MASS SPECTRUM ANALYSIS OF A SUBSTANCE SAMPLE PLACED INTO LIQUID SOLUTION

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Yunli Wang

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

June 2011

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Mass Spectrum Analysis of a Substance

Placed in Liquid Solution

By

Yunli Wang

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

ABSTRACT

Wang, Yunli, M.S., Department of Statistics, College of Science and Mathematics, North Dakota State University, June 2011. Mass Spectrum Analysis of a Substance Sample Placed into Liquid Solution. Major Professor: Dr. Volodymyr Melnykov.

Mass spectrometry is an analytical technique commonly used for determining elemental composition in a substance sample. For this purpose, the sample is placed into some liquid solution called liquid matrix. Unfortunately, the spectrum of the sample is not observable separate from that of the solution. Thus, it is desired to distinguish the sample spectrum. The analysis is usually based on the comparison of the mixed spectrum with the one of the sole solution. Introducing the missing information about the origin of observed spectrum peaks, the author obtains a classic set up for the Expectation-Maximization (EM) algorithm. The author proposed a mixture modeling the spectrum of the liquid solution as well as that of the sample. A bell-shaped probability mass function obtained by discretization of the univariate Gaussian probability density function was proposed or serving as a mixture component. The E- and M- steps were derived under the proposed model. The corresponding R program is written and tested on a small but challenging simulation example. Varying the number of mixture components for the liquid matrix and sample, the author found the correct model according to Bayesian Information Criterion. The initialization of the EM algorithm is a difficult standalone problem that was successfully resolved for this case. The author presents the findings and provides results from the simulation example as well as corresponding illustrations supporting the conclusions.

ACKNOWLEDGMENTS

I greatly appreciate the support and help of my major advisor, Dr. Volodymyr Melnykov, in this work. I would also like to thank Dr. Rhonda Magel for her assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Statistics is a field of science with many applications in business, agriculture, chemistry and other areas. This paper presents a statistical methodology applicable in chemistry. It can be used to find the mass spectrum of a substance sample placed into a liquid solution. Mass Spectrometry (MS) is a technique commonly used in analytical laboratories that study chemical, biochemical or physical properties of a wide variety of compounds for determining elemental composition in a substance sample. For this purpose, the sample is usually placed into some liquid solution called liquid matrix. Unfortunately, the spectrum of the sample is not observable separately from the spectrum of the liquid solution. Thus, it is desired to separate the sample spectrum from the spectrum of the liquid solution. There are special tools and methods to identify unknown composition of elements in a molecule or chemical solution. To learn more about these and other methods, we refer the readers to [1,17]. In literature [2,3,22,30], there are detailed introductions about the principle of MS and the application of MS instruments. These instruments are widely used in chemistry and physics. However, they are not flexible enough to be used out of laboratory or during a short period. There are various methods of mass spectrum investigating the peaks. However, there is a lack of procedures for extracting the entire mass spectrum of the substance. Therefore, it is necessary to find inexpensive and flexible methods which can be simply applied to separate the spectrum of the substance sample from the spectrum of the liquid solution.

The purpose of this paper is to derive a procedure which can find the mass spectrum of a sample placed into a liquid solution. Here, the author present the methodology executed in the research to find the spectrum of the sample from the spectrum of the liquid solution. The analysis is based on the comparison of the mixed spectrum with the spectrum of the sole liquid solution. However, the origins of all

spectrum peaks are unknown. Therefore, by introducing the missing information about the origin of observed spectrum peaks, the author obtain a classic set up for the Expectation-Maximization (EM) algorithm. A mixture model of the spectrum of the liquid solution as well as that of the sample is proposed. According to the pattern of the spectrum, a bell-shaped probability mass function obtained by discretization of the univariate Gaussian probability density function was designed for serving as a mixture component. The author have tested the methodology in a challenging simulation example in R environment. In Chapter 2, the author will consider necessary background and present the derivations for the technique. Chapter 3 presents a simulation study. Finally, in Chapter 4, the author conclude the paper and discuss the future directions of this research.

CHAPTER 2. METHODOLOGY

2.1. Finite Mixture Models

Since nineteenth century, finite mixture models are applied frequently in modern statistics. More and more scholars in science noticed that finite mixture models bring great flexibility and convenience when multivariate datasets are met in their research. Furthermore, a wide variety of probability distribution functions are learned and used as mixture components in finite mixture models. Diverse algorithms and methods are developed to estimate the unknown parameters. An increasing number of books and research papers [16, 18, 31] have come out introducing and interpreting the definition and application of finite mixture models.

In statistical literature, the first appearance of finite mixture models, which was used for the purpose of modeling outlier, was in paper [27] in 1886. In the paper, however, there was no complete definition and interpretation of finite mixture models. [26] includes the comprehensive explanation and summary of the application and development of finite mixture models. In literature, the mixtures of Gaussian densities are most commonly used and popular [10]. Now, we provide a definition of a finite mixture models.

Definition : Let $X_1, X_2, X_3, \ldots, X_n$ to be independent and identically distributed random variables from a distribution with probability density function $f(x;\pi)$ given by

$$f(x;\pi) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{1}$$

Then, this probability density function represents a distribution of a finite mixture model with K components. This is the most general form of mixture models. Here, K represents the total number of components contained in the mixture. $\pi = (\pi_1, \pi_2, \ldots, \pi_k)'$ is the vector of mixing proportions; the $k-th$ mixing proportion

π_k is the prior probability that an observation belongs to the $k-th$ component. The sum of all mixing proportions from different components must be equal to 1. Therefore, for mixing proportions $\pi_k s$, the restrictions are $0 < \pi_k \leq 1, k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$. Here, the proportions of distinct components in the sample can be equal or completely different. In the form given in equation (1), $f_k(x)$ is called the $k-th$ mixture component or mixing density. $f_k(x)$ represents the probability density function of the $k-th$ component. Mixing densities are usually assumed to have a parametric form. The functional form of f_k can be different or the same for different components and is assumed to be known. In the parametric form, unknown parameters of each mixture component need to be estimated to define the probability density function. For this reason, we refer to $f(x;\theta)$ given by

$$f(x;\theta) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \tag{2}$$

where, θ is the parameter vector, $\theta = (\pi', \theta_1', \theta_2', \ldots, \theta_k')'$, with θ_k representing the unknown parameters corresponding to the $k-th$ functional form of f_k. Here, the mixing proportions are also included into the vector of unknown parameters. For the future derivation and computation, it is convenient to use one vector θ to represent all unknown parameters in the form of probability density function.

Finite mixture models can be applied to various problems. In particular, it provides a convenient formal setting for model-based clustering whose purpose is to classify homogeneous observations into groups. In model-based clustering, each of the observations is assigned to different groups according to some pre-specified rule. Let sample $X_1, X_2, \ldots, X_n$ be drawn from the parametric mixture model (2). Observations from the $k-th$ group have the mixture component $f_k(x;\theta_k)$ with the corresponding mixing proportion π_k. To assign observations to clusters, the author use Bayes rule based on their obtained posterior probabilities.

