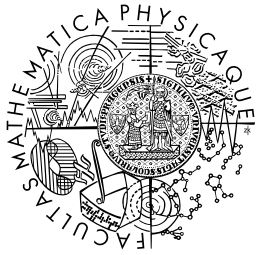

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Adéla Nguyenová

EM algorithm for truncated Gaussian mixtures

Department of Probability and Mathematical Statistics

Supervisor of the master thesis: RNDr. Jiří Dvořák, Ph.D.

Study programme: Mathematics

Study branch: MPMSE

Prague 2022

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to thank my supervisor RNDr. Jiří Dvořák, Ph.D. for sharing his knowledge with me and for providing me a huge support not only throughout the time I was writing my thesis but also throughout my whole studies.

I would love to dedicate this work to Daniel, who stood by my side not only during the beautiful but also the difficult times of my studies and my life. He made me a better person.

I would also like to thank my entire family for their support, I wouldn't have made it this far without them. I am grateful for the endless amount of time I spent with my dad and math during my high school years. I'm not sure I would have fallen in love with math without it.

Last but not least I thank all my close friends for their patience and support, all my colleagues at work and all my teachers who enriched me with valuable knowledge on my academic journey.

Title: EM algorithm for truncated Gaussian mixtures

Author: Adéla Nguyenová

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jiří Dvořák, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The expectation-maximization iterative algorithm is widely used in parameter estimation when dealing with missing information. Such a situation can naturally arise when we observe the data of our interest on a bounded observation window. This thesis focuses on the application of the EM algorithm for truncated Gaussian mixtures and compares the proposed algorithm with the approach in a previously published article, see Lee and Scott [2012], where it uses a heuristic simplification and is not sufficiently supported mathematically. We also compare the behavior of the proposed algorithm with the procedure from the article in a series of simulated experiments, as well as in analyzing a real dataset. We also provide `Python` implementation of the EM algorithm for truncated Gaussian mixtures.

Keywords: EM, Expectation-Maximization Algorithm, Gaussian Mixtures, Truncation

Contents

List of Abbreviations and Notation

$\mathcal{L}_C$	complete-data likelihood function
ℓ_C	complete-data log-likelihood function
ℓ_{obs}	observed-data log-likelihood function
EM	expectation-maximization algorithm
$\mathbb{E}_p$	expectation with respect to the distribution p
$\mathcal{M}^1(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}])$	first moment of Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ truncated to the rectangle window $[\boldsymbol{s}, \boldsymbol{t}]$
$\mathcal{M}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}])$	second moment of Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ truncated to the rectangle window $[\boldsymbol{s}, \boldsymbol{t}]$
$D_{KL}(p\|\|q)$	Kullback-Leibler divergence for distributions p and q
$\det(A)$	determinant of square matrix A
AIC	Akaike's information criterion
BIC	Bayesian information criterion

Introduction

In spatial statistics, we often deal with cluster point processes observed on some restricted observation window. The truncated Gaussian mixture is a simple but powerful model for such situtations. In order to estimate parameters of a mixture, the expectation-maximization algorithm (EM algorithm for short) is often used, see for example the resources we used for this thesis Dempster [1977], McLachlan and Rathnayake [2014], Kushary [1998] or Figueiredo and Jain [2002]. This work derives the version of the EM algorithm for truncated Gaussian mixtures, discusses practical issues arising from fitting the model to data and then studies the properties of the results using both simulated and real data.
This thesis is divided into four chapters. In the first chapter, we discuss the expectation-maximization algorithm for Gaussian mixtures. Firstly, we define a finite mixture distribution via its probability density function, a finite linear combination of probability density functions, and its commonly used example, the Gaussian mixture. Then, we introduce the expectation-maximization algorithm in general. We state two important theorems regarding the EM algorithm. The first one states that the log-likelihood does not decrease in each iteration. The second proves the convergence to a local maximum of the log-likelihood function. Finally, we combine the previous two sections of this chapter and apply the EM algorithm for Gaussian mixtures. By calculating the partial derivatives of the derived Lagrange function, we arrive at the updated parameters obtained in M step of the EM algorithm. We also show the simplified formulas for the updated parameters for homoscedastic components and/or isotropic components.

