, 1 in 3 women over age 50 will encounter osteoporotic breaks, as will 1 in 5 men aged over 50 (Melton III, 1995). Due to the clear impact of osteoporosis on older women, we focus on BMD data on women aged 50 and over in the NHANES III.

In the data set, we have access to BMD measurements of 3299 women aged 50 and older. We treat these 3299 BMD measurements as our underlying population. Using BIC model selection for the population, a mixture of two-component was suggested the best fit for the population. According to the symmetric properties of the component densities of the population, using the entire information of the population, we estimate the population by a mixture of two normal distributions as follows,

$$g(x;\mathbf{\Psi}) = 0.864\ \ \phi(x;0.662,0.116) + 0.136\ \ \phi(x;0.895,0.116). \tag{4.2}$$

Similarly to the simulation study, the component density functions and $\mathbf{\Psi} = (\pi, \mu_1, \mu_2)$ are unknown parameters of the population, while $\sigma = 0.116$ is treated as known in the estimation process.

Many research works have studied the association between the risk of osteoporosis and other characteristics of the patients. These characteristics include age, weight, BMI. For more information, readers are referred to Nahhas et al. (2002). In this real data example, we consider the weight of the patients as the contaminant variable and used it for ranking the sampling units in RSS data collections when the correlation between weight and BMD is $\rho = 0.53$. Similar to Section 4.1, we generated 3000 SRS and RSS samples of sizes $\{90, 150\}$ with replacement from the population.

Applying the semi-parametric EM-algorithms based on RSS data, we computed the biases and MSEs of the RSS estimates and SRS estimates. In each replication, we used weight characteristic to rank the patients in each set and collected ranked set samples with set size $H = \{3, 5\}$ and cycle size $n = \{30, 50\}$ (for $H = 3$)

and $n = \{18, 30\}$ (for $H = 5$). Similar to the simulation study, we obtained the biases and MSEs of the estimators of misplacement probabilities and population parameters. Table 4.6 shows the average, the bias and MSEs of the estimates of the misplacement probabilities. Although the correlation between weight characteristic and BMD is $\rho = 0.53$, we observe that, on average, we can assign the correct rank to the units in each set. For example, the probability that we assign correct ranks 1 and 2 to units in each set are 0.99 and 0.97, respectively. Table 4.7 shows the biases and MSEs of the estimates of (π, μ_1, μ_2) under the SRS and RSS data of sizes $N \in \{90, 150\}$ when set size $H \in \{3, 5\}$.

Table 4.6: The $\boldsymbol{\alpha}_{\hat{MLE}}$ and their biases and MSEs values when the set size $H = 3$.

n		α_{11}	α_{12}	α_{21}	α_{22}
30	Estimation	0.99150	0.00750	0.00750	0.97238
	Bias	0.44350	-0.28350	-0.28350	0.56738
	MSE	0.05233	0.03581	0.03581	0.09462
50	Estimation	0.93313	0.05213	0.05213	0.82887
	Bias	0.40213	-0.24537	-0.24537	0.44487
	MSE	0.14560	0.09126	0.09126	0.21702

From Table 4.7, we observe that the biases associated with almost all of the estimators are very small such that we can consider the estimators are practically unbiased in the estimation of the BMD population. We also observe that RSS estimates are more efficient than their SRS estimates. Note that this superiority is marginal. This is because the correlation between the weight characteristic and BMD response was to same extent low such that many ranking errors are produced in the collection of RSS estimates.

We show the kernel density estimates of the component density f, the mixture distribution g and the histogram based on SRS and RSS samples of size N from

Table 4.7: The biases and MSEs of the estimates of (π, μ_1, μ_2) under the SRS and RSS data of sizes $N \in \{90, 150\}$ when set size $H \in \{3, 5\}$.