In form (2), when the number of components K is unknown, it has to be estimated. In many applications of finite mixture models, the author assume that the probability density function of mixture components is the same for all clusters. In this paper, the author also relay on this assumption. Estimating unknown parameters of mixture components is an important statistical problem. the author discusses how to find posterior probabilities and estimate unknown parameters in the next section.

2.2. Expectation-Maximization (EM) Algorithm

In order to estimate the unknown parameters of the mixture components, the author needs to maximize the likelihood function which is constructed based on the probability density function (2). The estimates are called maximum likelihood estimates (MLEs). However, in finite mixture models, maximum likelihood (ML) estimation is difficult to implement because the form of the likelihood function form (2) is typically complicated and multi-modal. Obtaining the closed-form solution or conducting numerical optimization of the direct likelihood function is impossible or troublesome. Fortunately, ML estimation can be implemented via the EM algorithm [4,15]. The EM algorithm is an efficient method for estimation in finite mixture model setting. The EM algorithm assumes that there are missing observations called group identifiers. It is an iterative procedure that allows estimating unknown parameters θ. It iteratively alternates between the expectation step called E-step and the maximization step called M-step. In finite mixture models, the corresponding complete-data log likelihood function usually can be easily maximized. Then, at E-step, the EM algorithm computes the expected log likelihood for the complete data, denoted as $Q-$ function, and obtains the posterior probabilities π_{ih}.

$$\pi_{ih}^{(r)} = Prob\{X_i \in h-th \quad cluster | X_i; \theta^{(r-1)}\} = \frac{\pi_k^{(r-1)} f(x_i; \theta_k^{(r-1)})}{\sum_{h=1}^{K} \pi_h^{(r-1)} f(x_i; \theta_h^{(r-1)})} \qquad (3)$$

Here, r is the number of iteration. At the M-step, the algorithm maximizes the $Q-$ function $Q(\theta; \theta^{(r-1)}, x_1, x_2, \ldots, x_n)$ with respect to the parameter vector θ to re-estimate all parameters. Once the author has new parameter values, the author repeats E and M-steps until the likelihood converges. In this paper, the EM algorithm is implemented in the proposed model. The author obtained the closed-form solutions for the majority of unknown parameters. The means and variances of different mixture components need to be optimized numerically. In section 2.3, the author discusses an appropriate model for the spectrum modeling. In section 2.4, the solutions for the estimates of parameters are presented.

2.3. Proposed Model

In the research, the author wants to separate the spectrum of the sample from the spectrum of liquid. This can be done by comparing the spectrum of the sole liquid with the mixed spectrum. However, the author does not know the number of components needed to model both spectrum. The functional form of mixture component has to be proposed based on the pattern in the spectrum. As mentioned before, the author assumes that the functional form of mixture components is the same for all components in the paper.

In Figure 1., the X-axis represents the location of observations, while Y-axis represents the height of peaks which means the number of observations at each location. For instance, at location 8, there are almost 100 observations located at the same point.

After taking a look at the mass spectrum in Figure 1., the author can notice some unusual patterns and multiple local modes. Thus, standard distributions cannot be applied as mixture components for modeling mass spectrums. The observations in the spectrum are discrete. For the purpose of finding an appropriate probability mass function for mixture components, the author considered several standard discrete

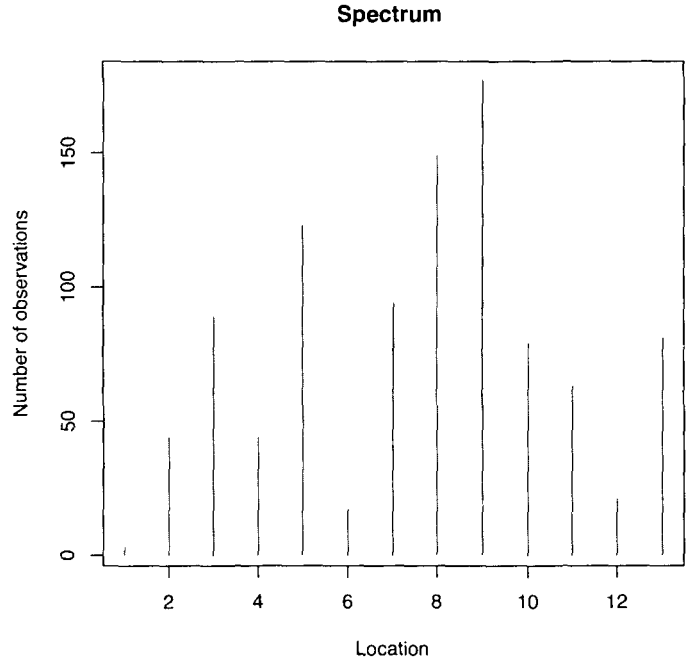

Figure 1. Example of mass spectrum

distributions such as Poisson, Binomial and negative Binomial distribution. Neither of them is appropriate for modeling the mixture components in the case of study. The bell-shaped patterns in spectrum remind a Gaussian distribution. Furthermore, mixtures of Gaussian densities are the most commonly used in finite mixture models. the author proposes the discretization of the univariate Gaussian probability density function to obtain a bell-shaped probability mass function $f(x; \mu_k, \sigma_k^2)$ given by

$$f(x; \mu_k, \sigma_k^2) = \Phi(\frac{x + 0.5 - \mu_k}{\sigma_k}) - \Phi(\frac{x - 0.5 - \mu_k}{\sigma_k}), x = 0, \pm 1, \pm 2, \ldots. \qquad (4)$$

Here, $\mu_k$and σ_k represent the mean and also the standard deviation in the $k-th$ component from univariate Gaussian distribution, and Φ represents the cumulative distribution function of the standard normal distribution. Figure 2. shows how Gaussian density is used to produce a bell-shaped probability mass function.