The second chapter applies the EM algorithm to the truncated Gaussian mixtures, bounded to a rectangular observation window. First we define the truncated Gaussian distribution and calculate its first two moments. Then we extend the defined distribution into a mixture of such distributions. Finally, we can apply the EM algorithm for truncated Gaussian mixtures. We derive the formulas for both steps of the EM algorithm, however we are not able to express the updated parameters in a closed form as for the standard EM algorithm because the unknown parameters occur inside an integral which makes it impossible to separate them from the corresponding equations. The method used in Lee and Scott [2012] is summarised in the next section of this chapter. We describe what simplification has been made in the approach used in this article and try to explain the main reason why this heuristic method has been introduced. We also address the major weak point of such approach: we cannot rely on the theory behind the EM algorithm. In the penultimate subchapter, the effect of truncation on the EM algorithm is shown in the example of a one-dimensional Gaussian mixture with two clusters where one of them is affected by truncation to a considerable degree. The mean estimates in cases when truncation is not considered and in case when truncation is taken into account are shown. At the end of this chapter, we outline how the algorithm can be generalised for an arbitrary observation window.

In the third chapter, we discuss the practical issues related to the EM algorithm. The first of them is the problem of the unknown number of components. We then show a few options on how to initialize the EM algorithm. In the third section, we mention several stopping criteria. In the end we briefly describe the implementation of the EM algorithm in the `Python` programming language.

In the last chapter, we apply the proposed algorithm and selected methods from the third chapter to the simulated datasets and the real dataset. We compare the performance of the proposed algorithm with the approach introduced in Lee and Scott [2012]. Apart from the comparative analysis, we performed experiments showing us some interesting practical results.

One of the main contributions of this master thesis is the detailed derivation of the EM algorithm in application on truncated Gaussian mixtures. Furthermore, it clarifies that Lee and Scott [2012] does not use the EM algorithm for finding estimates of truncated Gaussian mixtures, it uses its heuristic version instead. With such simplification, we cannot rely on the theory behind the EM algorithm. In practical examples, we show the comparison of the proposed rigorous use of the EM algorithm for truncated Gaussian mixtures with the approach described in Lee and Scott [2012]. To perform such a comparison, we implemented both approaches in the `Python` programming language. In addition to the theoretical part, we also provide a Python library capable of simulating truncated Gaussian mixtures (with the help of the `R` library `truncnorm`) in one and two dimensions, as well as performing the proposed EM algorithm on simulated and real datasets.

1. EM algorithm for Gaussian mixture model

We introduce the expectation-maximization algorithm (generally abbreviated as the EM algorithm) for Gaussian mixture data. This is an iterative method for computing the maximum likelihood estimate of the model parameters, widely used not only for incomplete data problems. A Gaussian mixture is a probabilistic model represented by a finite number of Gaussian distributions [1]. At first glance, this does not seem to be an incomplete-data problem however we can formulate it as such as well.

1.1 Gaussian mixture distribution

Often we observe data that are grouped in so-called clusters. This could be due to the similarity between observations belonging to one cluster in some property or location. When analysing such dataset, we would like to classify the data into such groups. There are three major categories of clustering methods which deal with this problem, partitioning algorithms (for example K-means), distance-based algorithms (hierarchical methods) and parametric model-based methods, see Melnykov and Melnykov [2012]. The last is usually based on finite mixture models where we assume a specific distribution of clusters which depends on some unknown parameters. Let us define a finite mixture formally.

Definition 1. *Let $K \in \mathbb{N}$, $\boldsymbol{Y}$ be a d-dimensional real random vector with the probability density function*

$$f(\boldsymbol{y}) = \sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{y}), \quad \boldsymbol{y} \in \mathbb{R}^d, \tag{1.1}$$

where $f_k(\boldsymbol{y})$, $k = 1, \ldots, K$, are probability density functions and $0 \leq \pi_k \leq 1$, $k = 1, \ldots, K$, satisfy $\sum_{k=1}^{K} \pi_k = 1$. Then we say $\boldsymbol{Y}$ is distributed as a finite mixture with K components, $f_k(\boldsymbol{y})$ is the k-th component density function and π_k is the k-th mixing weight.