Design	H	n	N		π	μ_1	μ_2
SRS	—	—	90	Bias	0.0268	0.0095	0.0954
				MSE	0.1266	0.0244	0.1561
	—	—	150	Bias	0.0379	0.0116	0.0889
				MSE	0.1051	0.0202	0.1486
RSS	3	30	90	Bias	-0.0906	0.0286	-0.1845
				MSE	0.1567	0.0146	0.0503
		50	150	Bias	-0.0211	0.0109	-0.0757
				MSE	0.0855	0.0194	0.0902
	5	18	90	Bias	-0.0459	0.0170	-0.1097
				MSE	0.1323	0.0214	0.1194
		30	150	Bias	-0.0273	0.0087	-0.0697
				MSE	0.0873	0.0195	0.0897

BMD data for one sample in Figures 4.7 to 4.12. One practically may miss the second component of the bone mixture through the SRS estimation method; but RSS estimator still very well detects the two components. While the difference between the two subpopulations is very close and it is hard for most estimators to distinguish them.

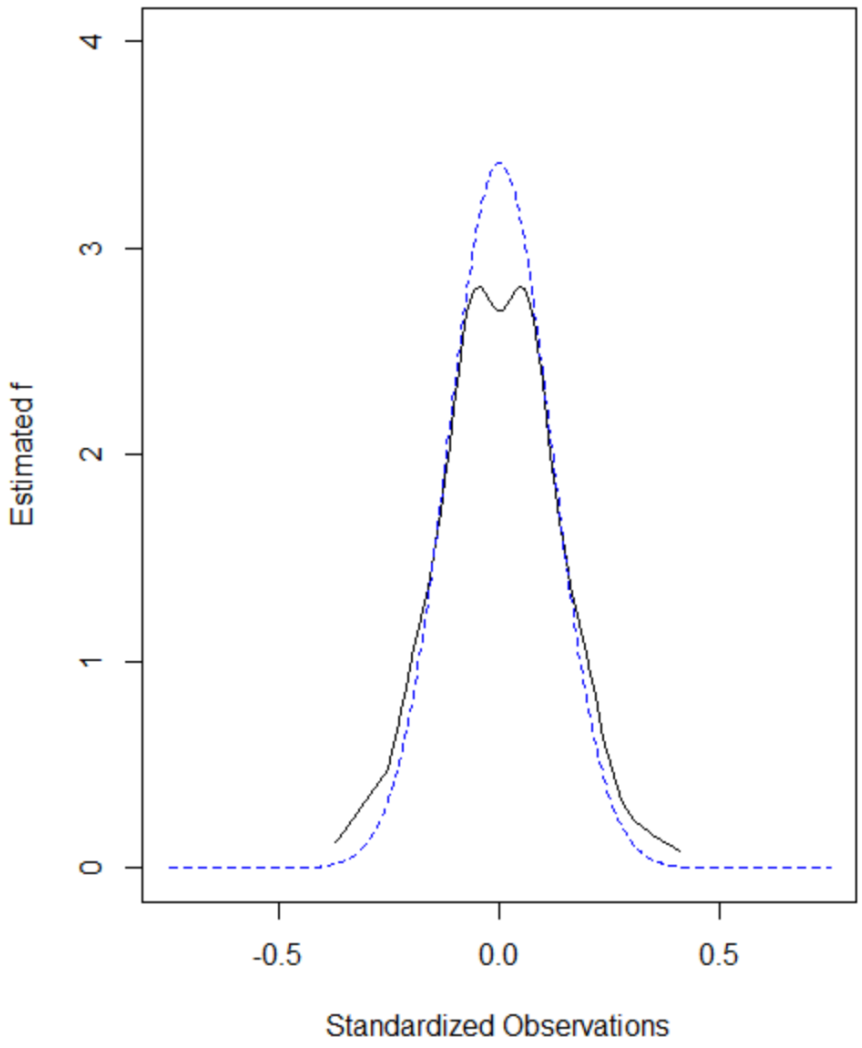

Figure 4.7: Estimated f distribution based on the SRS data and true f density (dashed blue line).