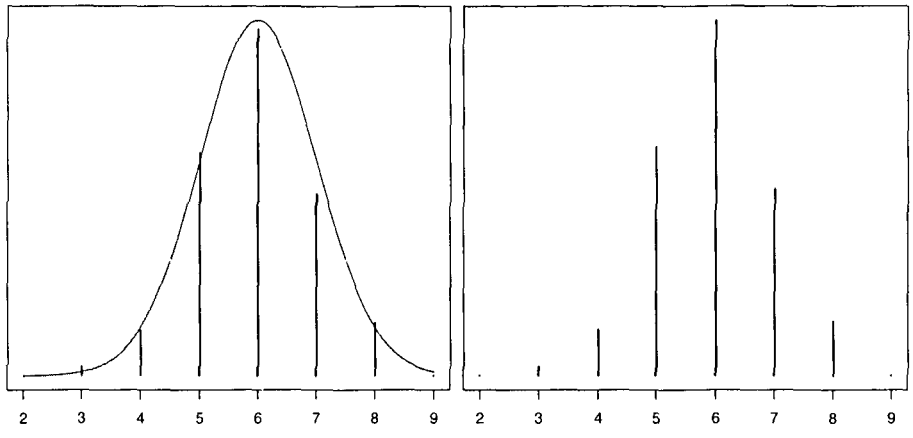

Figure 2. Illustrations of how to use univariate Gaussian probability density function to obtain the discrete probability mass function

Since the author compares the spectrum of the liquid and mixed spectrum, the author must work with two independent samples simultaneously. The author also needs to estimate the number of components in both specified spectrums. The following mixture models are proposed for the spectrum of liquid and mixed spectrum, respectively.

$$X_1, X_2, \ldots, X_{n_x} \sim^{iid} h(x) = \sum_{k=1}^{K} \pi_k f(x_i; \mu_k, \sigma_k^2) \tag{5}$$

$$Y_1, Y_2, \ldots, Y_{n_y} \sim^{iid} g(y) = c \sum_{k=1}^{K} \pi_k f(y_j; \mu_k, \sigma_k^2) + \sum_{m=1}^{M} \gamma_m f(y_j; \upsilon_m, \tau_m^2) \tag{6}$$

Expression (5) represents the proposed model for the spectrum of the liquid. n_x is the number of observations in the spectrum of liquid , while n_y is the number of observations in the mixed spectrum. Unknown parameters μ_k, σ_k^2 are mean and variance of the $k-th$ mixture component of the liquid from the univariate Gaussian distribution. The expression (6) is the proposed model for the mixed spectrum when the substance is placed into the liquid. Unknown parameters υ_m, τ_m^2 represent

the mean and variance of $m-th$ mixing component of the substance in the mixed spectrum. In section 2.1, the author has mentioned that, in finite mixture models, the sum of all mixing proportions should be equal to 1, $\sum_{k=1}^{K} \pi_k = 1$. For (6), a similar condition has to be implemented. The constant c is introduced to guarantee that $c\sum_{k=1}^{K} \pi_k + \sum_{m=1}^{M} \gamma_m = 1$. Thus, the sum of all mixing proportions in the mixed spectrum is restricted to be 1. Then, the author can derive the expression for the constant c: $c = 1 - \sum_{m=1}^{M} \gamma_m$. Now, the bell-shaped probability mass function for the mixture component is proposed, and the unknown parameters and number of components are estimated.

2.4. Parameter Estimation

The unknown parameters need to be estimated from the likelihood function based on the functional form of both proposed models. The likelihood function is obtained by considering the spectrum of liquid and the mixed spectrum simultaneously. Thus, the likelihood function is given by:

$$L(\theta) = \prod_{i=1}^{n_x} \sum_{k=1}^{K} \pi_k f(x_i; \mu_k, \sigma_k^2) \times \prod_{j=1}^{n_y} [\sum_{k=1}^{K} c\pi_k f(y_j; \mu_k, \sigma_k^2) + \sum_{m=1}^{M} \gamma_m f(y_j; v_m, \tau_m^2)] \quad (7)$$

The log-likelihood function obtained from equation (7) is multi-modal. Also, it is hard to take partial derivatives of the maximum log-likelihood function directly. Usually, deriving the closed-form solutions for unknown parameters is complicated or not possible. Because the EM algorithm is flexible and available for treating complicated, multi-modal incomplete data, by introducing the missing information about the group

identifiers, the complete-data likelihood function L_c can be obtained:

$$\begin{aligned} L_c(\theta) = & \prod_{i=1}^{n_x} \prod_{k=1}^{K} [\pi_k f(x_i, \mu_k, \sigma_k^2)]^{I(x_i \in k^{th})} \\ & \times \prod_{j=1}^{n_y} \{ \prod_{k=1}^{K} [c\pi_k f(y_j; \mu_k, \sigma_k^2)]^{I(y_j \in k^{th})} \\ & \times \prod_{m=1}^{M} [\gamma_m f(y_j; v_m, \tau_m^2)]^{I(y_j \in m^{th})} \} \end{aligned} \tag{8}$$

where the author assume that the origin of each peak is known. $I(x_i \in k^{th})$ is the indicator function that the $i-th$ peak belongs to the $k-th$ component in the spectrum of the sole liquid and $I(y_j \in k^{th})$ indicates that the $j-th$ peak is from the $k-th$ component of the model for the mixed spectrum. Similarly, $I(y_j \in m^{th})$ indicates that the $j-th$ peak belongs to the $m-th$ component of the spectrum of the model for the mixed spectrum.

Next, the author can obtain the corresponding complete-data log-likelihood function $l_c(\theta)$:

$$\begin{aligned} l_c(\theta) = log L_c(\theta) = & \sum_{i=1}^{n_x} \sum_{k=1}^{K} I(x_i \in k^{th}) [log\pi_k + log f(x_i; \mu_k, \sigma_k^2)] \\ & + \sum_{j=1}^{n_y} \{ \sum_{k=1}^{K} I(y_j \in k^{th}) [log c\pi_k + log f(y_j; \mu_k, \sigma_k^2)] \\ & + \sum_{m=1}^{M} I(y_j \in m^{th}) [log\gamma_m + log f(y_j; v_m, \tau_m^2)] \} \end{aligned} \tag{9}$$

From EM algorithm, the expectation of the conditional complete-data log-likelihood function given observed data is obtained at the E-step. Thus, the $Q-$

function is given by

$$Q(\theta) = E(l_c(\theta)|x_1, \ldots, x_{n_x}, ; y_1, \ldots, y_{n_y}) = \sum_{i=1}^{n_x} \sum_{k=1}^{K} \pi_{ik}[log\pi_k + logf(x_i; \mu_k, \sigma_k^2)]$$
$$+ \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^{K}[logc\pi_k + logf(y_j; \mu_k, \sigma_k^2)] + \sum_{j=1}^{n_y} \sum_{m=1}^{M} \gamma_{jm}^{M}[log\gamma_m + logf(y_j; v_m, \tau_m^2)] \tag{10}$$

where π_{ik} is the posterior probability that observation x_i belongs to the $k-th$ component of the spectrum of the sole liquid. γ_{jk}^{K} is the posterior probability that observation y_j originates from the $k-th$ component of the liquid in the mixed spectrum, and γ_{jm}^{M} is the posterior probability that observation y_j belongs to the $m-th$ component of the sample in the mixed spectrum.