When dealing with data distributed as a mixture, our aim often is to estimate the mixing weights and component density functions. Usually those component densities $f_k(\boldsymbol{y})$, $k = 1, \ldots, K$, can be parametrized by $\boldsymbol{\theta}_k$, so we will write $f_k(\boldsymbol{y}) = f_k(\boldsymbol{y}; \boldsymbol{\theta}_k)$ for $k = 1, \ldots, K$. Then the probability density function of the mixture can be written as

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{y}; \boldsymbol{\theta_k}), \quad \boldsymbol{y} \in \mathbb{R}^d, \tag{1.2}$$

[1] Note that there is an infinite Gaussian mixture model where the number of clusters tends to infinity which can be useful in many applications where we do not want to limit the number of clusters. Since the condition for the mixing weights remains the same (all weights should sum up to one), they have to follow a distribution which ensures that the majority of clusters will have negligible weight. Then the model will be well defined. See for example Rasmussen [1999]. Nevertheless, we will assume a finite number of clusters in this thesis.

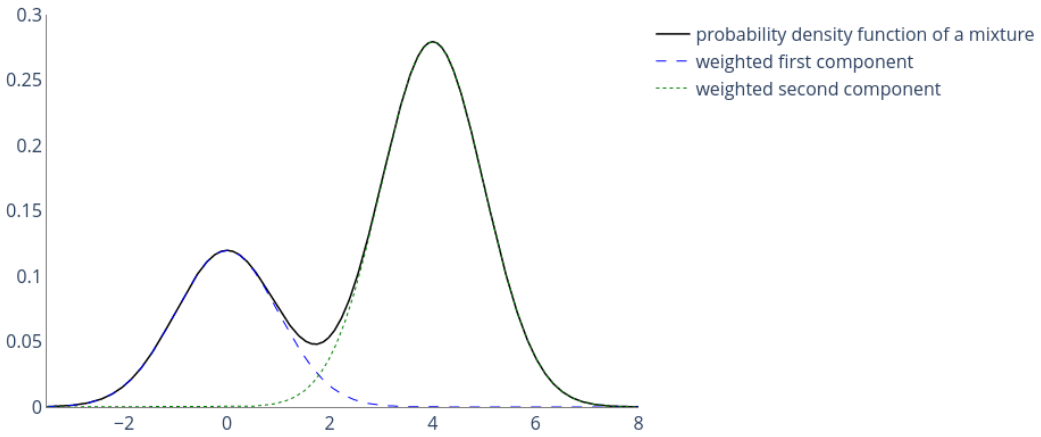

Figure 1.1: Example of a Gaussian mixture with two clusters with means $\mu_1 = 0$ and $\mu_2 = 4$, common variance $\sigma_1^2 = \sigma_2^2 = 1$ and weights $\pi_1 = 0.3$, $\pi_2 = 0.7$.

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is the vector of all unknown parameters including the mixing weights.

Note that we assume K to be fixed in the definition of a finite mixture, however as we will see later, the value K is unknown in many applications and has to be estimated as well.

The most common finite mixture model is a mixture where each component has a multivariate or univariate normal density function. Then we call it a Gaussian mixture model. Its k-th component density function f_k depends on parameter $\boldsymbol{\theta_k} = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k \in \mathbb{R}^d$ is the mean and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ is a positive definite matrix representing the variance matrix, $d \in \mathbb{N}$. We have

$$f_k(\boldsymbol{y}; \boldsymbol{\theta_k}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k)\right\}, \boldsymbol{y} \in \mathbb{R}^d.$$

The density function of the mixture is then

$$f(\boldsymbol{y}; \boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}} \sum_{k=1}^{K} \pi_k \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_k)\right\}, \ \boldsymbol{y} \in \mathbb{R}^d,$$

where $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K})$ is a vector of all model parameters.

An example of the density of univariate Gaussian mixture with two components is shown in Figure 1.1. We can see the weighted density function for each component and the resulting sum of these densities.