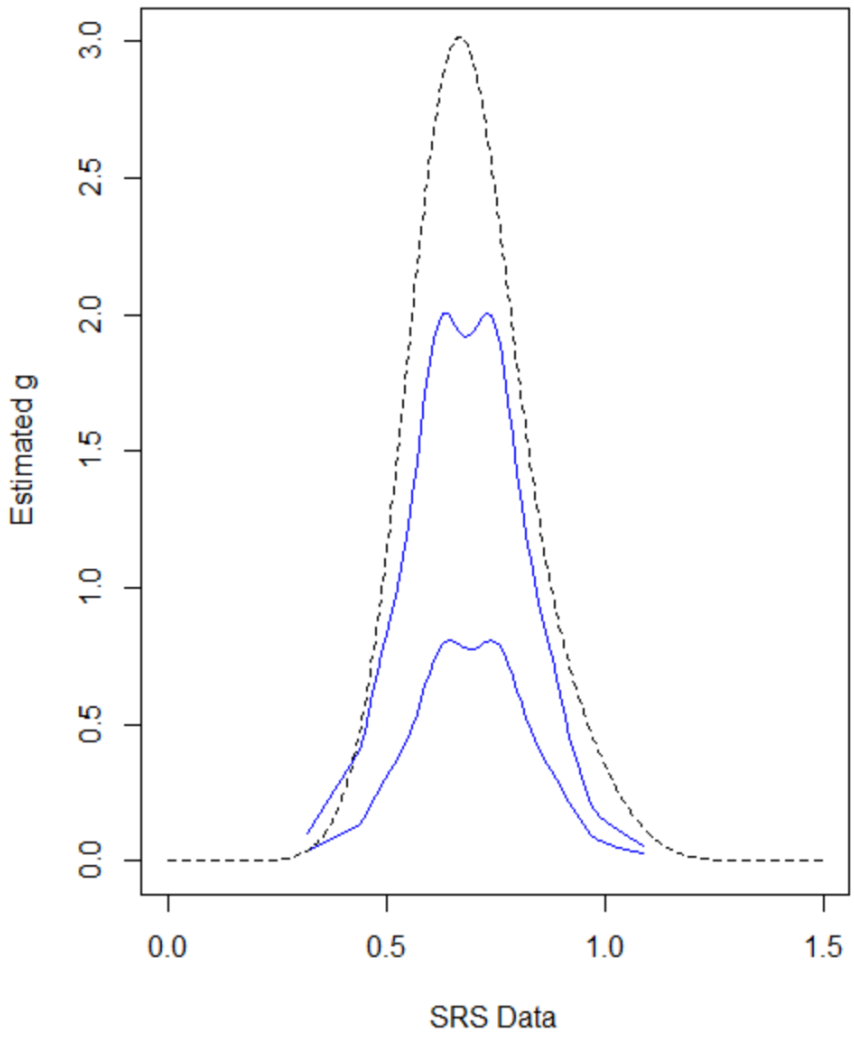

Figure 4.8: Estimated g distribution based on the SRS data (blue line) and the mixture is shown by dashed line.