The posterior probabilities can be estimated assuming that the parameters from the functional form of the proposed models are known. Below, r represents the iteration number of the EM algorithm. $\theta^{(r-1)}$ is the parameter vector estimates calculated at the $r-th$ iteration.

$$\begin{aligned} \pi_{ik}^{(r)} &= Prob\{X_i \in k-th \quad component|X_i; \theta^{(r-1)}\} \\ \gamma_{jk}^{K(r)} &= Prob\{Y_j \in k-th \quad component|Y_j; \theta^{(r-1)}\} \\ \gamma_{jm}^{M(r)} &= Prob\{Y_j \in m-th \quad component|Y_j; \theta^{(r-1)}\} \end{aligned} \tag{11}$$

The E-step consists of updating the posterior probabilities $\pi_{ik}^{(r)}$, $\gamma_{jk}^{K(r)}$ and $\gamma_{jm}^{M(r)}$ given the current parameter estimate $\theta^{(r-1)}$ for all $r = 1, 2, 3, \ldots$.

The posterior probabilities can be calculated as follows:

$$\pi_{ik}^{(r)} = \frac{\pi_k^{(r-1)} f(x_i; \mu_k^{(r-1)}, \sigma_k^{2(r-1)})}{\sum_{k'=1}^{K} f(x_i; \mu_{k'}^{(r-1)}, \sigma_{k'}^{2(r-1)})} \tag{12}$$

$$\gamma_{jk}^{K(r)} = \frac{c^{(r-1)}\pi_k^{(r-1)} f(y_j; \mu_k^{(r-1)}, \sigma_k^{2(r-1)})}{\sum_{k'=1}^{K} c^{(r-1)}\pi_{k'}^{(r-1)} f(y_j; \mu_{k'}^{(r-1)}, \sigma_{k'}^{2(r-1)}) + \sum_{m'=1}^{M} \gamma_{m'} f(y_j; v_{m'}^{(r-1)}, \tau_{m'}^{2(r-1)})} \tag{13}$$

$$\gamma_{jm}^{M(r)} = \frac{\gamma_m^{(r-1)} f(y_j; \mu_m^{(r-1)}, \sigma_m^{2(r-1)})}{\sum_{k'=1}^{K} c^{(r-1)}\pi_{k'}^{(r-1)} f(y_j; \mu_{k'}^{(r-1)}, \sigma_{k'}^{2(r-1)}) + \sum_{m'=1}^{M} \gamma_{m'}^{(r-1)} f(y_j; v_{m'}^{(r-1)}, \tau_{m'}^{2(r-1)})} \tag{14}$$

At the M-step, $Q-$ function is maximized with respect to the parameters. Then, the author can consider several simpler versions of the $Q-$ function depending on parameters $\pi_k, \gamma_m, c, \mu_k, \sigma_k^2, v_m$ and τ_m^2 , with respect to the which the author maximizes the function.

First, the author derives the closed-form solution for parameters π_k, γ_m, and c. Since the author has two restrictions on mixing proportions, the author needs to introduce two Lagrange multipliers: λ_1 and λ_2 . After the author finds the expression for π_k, γ_m, and c can be obtained by $c = 1 - \sum_{m=1}^{M} \gamma_m$.

Therefore, the Q^*- function of the interest that needs to be maximized over π_k, γ_m and c is given by

$$\begin{aligned} Q^*(\theta) = \sum_{i=1}^{n_x}\sum_{k=1}^{K} \pi_{ik} log \pi_k + \sum_{j=1}^{n_y}\sum_{k=1}^{K} \gamma_{jk}^{K} log c \pi_k + \sum_{j=1}^{n_y}\sum_{m=1}^{M} \gamma_{jm}^{M} log \gamma_m \\ -\lambda_1(\sum_{k=1}^{K} \pi_k - 1) - \lambda_2(\sum_{m=1}^{M} \gamma_m - 1 + c) + constant \end{aligned} \tag{15}$$

Here, the constant does not affect the derivation for π_k, γ_m,and c. Based on this Q^*- function, the author continuess taking partial derivatives with respect to $\pi_k, \gamma_m, and c.$, separately. First, λ_1 and λ_2 need to be estimated to help us achieve the closed-form solutions for π_k, γ_m, and c. The derivations is shown as below:

$$\frac{\partial Q^*}{\partial \pi_k} = \frac{\sum_{i=1}^{n_x} \pi_{ik}}{\pi_k} + \frac{\sum_{j=1}^{n_y} \gamma_{jk}^K}{\pi_k} - \lambda_1 \tag{16}$$

Thus, $\lambda_1 = \sum_{i=1}^{n_x} \sum_{k=1}^{K} \pi_{ik} + \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K$. Next, the author obtains the solution for π_k . Substituting the expression for λ_1 into the equation (16), the following expression can be obtained:

$$\frac{\sum_{i=1}^{n_x} \pi_{ik}}{\pi_k} + \frac{\sum_{j=1}^{n_y} \gamma_{jk}^K}{\pi_k} - \sum_{i=1}^{n_x} \sum_{k=1}^{K} \pi_{ik} - \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K = 0 \tag{17}$$

In the spectrum of the liquid, the sum of probabilities that each peak assigned to different components is equal to 1. So, the author can find the following equation: $\sum_{k=1}^{K} \pi_{1k} = \sum_{k=1}^{K} \pi_{2k} = \ldots \sum_{k=1}^{K} \pi_{ik} = 1$. Then, the author can conclude that $\sum_{i=1}^{n_x} \sum_{k=1}^{K} \pi_{ik} = n_x$. For this reason, the equation (17) can be simplified to:

$$\frac{\sum_{i=1}^{n_x} \pi_{ik}}{\pi_k} + \frac{\sum_{j=1}^{n_y} \lambda_{jk}^K}{\pi_k} - n_x - \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K = 0 \tag{18}$$

Finally, the convenient closed-form solution for $\pi_k^{(r)}$ can be obtained:

$$\pi_k^{(r)} = \frac{\sum_{i=1}^{n_x} \pi_{ik}^{(r)} + \sum_{j=1}^{n_y} \gamma_{jk}^{K^{(r)}}}{n_x + \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^{K^{(r)}}} \tag{19}$$

By taking partial derivatives of Q^*- function with respect to γ_m, c, λ_2 and setting the derivatives equal to zero, two equations are given by

$$\frac{\partial Q^*}{\partial \gamma_m} = \frac{\sum_{j=1}^{n_y} \gamma_{jm}^M}{\gamma_m} - \lambda_2 = 0$$
$$\frac{\partial Q^*}{\partial c} = \frac{\sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K}{c} - \lambda_2 = 0 \tag{20}$$

For these equations, the author goes further to simplify them and obtain the following results:

$$\gamma_m = \frac{\sum_{j=1}^{n_y} \gamma_{jm}^M}{\lambda_2}$$
$$c = \frac{\sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K}{\lambda_2} \tag{21}$$

The author derived the closed-form solutions for γ_m, c and λ_2 by combining the above equations together. The author can find that λ_2 is the number of peaks in the mixed spectrum. Thus, the solution for λ_2 is given by

$$\lambda_2 = \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K + \sum_{j=1}^{n_y} \sum_{m=1}^{M} \gamma_{jm}^M = n_y \tag{22}$$

Then, the parameters $c^{(r)}$ and $\gamma_m^{(r)}$ can be estimated by the following expressions:

$$c^{(r)} = \frac{\sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^{K^{(r)}}}{n_y}$$
$$\gamma_m^{(r)} = \frac{\sum_{j=1}^{n_y} \gamma_{jm}^{M^{(r)}}}{n_y} \tag{23}$$

As we can see, the M-step can provide us with convenient closed-form solutions for π_k, γ_m and c:

The unknown parameters that still have to be estimated are μ_k, σ_k^2, v_m, and τ_m^2.