1.2 EM algorithm

We will describe the EM algorithm for a general finite mixture model described in Definition 1. Our aim is to estimate the vector of unknown parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter space, a finite-dimensional subset of Euclidean space, based on observed data $\boldsymbol{Y}_n \in \mathbb{R}^d,\ n = 1, \ldots, N$, generated from a mixture $\boldsymbol{Y}$ with K components. Since direct maximization of the observed log-likelihood function with densities in the form of (1.2) would be difficult to handle numerically, we approach this problem in a different way.

Let us define new indicator variables $\boldsymbol{Z} = (Z_1, \ldots, Z_K)^T$ where $Z_k = 1$ if and only if $\boldsymbol{Y}$ is generated from a distribution with the density function f_k and $Z_k = 0$ otherwise. Thus, vector $\boldsymbol{Z}$ has a multinomial distribution with one trial and K possible events. With $\boldsymbol{Z}_n = (Z_{n,1}, \ldots, Z_{n,K}),\ n = 1, \ldots, N$, we obtain a new dataset $\left((\boldsymbol{Y}_n)^T, (\boldsymbol{Z}_n)^T\right)^T \in \mathbb{R}^{d+K}, n = 1, \ldots, N$, forming a random sample from a distribution with joint density function

$$f_{(\boldsymbol{Y},\boldsymbol{Z})}\left(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta}\right) = f_{(\boldsymbol{Y}|\boldsymbol{Z})}\left(\boldsymbol{y}|\boldsymbol{z}; \boldsymbol{\theta}\right) f_{\boldsymbol{Z}}\left(\boldsymbol{z}; \boldsymbol{\theta}\right), \quad \boldsymbol{y} \in \mathbb{R}^d,\ \boldsymbol{z} \in \{0,1\}^K,$$

where $f_{\boldsymbol{Z}}$ denotes the density of Z with respect to the counting measure. Obviously, we do not observe $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$. When dealing with incomplete-data parameters estimation, the EM algorithm performs well.

The complete-data likelihood function is

$$\mathcal{L}_C(\boldsymbol{\theta}) = \prod_n f_{(\boldsymbol{Y},\boldsymbol{Z})}\left(\boldsymbol{Y}_n, \boldsymbol{Z}_n; \boldsymbol{\theta}\right).$$

We take a logarithm of the complete-data likelihood function and obtain the complete-data log-likelihood function

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \ln\left(\prod_{n=1}^{N} f_{(\boldsymbol{Y},\boldsymbol{Z})}\left(\boldsymbol{Y}_n, \boldsymbol{Z}_n; \boldsymbol{\theta}\right)\right) = \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{Y}_n|\boldsymbol{Z}_n; \boldsymbol{\theta})\ f_{\boldsymbol{Z}}(\boldsymbol{Z}_n; \boldsymbol{\theta})\right) \\
&= \ln\left\{\prod_{n=1}^{N}\left[\left(\sum_{k=1}^{K} Z_{n,k} f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k)\right)\left(\prod_{k=1}^{K} (\pi_k)^{Z_{n,k}}\right)\right]\right\} \\
&= \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} Z_{n,k} f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \pi_k \\
&= \sum_{n=1}^{N} \ln\left(\prod_{k=1}^{K} (f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k))^{Z_{n,k}}\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \pi_k \qquad (1.3) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln\left(f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \pi_k. \qquad (1.4)
\end{aligned}$$

Remark. The equality on the line (1.3) is well-defined for $f_k,\ k = 1 \ldots, K$, being strictly greater than zero. In cases where $f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k) = 0$ for some k and $\boldsymbol{Y}_n$ we introduce the convention $0^0 = 1$. The next equality on the line (1.4) is again well-defined for $f_k,\ k = 1 \ldots, K$ strictly greater than zero. In cases where $f_k(\boldsymbol{Y}_n; \boldsymbol{\theta}_k) = 0$ for some k and $\boldsymbol{Y}_n$ we introduce the convention $0 * (-\infty) = 0$.

In the classical approach of maximum likelihood estimation, we would maximize the log-likelihood function. However, here we have variables which are not observed and therefore we first have to deal with them before we proceed to maximization. This leads us to the first step of the EM algorithm, called the expectation step, E step in short. We take the expected value of the complete-data log-likelihood function with respect to the conditional distribution of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$ given $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N$. This distribution depends also on the vector of unknown parameters. Here we assume it is known and we use the estimate from the previous iteration $\hat{\theta}^{old}$ (or the initial estimate $\boldsymbol{\theta}^{init}$ if it is the first iteration).