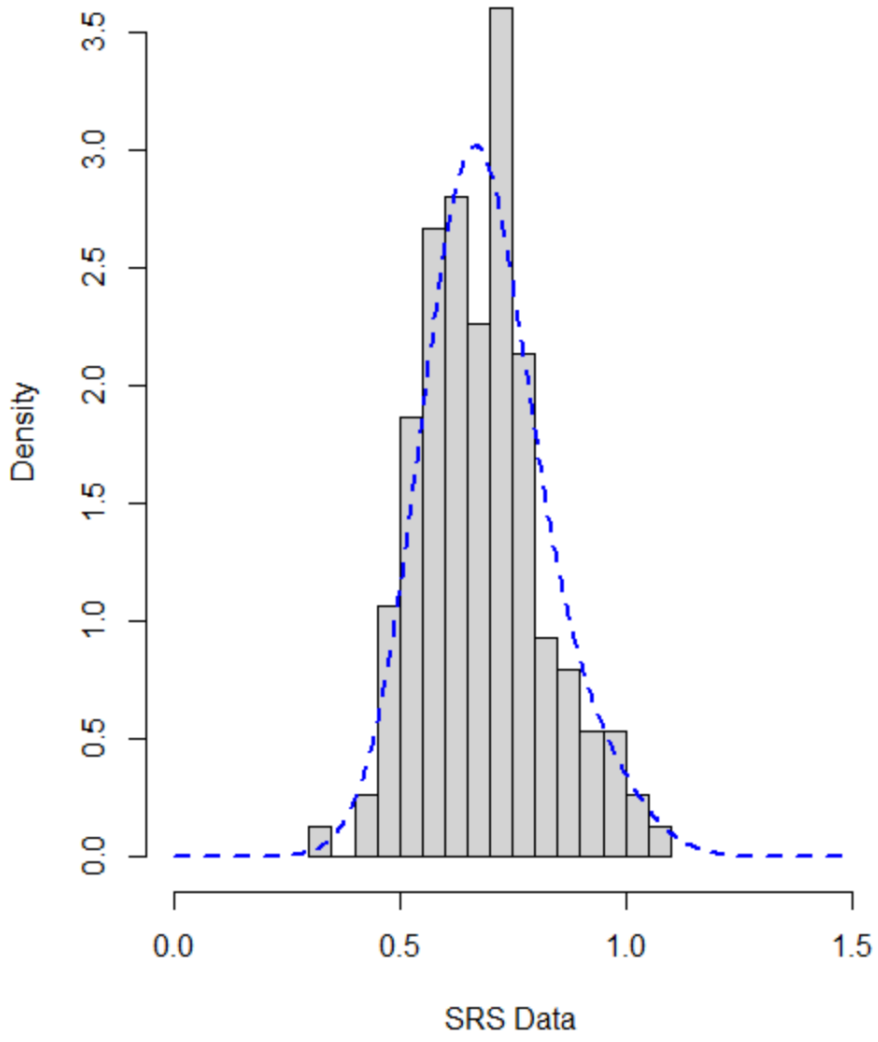

Figure 4.9: Histogram of the SRS data and the estimated mixture (dashed blue line).

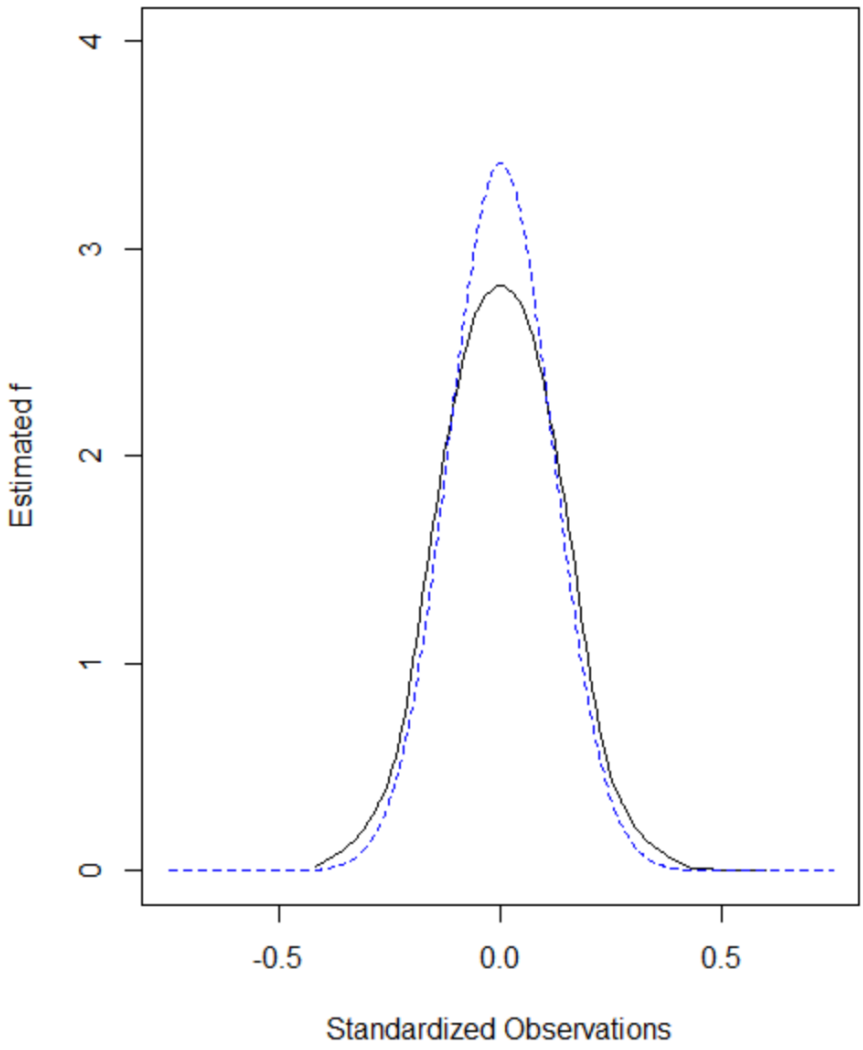

Figure 4.10: Estimated f distribution based on the RSS data and true f density (dashed blue line).