However, for maximizing $Q-$ function with respect to μ_k, σ_k^2, and v_m, τ_m^2 separately, we can consider two versions of Q, given by the following forms:

$$Q_1^*(\theta) = \sum_{i=1}^{n_x} \sum_{k=1}^{K} \pi_{ik} log f(x_i; \mu_k, \sigma_k^2) + \sum_{j=1}^{n_y} \sum_{k=1}^{K} \gamma_{jk}^K log f(y_j; \mu_k, \sigma_k^2) + constant \quad (24)$$

$$Q_2^*(\theta) = \sum_{j=1}^{n_y} \sum_{m=1}^{M} \gamma_{jm}^M log f(y_j; v_m, \tau_m^2) + constant \quad (25)$$

Unfortunately, the closed-form solutions for the parameters $\mu_k, \sigma_k^2, v_m$and τ_m^2 cannot be obtained. These two Q^*- functions need to be maximized numerically. The author used an R-function optim( ) to conduct numerical optimization.

2.5. Spectrum Estimation

How can we pick the peaks which belong to the spectrum of the sample out of the mixed spectrum? We do not know the number of components in the model. We do not know which peaks come from the spectrum of the substance either. Therefore, the optimal number of components in the mixed spectrum should be found. First, we can find the number of components for model of the liquid. Then, using the estimated number of components in the liquid solution model, we can find the number of components of the substance sample in the mixed spectrum. The author assign observations to cluster according to the largest posterior probabilities γ_{jk}^K and γ_{jm}^M . In Section 2.3, the author have provided the function (4), which is an appropriate choice for being the functional form of the mixture component in the framework of the problem. Also, in Section 2.4, by the EM algorithm, the closed-form solutions for unknown parameters are obtained. The author incorporate the probability mass function (4) into the proposed model to be the mixture component. Eventually, by the final results for posterior probabilities, the observations can be allocated to their

estimated components. Then, we can separate the spectrum of the sample from the liquid spectrum.

2.5.1. Model Selection

Model selection in finite mixture models has often referred to the problem of choosing the optimal number of components [13,14]. In this paper, we have $K + M$ components in the mixed spectrum. First, K needs to be estimated. In this section, we briefly review the history related to choosing the optimal number of components in mixture models. There is vast literature [20,23,32] contributing to the problem. We refer to [6] who provide a detailed description of different and available approaches to address this issue. According to [9,28], most methods devoted to estimating the number of components can be divided into two categories. One group of methods is parsimony-based while another category depends on testing procedures. The former has been widely used and discussed by [25]. In this paper, the method we employ belongs to the parsimony-based category. The majority of parsimony-based approaches choose a number of components to minimize a penalized negative log likelihood function by trying different values of the number of components. A variety of information-based criteria such as Akaike Information Criterion (AIC) [5], Bayesian Information Criterion (BIC) [29] and their modifications fall into this category. BIC has been known in finite mixture models for demonstrating good performance. It also can be implemented easier than many other methods in this group. In this paper, the author use BIC to select the best model. The author vary the values of M and K in order to minimize the value of BIC. The computational form of BIC is $BIC = -2 \times [logL(\theta)] + p \times log(n)$, where p is the number of parameters, $p = 3K + 3M - 1$, and n is the total number of observations from all peaks in the liquid matrix and the mixed matrix. The values of M and K should be chosen so that they correspond to the smallest value of BIC. They represent the optimal numbers

of components according to BIC. $l(\theta)$ is log-likelihood function which is given by the following form:

$$l(\theta) = logL(\theta) = \sum_{i=1}^{n_x} log \sum_{k=1}^{K} [\pi_k f(x_i; \mu, \sigma_k^2)] + \sum_{j=1}^{n_y} log \sum_{k=1}^{K} [\pi_k f(y_j; \mu_k, \sigma_k^2)] + \sum_{j=1}^{n_y} log \sum_{m=1}^{M} [\gamma_m f(y_j; v_m, \tau_m^2)] \tag{26}$$

2.5.2. Groups Identifiers

Group identifier is the basic information needed to allocate the peaks into different groups. By the EM algorithm, at E- and M-steps, the author obtained the parameter estimates of $\pi_{ik}, \gamma_{jk}^K, \gamma_{jm}^M, \pi_k$ and γ_m. Therefore, the posterior probabilities γ_{jk}^K and γ_{jm}^M are available for the group identifiers. The spectrum matrix has two columns. The first one presents the peaks heights and the other one provides the locations of peaks. Using posterior probabilities γ_{jm}^M multiplied by the heights of the peaks (number of observations), we can estimate how many observations came from the components of the spectrum of the sample. Based on the Bayes rule, observations are assigned to these groups according to the highest posterior probability. Under some conditions, there might be several posterior probabilities with the same value, so it is unclear how to assign an observation to a group. In [24], randomization is recommended to break the ties among competing clusters

Thus, the author implements the EM algorithm and allocating the peaks to the estimated clusters. The author starts the procedure by finding reasonable starting values. We also need to specify the stopping rule for detecting the convergence of the EM algorithm.

2.6. Initialization and Stopping Criterion

The initialization of the EM algorithm is a step that might be challenging to implement in research. We need to select the best possible starting values to make

sure that we can obtain the correct estimates of parameters and the right number of components. But how do we decide on which combination of starting values is good? Fortunately, the likelihood function is a convenient tool that can be used for both, initialization and stopping steps.

2.6.1. Initialization of the EM Algorithm

Good initialization strategy is crucial for finding ML estimators. In papers of [11,19], many different initialization procedures have been mentioned and considered. However, there is no any single method that can outperform the others in all cases. A model-based hierarchical clustering approach is seen to work well when components are well-separated [21], but not as well in other situations. This method is proposed by [7] and included in the R package Mclust [12], and is specifically designed for Gaussian mixtures. This hierarchical clustering method in initialization is very restrictive for larger datasets.