E-step (Expectation step):

$$
\begin{aligned}
Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{old}) &:= \mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}\right] \qquad (1.5)\\
&= \mathbb{E}_{\hat{\theta}^{old}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln\left(f_k(\boldsymbol{Y}_n;\boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k}\ln\pi_k \;\middle|\; \boldsymbol{Y}\right]\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} \ln\left(f_k(\boldsymbol{Y}_n;\boldsymbol{\theta}_k)\right)\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] + \sum_{n=1}^{N}\sum_{k=1}^{K}\ln\pi_k \mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] \qquad (1.6)
\end{aligned}
$$

We used a fact that the random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are mutually independent. Moreover, it is sufficient to only consider $\boldsymbol{Y}_n$ in the condition instead of whole vector $\boldsymbol{Y}$ because $\boldsymbol{Z}_n$ is independent with $\boldsymbol{Y}_i$ for all $n \neq i$. Let us calculate $\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right]$ by using the definition of conditional probability and Bayes' theorem, see Brémaud [2020], Chapter 1.2.2.

$$
\begin{aligned}
\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] &= \mathsf{P}_{\hat{\theta}^{old}}\left(Z_{n,k}=1|\boldsymbol{Y}_n\right) = f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{e}_k|\boldsymbol{Y}_n;\hat{\theta}^{old}\right)\\
&= \frac{f_{\boldsymbol{Y}|\boldsymbol{Z}}\left(\boldsymbol{Y}_n|\boldsymbol{e}_k;\hat{\theta}^{old}\right) f_{\boldsymbol{Z}}\left(\boldsymbol{e}_k;\hat{\theta}^{old}\right)}{f_{\boldsymbol{Y}}\left(\boldsymbol{Y}_n;\hat{\theta}^{old}\right)} = \frac{f_k\left(\boldsymbol{Y}_n;\hat{\theta}^{old}\right)\hat{\pi}_k^{old}}{\sum_{l=1}^{K}\hat{\pi}_l^{old} f_l\left(\boldsymbol{Y}_n;\hat{\theta}^{old}\right)} =: \langle z_{n,k}\rangle \qquad (1.7)
\end{aligned}
$$

Here $\boldsymbol{e}_k := (e_{k,1}, \ldots, e_{k,K})^T$ denotes a K-dimensional vector with $e_{k,j} = 1$ for $j = k$ and 0 otherwise. Now we obtain the updated value of vector $\hat{\theta}^{new}$ as the maximum over all possible $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ of $\mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}\right]$.

M-step (Maximization step):

$$
\begin{aligned}
\hat{\theta}^{new} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}\right]\\
&= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \ln\left(f_k(\boldsymbol{Y}_n;\boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\ln\pi_k
\end{aligned}
$$

such that $\sum_{k=1}^{K}\pi_k = 1$. This kind of optimization with constraints is usually solved using the Lagrange multipliers method. The Lagrange function is then

$$
\begin{aligned}
L\left(\boldsymbol{\theta},\lambda\right) &= \mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C\left(\boldsymbol{\theta}\right)|\boldsymbol{Y}\right] - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \ln f_k\left(\boldsymbol{Y}_n;\boldsymbol{\theta}_k\right) + \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\ln\pi_k - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right).
\end{aligned}
$$

We can see that the first double summation depends only on the parameters of the component densities $\boldsymbol{\theta}_k$ and the rest of the expression depends only on the weights π_k. Thus, we can maximize the terms separately, which makes the problem significantly easier.
First, we calculate the partial derivative of L with respect to the Lagrange multiplier λ and set it equal to zero. We get

$$\frac{\partial}{\partial \lambda} L\left(\boldsymbol{\theta}, \lambda\right) = 1 - \sum_{k=1}^{K} \hat{\pi}_k^{new} \stackrel{!}{=} 0. \tag{1.8}$$