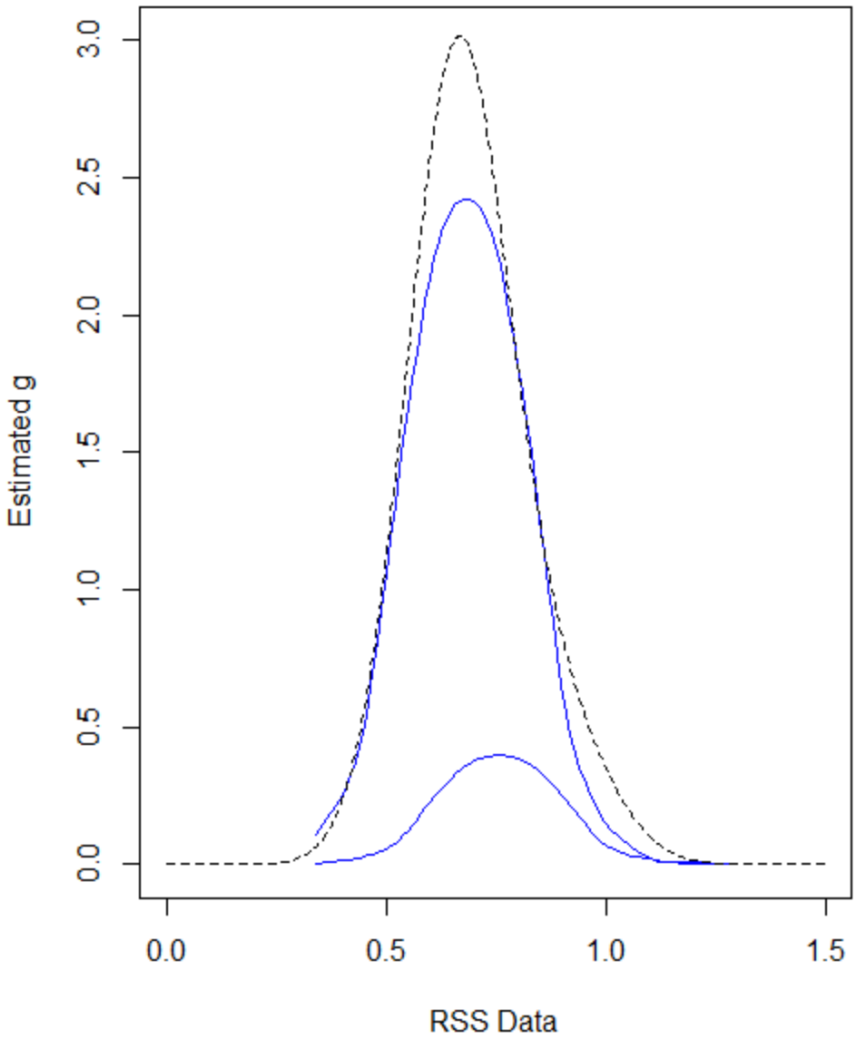

Figure 4.11: Estimated g distribution based on the RSS data (blue line) and the mixture is shown by dashed line.