The EM algorithm is an iterative and hill-climbing procedure whose performance can depend on particular starting observations called central points. Choosing and assigning observations to the closest central points in the initialization is an important step. The initialization of the EM algorithm involves starting from the central points and running the EM algorithm until the pre specific convergence criterion is satisfied. The EM algorithm implemented to find the best starting values in the initialization is usually call short em. A good choice of central points increases the chance to find the correct parameter estimates. Random selection of initial points may be a bad approach, because we may pick several initial central observations from the same component. Therefore, a pre specified rule is necessary to choose the best initial observations. The author will discuss this rule later in the next section. Once we obtain the best initial observations, representing the central points of each cluster, we continue allocating the observations to groups and proceed to find the unknown

parameters. At the short em, the solution producing the highest log likelihood is chosen as a starter for the long EM algorithm. Then the long EM algorithm runs until the convergence is achieved. Thus, the em-EM algorithm consists of two EM stages proposed by [8]. In this paper, the author use the locations of the peaks instead of considering all observations to find the best starting central locations. Then, based on the distance between pair wise locations of the peaks, all observations are assigned to their nearest group. After the short em is finished, the author saves the best locations and the corresponding parameters estimates: $\pi_k, \gamma_m, c, \mu_k, \sigma_k^2, v_m$and τ_m^2 as the starting parameters. The starting values are passed into the long EM algorithm to get the final parameter estimates. The initialization of the EM algorithm in this paper is a challenging step because the information about the spectrum of the liquid is also unknown. Therefore, first, the author runs the short em for the spectrum of the sole liquid to obtain the estimates of π_k, μ_k, σ_k^2. At the short em and the long EM steps, specific stopping criterion helps us decide when the EM algorithm should be stopped.

2.6.2. Stopping Criterion

The log-likelihood function is used in the convergence criterion to select the best starting points. If the difference between two consequent log-likelihood values is less than some pre specified error margin, we stop the short em algorithm and continue running the long EM with the best starting values of the parameters. After running the long EM algorithm, the main stopping criterion is also based on the log- likelihood function. When the long EM algorithm converges, we save the produced parameter estimates. Meanwhile, the maximized log-likelihood function is used for calculating the values of BIC for different numbers of components.

CHAPTER 3. SIMULATION EXAMPLE

In the paper, the author applied the methodology to a small but challenging simulated dataset. In this simulation example, the author uses a 3-component model to simulate 1000 observations from the liquid. the author also performs simulations of 1000 observations under the mixed spectrum model assuming 2 components for the substance sample. 500 observations were simulated from the model for the liquid solution. The following tables summarize the data simulated with an underlying Gaussian mixture distribution: $0.5 \times 0.5 \times N(6,1) + 0.5 \times 0.2 \times N(9,0.5) + 0.5 \times 0.3 \times N(14,1) + 0.3 \times N(7.5,0.5) + 0.2 \times N(12.2,0.5)$. Table 1. provides simulated data for the liquid solution. In Table 2., we can find simulated data for the mixed spectrum when the sample is placed into the liquid.

Table 1. Liquid matrix data

Location	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Peak Height	3	26	123	194	123	77	94	55	5	17	63	124	81	14	1

Table 2. Liquid matrix data+sample data

Location	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Peak Height	0	13	61	118	197	160	76	26	31	102	100	70	33	12	1

Figure 3. shows the spectrum of the liquid from the simulation example. Where, the x-axis represents the location of the peaks and the y-axis demonstrates the number of observations at each location in the liquid. Figure 4. represents the mixed spectrum from the simulation example when the sample is placed into the liquid. According to Figures 3. and 4., the difference between two spectrums is obvious, but the author cannot visually detect the number of components needed for fitting both datasets.

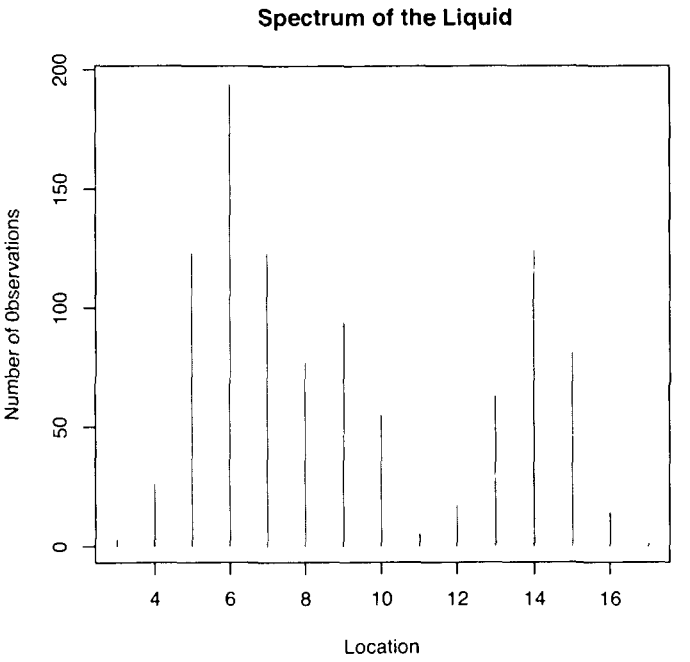

Figure 3. The spectrum of the liquid from the simulation example

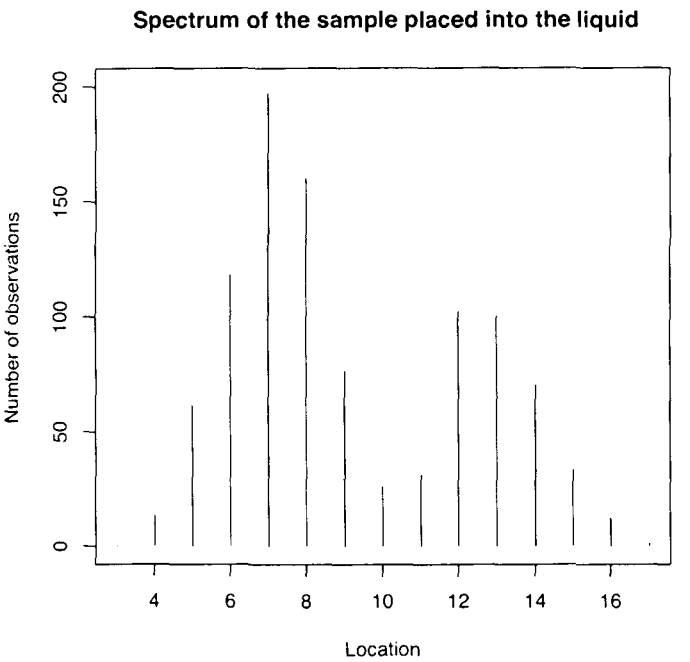

Figure 4. The spectrum of the mixed solution from the simulation example

The author implemented the methodology in R environment and applied it for the simulated data. Although the initialization of the EM algorithm is a difficult problem, we solved it successfully in the simulation example. As mentioned in section 2.5.1, by varying the number of mixture components for the liquid matrix and the sample, we obtained the correct number of components according to BIC. The following Table 3. represents the different values of BIC corresponding to the various values of the number of mixture components K and M.