We then calculate the partial derivatives of the Lagrange function with respect to each model parameter and set it equal to zero (or zero vector/matrix if the dimension is greater than 1). Then we are dealing with solving a set of equations. We calculate the partial derivative of (1.6) with respect to the mixing weight π_k which leads to a closed-form solution for the estimate of π_k. We have

$$\begin{aligned} \frac{\partial}{\partial \pi_k} L\left(\boldsymbol{\theta}, \lambda\right) = \frac{1}{\hat{\pi}_k^{new}} \sum_{n=1}^{N} \langle z_{n,k} \rangle - \lambda &\stackrel{!}{=} 0 \\ \frac{1}{\lambda} \sum_{n=1}^{N} \langle z_{n,k} \rangle &= \hat{\pi}_k^{new}. \end{aligned}$$

In order to calculate λ, we take the sum over all $k = 1, \ldots, K$ and together with (1.8) this leads us to

$$\begin{aligned} \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{n=1}^{N} \langle z_{n,k} \rangle &= \sum_{k=1}^{K} \hat{\pi}_k^{new}, \\ \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{n=1}^{N} \langle z_{n,k} \rangle &= 1, \\ \sum_{k=1}^{K} \sum_{n=1}^{N} \frac{f_k\left(\boldsymbol{Y}_n\right) \hat{\pi}_k^{old}}{\sum_{l=1}^{K} \hat{\pi}_l^{old} f_l\left(\boldsymbol{Y}_n\right)} &= \lambda, \\ \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{f_k\left(\boldsymbol{Y}_n\right) \hat{\pi}_k^{old}}{\sum_{l=1}^{K} \hat{\pi}_l^{old} f_l\left(\boldsymbol{Y}_n\right)} &= \lambda, \\ \sum_{n=1}^{N} 1 &= \lambda, \\ N &= \lambda. \end{aligned}$$

We obtain the expression for the mixing weights π_k

$$\hat{\pi}_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \langle z_{n,k} \rangle, \; k = 1, \ldots, K. \tag{1.9}$$

If we take a partial derivative of L with respect of $\boldsymbol{\theta}_k$ we get

$$\frac{\partial}{\partial \boldsymbol{\theta}_k} L\left(\boldsymbol{\theta}, \lambda\right) = \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\theta}_k} \left[\ln f_k\left(\boldsymbol{Y}_n; \boldsymbol{\theta}_k\right)\right].$$

Thus, if we can differentiate the logarithm of the component density function f_k with respect to its parameters $\boldsymbol{\theta}_k$, we may get a solution in a closed form. We will see such distribution in the next sub-chapter for Gaussian mixtures.

Let us summarize the whole EM algorithm process.

Initialization: Here, we have two options how to initialize the EM algorithm. First, we can choose the initial vector of all unknown parameters $\boldsymbol{\theta}^{init}$ and then proceed to E step. Second, we can just estimate coefficients $\langle z_{n,k} \rangle$ and then go directly to M step. For a discussion of how to select the initial values, see Section 3.2.

E-step: Calculate $\mathbb{E}_{\hat{\theta}^{old}} [\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}]$.

M-step: Find the updated values $\hat{\theta}^{new}$ as $\arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\hat{\theta}^{old}} [\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}]$.

Stopping criterion: Check if the stopping criterion is satisfied. If not, continue with the updated value to the E-step where $\hat{\theta}^{new}$ from the current iteration becomes $\hat{\theta}^{old}$ in the next iteration. If the stopping criterion is not met, the algorithm ends. See Section 3.3 for different stopping criteria.

1.2.1 Basic theory

Let us formulate some theory behind the EM algorithm. We are mainly using the course notes [Omelka, 2021] supported with Wu [1983] and Kushary [1998]. The main goal of the EM algorithm is to substitute direct maximization of the observed log-likelihood function, since it could easily lead to numerical instability, and to obtain a suitable estimate of the maximum likelihood estimator. Let us rewrite the observed log-likelihood function in the terms of the complete-data log-likelihood which is used in the EM algorithm.