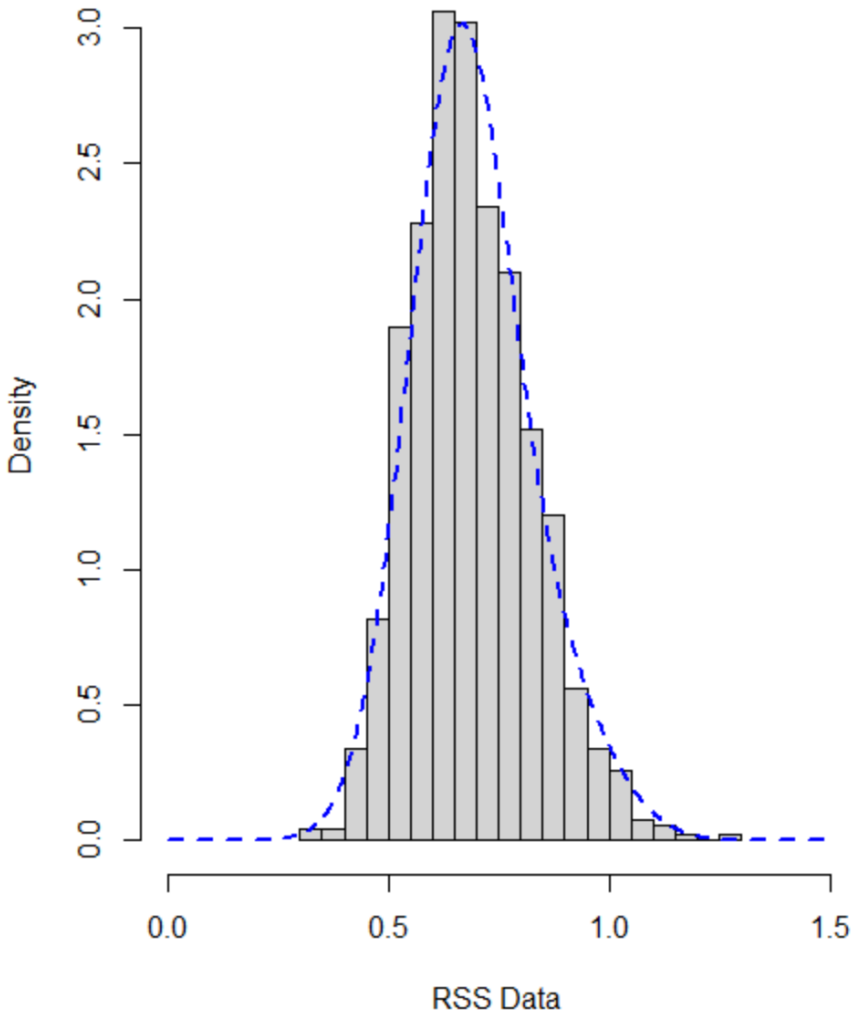

Figure 4.12: Histogram of the RSS data and the estimated mixture (dashed blue line).

Chapter 5

Summary and Future Work

5.1 Summary

This thesis investigated semi-parametric estimation of the finite mixture models utilizing ranked set sampling (RSS). We developed more efficient semi-parametric estimators for finite mixture models (FMMs) using RSS. FMMs, as key statistical tools in data analysis, play fundamental roles in mainstream statistical analysis. In addition, FMMs have found applications in different scientific fields such as genetics, medical studies, different engineering fields, etc.

Ranked set sampling is cost-effective sampling technique that can be applied in situations where the precise measurement of the variable of interest is expensive or hard to achieve; however, sampling units can be ranked via extra variables or judgment ranking, without actual measurements on the variable of interest. In the standard estimation methods for FMMs, the samples are typically extracted from the population using simple random sampling (SRS). In this thesis, we used RSS to collect more informative samples from the FMMs and developed more efficient semi-parametric estimations for the FMMs.

In Chapter 2, we studied the semi-parametric estimation of FMMs from SRS. We presented the likelihood function of the mixture model based on simple random samples. We used a semi-parametric version of the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood (ML) estimate of the parametric and non-parametric elements of the underlying mixture model. We described how one could develop the missing-data mechanism and EM-algorithm to obtain the ML estimate of the finite mixture model.

In Chapter 3, we developed the ML estimation of FMM with RSS data in a semi-parametric framework. A comprehensive EM algorithm was proposed to estimate the parameters of semi-parametric components of FMMs. One significant difficulty in obtaining the ML estimation based on RSS data from the semi-parametric FMMs was its computational burden. To overcome this problem, we presented a modified version of the developed EM algorithm, which reduces the computational burden to the level of the standard EM algorithm based on SRS data.

Our numerical studies, in Chapter 4, showed that the proposed EM-algorithm works properly in estimating the underlying FMM. Simulation studies showed that RSS estimators outperform their SRS competitors in the semi-parametric estimation of FMMs in terms of bias and MSE. The proposed methods were finally applied to analyze the bone mineral densities (BMD) data and obtain the ML estimate of the distribution of BMD data.