Table 3. BIC values of different models

BIC	M=1	M=2	M=3	M=4	M=5	M=6
K=1	10379.97	10110.85	10130.63	10149.60	10174.83	10192.92
K=2	9524.315	9508.549	9528.966	9538.091	9564.244	9584.943
K=3	9518.027	9397.326	9419.563	9445.313	9467.393	9493.468
K=4	9544.071	9444.088	9444.636	9466.44	9492.942	9513.477

Based on the values of BIC in Table 3., when $K = 3$ and $M = 2$, BIC reaches its smallest value: 9397.326. Therefore, we can conclude that there are 5 total mixture components which consist of 3 components from the liquid and 2 from the sample in the mixed spectrum. Thus, we are able to detect the correct model from the simulated data. Figures 5. and 6. illustrate the original spectrums along with the predicted values obtained based on the chosen model. As we can see, the model does an excellent job in predicting the peak heights.

The posterior probabilities γ_{jm}^{M} which is obtained from the EM algorithm specify proportion of particles from the substance in the $j-th$ peak of the mixed spectrum. Therefore, the particle counts of observations or the heights of peaks for the sample can be estimated using the total number of observations multiplied by the probabilities of the sample at each location in the mixed spectrum. The formula used is given: $\sum_{m=1}^{M} \gamma_{jm}^{M} \times n_{jy}$. Here, n_{jy} is the number of the observations in the $j-th$ peak from the

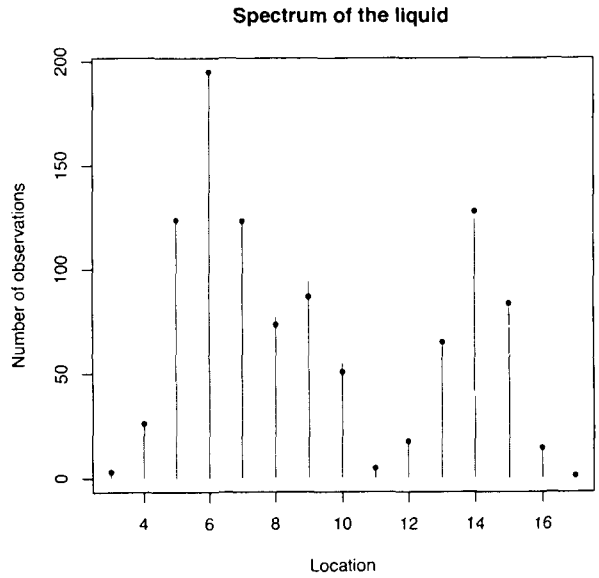

Figure 5. The points represent the predicted values of the number of observations at each location in the liquid spectrum

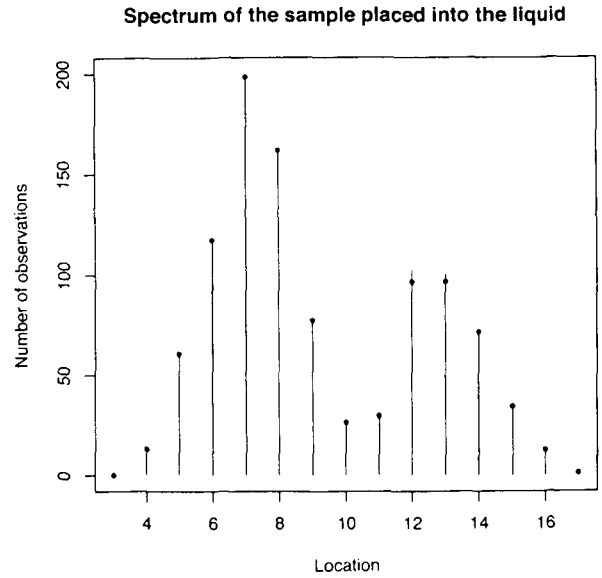

Figure 6. The points represent the predicted number of observations at each location in the mixed spectrum

mixed spectrum. Eventually, the author successfully finds the mass spectrum of the substance sample placed into the liquid. The correct number of components has been detected. Figure 7. represents the extracted sample spectrum. The author applies parametric bootstrapping to obtain the variability of the MLEs for the heights of the peaks in the spectrum of the substance. The author simulated 1000 datasets from random estimated mixtures.

The author ran the EM algorithm and then saved the posterior probabilities for all 1000 datasets. Then, we can obtain the 1000 different combinations of the heights for all peaks for the spectrums of the sample. Finally, the author dropped the first 25 minimum values of the heights out of all peaks and the largest 25 maximum values of the heights of all peaks for the spectrum of the sample. So, the author present the variability of MLEs for the heights of peaks in Figure 8., using 95% confidence intervals. In Figure 8., the dotted lines represent the 95% confidence intervals for the numbers of observations at every location of the simulated spectrum, where the circles are the predicted number of observations at each location.

In order to check the performance of the proposed method on other datasets and more challenging parameter settings, the author considered $c = 0.1, c = 0.5, c = 0.9$. For each value of c, 5 datasets were simulated. For each dataset, the entire analysis has been repeated. In all 15 cases, we successfully detected the number of components under liquid matrix and mixed spectrum mixture models. Thus, the author can conclude that the procedure is relatively robust to changes in c and can be used even for small concentrations of substance in the liquid. In the simulation example, the author successfully extracted the spectrum of the sample from the mixed spectrum. The author finds that there are two components of the sample in the mixed solution according to the BIC based model selection. The author present the variability in MLEs for the number of observations from the substance sample at each location.

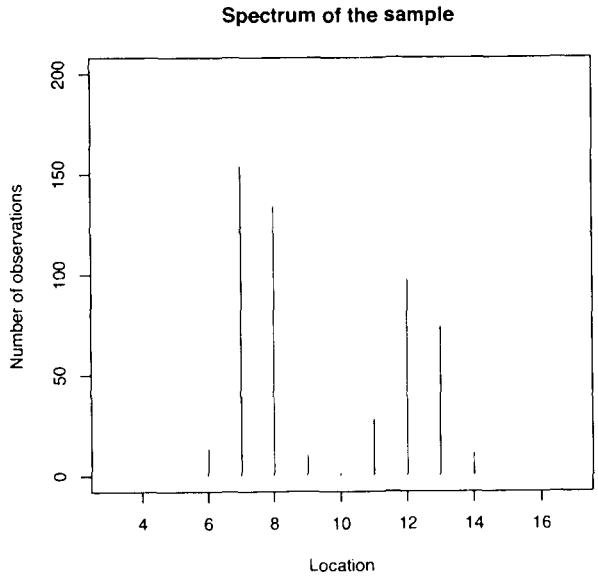

Figure 7. The substance spectrum extracted from the mixed spectrum

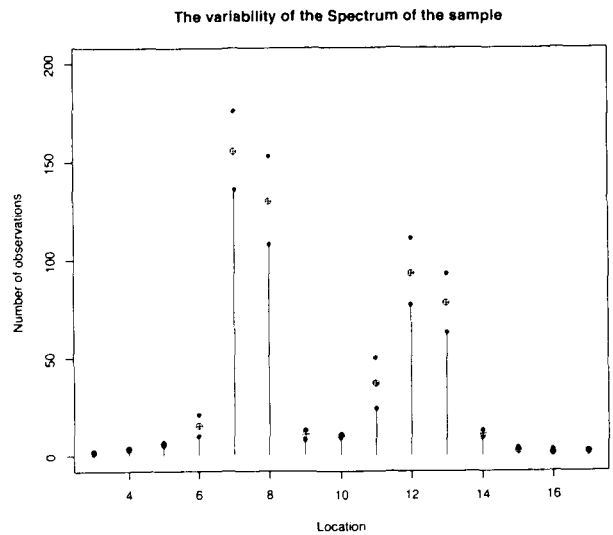

Figure 8. The variability of the estimated height of each peak according to bootstrap

The results of the simulation example indicate that the bell-shaped probability mass function which we have proposed is appropriate and can be employed. Furthermore, the author concludes that the statistical methodology works well for extracting the entire mass spectrum of the substance sample from the mixed spectrum.