5.2 Future Work

In the future, we will study the ML estimation of a semi-parametric FMM using judgment post-stratification (JPS) samples. One challenge of semi-parametric es-

timations of FMMs with RSS data is that the rank of RSS statistics cannot be separated from the observations. Hence, when the ranking error is increased, we are not able to recognize between observation information and ranking information. Semi-parametric estimation of FFMs based on JPS data can be considered as a remedy to this challenge. JPS method as post-stratification sampling enables us to separate ranking information from the observation if we find the ranking error is overwhelming in data collection. To obtain JPS data, sampled units are post-stratified on ranks by randomly choosing comparison sets for each unit from the underlying population and allocating ranks to them, applying judgment ranking. This happens in a set of independent order statistics from the underlying model, where the number of units in each rank class is random. We need to develop a new missing data mechanism to facilitate the likelihood function by introducing latent variables and estimate the unknown parameters of a semi-parametric FMM.

Bibliography

Arslan, G. and Ozturk, O. 2013. Parametric inference based on partially rank ordered set samples. *Journal of The Indian Statistical Association* 51:1–24.

Barabesi, L. 2001. The unbalanced ranked-set sample sign test. *Journal of Nonparametric Statistics* 13:279–289.

Barabesi, L. and El-Sharaawi, A. 2001. The efficiency of ranked set sampling for parameter estimation. *Statistics and probability letters* 53:189–199.

Benaglia, T., Chauveau, D., and Hunter, D. R. 2009. An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics* 18:505–526.

Bohn, L. L. 1996. A review of nonparametric ranked-set sampling methodology. *Communications in Statistics–Theory and Methods* 25:2675–2685.

Bordes, L., Chauveau, D., and Vandekerkhove, P. 2007a. A stochastic em algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis* 51:5429–5443.

Bordes, L., Chauveau, D., and Vandekerkhove, P. 2007b. A stochastic em algorithm for a semiparametric mixture model. *Comput. Stat. Data Anal.* 51:5429–5443.

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

Cassie, R. M. 1954. Some uses of probability paper in the analysis of size frequency distributions. *Marine and Freshwater Research* 5:513–522.

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

Chen, H. and Chen, J. 2003. Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica* 13:351–366.

Chen, J., Li, P., and Fu, Y. 2012. Inference on the order of a normal mixture. *Journal of the American Statistical Association* 107:1096–1105.

Chen, Z. 2000. The efficiency of ranked-set sampling relative to simple random sampling under multi-parameter families. *Statistica Sinica* 10:247–264.

Chen, Z., Bai, Z., and Sinha, B. 2003. Ranked set sampling: theory and applications, volume 176. Springer Science & Business Media.

Chen, Z. and Wang, Y.-G. 2004. Efficient regression analysis with ranked-set sampling. *Biometrics* 60:997–1004.

Cohen, A. C. 1967. Estimation in mixtures of two normal distributions. *Technometrics* 9:15–28.

Cruz-Medina, I. and Hettmansperger, T. 2004. Nonparametric estimation in semi-parametric univariate mixture models. *Journal of Statistical Computation and Simulation* 74:513–524.

Dacunha-Castelle, D., Gassiat, E., et al. 1999. Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *Annals of Statistics* 27:1178–1209.

Dell, T. and Clutter, J. 1972. Ranked set sampling theory with order statistics background. *Biometrics* pp. 545–555.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39:1–22.

Duda, R. O., Hart, P. E., et al. 1973. Pattern classification and scene analysis, volume 3. Wiley New York.

El Zaart, A., Ziou, D., Wang, S., and Jiang, Q. 2002. Segmentation of sar images. *Pattern Recognition* 35:713–724.

Furman, W. D. and Lindsay, B. G. 1994. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational statistics & data analysis* 17:493–507.

Hall, P. 1981. On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society: Series B (Methodological)* 43:147–156.

Hall, P., Zhou, X.-H., et al. 2003. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics* 31:201–224.

Halls, L. K. and Dell, T. R. 1966. Trial of ranked-set sampling for forage yields. *Forest Science* 12:22–26.

Harding, J. 1949. The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the United Kingdom* 28:141–153.