CHAPTER 4. DISCUSSION

MS is the technique used to determine the elemental composition in a substance sample. Usually, the substance sample is placed into some liquid solution. Therefore, the origin of the peaks from the obtained mixed spectrum is unknown. In order to separate the spectrum of the sample from the spectrum of the liquid solution, the author propose and implement a flexible statistical model formulated in term of finite mixtures. Then, the EM algorithm is used for the purpose of maximum likelihood estimation. Although, the initialization of the EM algorithm is a difficult problem, the author propose a strategy that successfully resolves all issues. The author demonstrate that the methodology can successfully separate the spectrum of the substance and liquid spectrum. The variability of the obtained estimates of the heights of peaks in spectrums can be assessed by the parametric bootstrap.

The reader can think about using a naive approach for finding the spectrum of the sample that would simply subtract the liquid matrix spectrum from the mixed spectrum. However, this approach is troublesome and cannot provide desired results because the scaling coefficient for mass spectrum peaks is not known. Also, the obtained spectrum might have negative peaks that are difficult to interpret. In addition to that, this approach does not provide us with a reasonable model that can be used, for example, for assessing variability.

For the future work, the author would like to analyze several real mass spectrums and compare the results with results obtained by other researchers.

REFERENCES

[1] *Multivariate least-squares methods applied to the quantitative spectral analysis of multicomponent samples*, http://www.opticsinfobase.org/abstract.cfm?id=125821.

[2] *Standard definitions of terms relating to mass spectrometry. a report from the committee on measurements and standards of the american society for mass spectrometry*, Journal of the American Society for Mass Spectrometry **4** (2), 336–348.

[3] *Definition of spectrograph*, Merriam Webster (2008).

[4] N. M. Larid A. P. Dempster and D. B. Rubin, *Maximum likelihood for incomplete data via the em algorithm*, Jounal of the Royal Statistical Society, 1–38.

[5] H. Akaike, *Information theory and an extension of the maximum likelihood principle*, In Second international symposium on information theory **6** (1973), 267–281.

[6] D. R. Anderson, *Model based inference in the life sciences*, Springer.

[7] J. D. Banfield and A. E. Raftery, *Model-based gaussian and non-gaussian clustering*, Biometrics **49** (1993), 803–821.

[8] G. Celeux C. Biernacki and G. Govaert, *Choosing start- ing values for the em algorithm for getting the highest likelihood in mul- tivariate gaussian mixture models*, Computational Statistics and Data Analysis **413** (2003), 561–575.

[9] A. Dasgupta and A. E. Raftery, *Detecting features in spatial point processes with clutter via model-based clustering*, Journal of the American Statistical Association **93** (1998), 294–302.

[10] N. Day, *Estimating the components of a mixture of two normal distributions*, Biometrika **56** (1969), 463–474.

[11] M. A. T. Figueiredo and A. K. Jain, *Unsupervised learning of finite mixture models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002), no. 3, 381–396.

[12] C. Fraley and A. E. Raftery, *Normal mixture modeling and model-based clustering*. Tech. Rep at Department of Statistics of University of Statistics, Seattle, WA (2006).

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

[14] K. Basford G. McLachlan, G. Peel and P. Adams, *Fitting of mixtures of normal and t-components*, Journal of Statistical Software.

[15] T. Krishnan G. McLachlan, *The em algorithm and extensions*, Wiley, New York (1997).

[16] J. Francos H. Permuter and I.H. Jermyn, *Gaussian mixture models of texture and colour for image database retrieval*, IEEE International Conference on Acoustics, Speech, and Signal Processing (2003).

[17] S. A. Lammert, A. A. Rockwood, M. Wang, and M. L. Lee, *Miniature toroidal radio frequency ion trap mass analyzer*, Journal of the American Society for Mass Spectrometry **17** (2006), no. 7, 916–922.

[18] B. G. Lindsay, *Mixture models: Theory, geometry, and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics **5** (1995).

[19] R. Maitra, *Initializing partition-optimization algorithms.*, IEEE/ACM Transactions on Computational Biology and Bioinformatics **6** (2009), no. 144-157.

[20] R. Maitra and V. Melnykov, *Assessing significance in finite mixture models*, Tech. Rep at Department of Statistics of Iowa State University (2010).

[21] R. Maitra and V. Melnykov, *Simulating data to study performance of finite mixture modeling and clustering algorithms.* Journal of Computational and Graphical Statistics (in press).

[22] R. E. March, *Quadrupole ion trap mass spectrometry: a view at the turn of the century*, International Journal of Mass Spectrometry **200**, 285–312.

[23] G. McLachlan, *On bootstrapping the likelihood ratio test statis- tic for the number of components in a normal mixture*, Applied Statistics **36**, 318–324.

[24] G. McLachlan and T. Krishnan, *The em algorithm and extensions*, Wiley, New York **MR1417721**.

[25] G. McLachlan and D. Peel, *Finite mixture models*, John Wiley and Sons, Inc., New York.

[26] G. J. McLachlan and D. Peel, *Finite mixture models*, Wiley **ISBN 0471006262**.

[27] S. Newcomb, *A generalized theory of the combination of observa- tions so as to obtain the best result*, American Journal of Mathematics **8** (1886), 343–366.

[28] K. Roeder and L. Wasserman, *Practical bayesian density estimation using mixtures of normals*, Journal of the American Statistical Association **92** (1997), no. 894-902.

[29] G. Schwarz, *Estimating the dimensions of a model*, Annals of Statistics **6** (1978), 461–464.

[30] Gordon Squires, *Francis aston and the mass spectrograph*, Dalton Transactions **23**, 3893–3900.

[31] D. Titterington, A. Smith, and U. Makov, *Statistical analysis of finite mixture distributions*, John Wiley and Sons **ISBN 0471907634**.

[32] Melnykov V. and Maitra R., *Finite mixture models and model-based clustering*, Statist. Surv. **4** (2010), 80–116.