Hatefi, A. and Jozani, M. J. 2013. Fisher information in different types of perfect and imperfect ranked set samples from finite mixture models. *Journal of Multivariate Analysis* 119:16–31.

Hatefi, A., Jozani, M. J., and Ozturk, O. 2015. Mixture model analysis of partially rank-ordered set samples: age groups of fish from length-frequency data. *Scandinavian Journal of Statistics* 42:848–871.

Hatefi, A., Jozani, M. J., and Ziou, D. 2014. Estimation and classification for finite mixture models under ranked set sampling. *Statistica Sinica* pp. 675–698.

Hettmansperger, T. and Thomas, H. 2000. Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62:811–825.

Hosmer Jr, D. W. 1973. On mle of the parameters of a mixture of two normal distributions when the sample size is small. *Communications in Statistics-Theory and Methods* 1:217–227.

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

Johnson, R. A., Mehrotra, K. G., et al. 1972. Locally most powerful rank tests for the two-sample problem with censored data. *The Annals of Mathematical Statistics* 43:823–831.

Kanis, J. A. 1997. Bone density measurements and osteoporosis.

Lange, K. 1995. A quasi-newton acceleration of the em algorithm. *Statistica sinica* pp. 1–18.

Lemdani, M., Pons, O., et al. 1999. Likelihood ratio tests in contamination models. *Bernoulli* 5:705–719.

Leroux, B. G. 1992. Consistent estimation of a mixing distribution. *The Annals of Statistics* pp. 1350–1360.

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

McIntyre, G. 1952. A method for unbiased selective sampling, using ranked sets. *Crop and Pasture Science* 3:385–390.

McLachlan, G. and Peel, D. 2004. Finite mixture models. Wiley. com.

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

Mehrotra, K. and Nanda, P. 1974. Unbiased estimation of parameters by order statistics in the case of censored samples. *Biometrika* 61:601–606.

Melton III, J. L. 1995. Perspectives: how many women have osteoporosis now? *Journal of Bone and Mineral Research* 10:175–177.

Muttlak, H. and McDonald, L. 1992. Ranked set sampling and the line intercept method: A more efficient procedure. *Biometrical Journal* 34:329–346.

Nahhas, R. W., Wolfe, D. A., and Chen, H. 2002. Ranked set sampling: cost and optimal set size. *Biometrics* 58:964–971.

Omidvar, S., Jafari Jozani, M., and Nematollahi, N. 2018. Judgment post-stratification in finite mixture modeling: An example in estimating the prevalence of osteoporosis. *Statistics in medicine* 37:4823–4836.

Öztürk, Ö. 1999. Two-sample inference based on one-sample ranked set sample sign statistics. *Journal of Nonparametric Statistics* 10:197–212.

Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A* 185:71–110.

Peters, Jr, B. C. and Walker, H. F. 1978. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal on Applied Mathematics* 35:362–378.

Presnell, B. and Bohn, L. L. 1999. U-statistics and imperfect ranking in ranked set sampling. *Journal of nonparametric statistics* 10:111–126.

Roeder, K. 1994. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association* 89:487–495.

Schlattmann, P. 2009. Medical applications of finite mixture models. Springer.

Schork, N. J., Allison, D. B., and Thiel, B. 1996. Mixture distributions in human genetics research. *Statistical Methods in Medical Research* 5:155–178.

Sinha, B. K., Sinha, B. K., and Purkayastha, S. 1996. On some aspects of ranked set sampling for estimation of normal and exponential parameters. *Statistics and Decisions-International Journal for Stochastic Methods and Models* 14:223–240.

Tan, W. and Chang, W. 1972. Convolution approach to the genetic analysis of quantitative characters of self-fertilized populations. *Biometrics* pp. 1073–1090.

Titterington, D. M. 1983. Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society: Series B (Methodological)* 45:37–46.

Titterington, D. M., Smith, A. F., Makov, U. E., et al. 1985. Statistical analysis of finite mixture distributions, volume 7. Wiley New York.

Wang, Y.-G., Ye, Y., and Milton, D. A. 2009. Efficient designs for sampling and subsampling in fisheries research based on ranked sets. *ICES Journal of Marine Science: Journal du Conseil* 66:928–934.