$$\begin{aligned}
\ell_{obs}(\boldsymbol{\theta}) &= \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Y}}\left(\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right) = \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Y}}\left(\boldsymbol{Y}_n;\boldsymbol{\theta}\right) \frac{f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)}{f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)}\right) \\
&= \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Y}}\left(\boldsymbol{Y}_n;\boldsymbol{\theta}\right) f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right) - \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right) \\
&= \ln\left(\prod_{n=1}^{N} f_{(\boldsymbol{Y},\boldsymbol{Z})}\left(\boldsymbol{Y}_n,\boldsymbol{Z}_n;\boldsymbol{\theta}\right)\right) - \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right) \\
&= \ell_C(\boldsymbol{\theta}) - \ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right) \qquad (1.10)
\end{aligned}$$

With the use of the expression above, we will prove that the observed log-likelihood in each step does not decrease.

Theorem 1. *Let $\ell_{obs}(\boldsymbol{\theta})$ be the observed log-likelihood and $\hat{\theta}^{(k)}$ be a result of the k-th iteration of the EM algorithm. Then*

$$\ell_{obs}(\hat{\boldsymbol{\theta}}^{(k+1)}) \geq \ell_{obs}(\hat{\boldsymbol{\theta}}^{(k)}).$$

Proof. We apply $\mathbb{E}_{\hat{\theta}^{(k)}}[\,\cdot\,|\boldsymbol{Y}]$ on both sides of (1.10) and from the independence of $\ell_{obs}(\boldsymbol{\theta})$ on $\boldsymbol{Z}$ we get

$$\begin{aligned}\ell_{obs}(\boldsymbol{\theta}) &= \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ell_{obs}(\boldsymbol{\theta})|\boldsymbol{Y}\right] \\ &= \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}\right] - \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)\right)\,\middle|\,\boldsymbol{Y}\right] \\ &=: Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) - H(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}). \end{aligned} \tag{1.11}$$

From the M step, we immediately obtain the following implication.

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\arg\max}\, Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) \Rightarrow Q(\hat{\boldsymbol{\theta}}^{(k+1)},\hat{\boldsymbol{\theta}}^{(k)}) \leq Q(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)})$$

Now, let us restrict $H(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)})$ from above.

$$\begin{aligned} H(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}^{(k)}) &= \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ln\left(\frac{\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)}{\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}^{(k)}\right)}\right)\,\middle|\,\boldsymbol{Y}\right] \\ &\quad + \mathbb{E}_{\hat{\theta}^{(k)}}\left[\ln\left(\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}^{(k)}\right)\right)\,\middle|\,\boldsymbol{Y}\right] \\ &\leq \ln\left(\mathbb{E}_{\hat{\theta}^{(k)}}\left[\frac{\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\boldsymbol{\theta}\right)}{\prod_{n=1}^{N} f_{\boldsymbol{Z}|\boldsymbol{Y}}\left(\boldsymbol{Z}_n|\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}^{(k)}\right)}\,\middle|\,\boldsymbol{Y}\right]\right) + H(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)}) \\ &= \ln(1) + H(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)}) = H(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)}) \end{aligned}$$

In the inequality above we use Jensen's inequality for conditional expectations. Altogether we get the desired inequality.

$$\begin{aligned}\ell_{obs}(\hat{\boldsymbol{\theta}}^{(k+1)}) - \ell_{obs}(\hat{\boldsymbol{\theta}}^{(k)}) &= Q(\hat{\boldsymbol{\theta}}^{(k+1)},\hat{\boldsymbol{\theta}}^{(k)}) - Q(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)}) \\ &\quad - \left[H(\hat{\boldsymbol{\theta}}^{(k+1)},\hat{\boldsymbol{\theta}}^{(k)}) - H(\hat{\boldsymbol{\theta}}^{(k)},\hat{\boldsymbol{\theta}}^{(k)})\right] \geq 0\end{aligned}$$

□

To obtain some convergence characteristic, we need to make the following regularity assumptions.

- the parameter space Θ is a subset of $\mathbb{R}^p$.
- the set $\Theta_0 = \{\boldsymbol{\theta}\in\Theta : \ell_{obs}(\boldsymbol{\theta}) \geq \ell_{obs}(\boldsymbol{\theta}_0)\}$ is compact for any $\boldsymbol{\theta}_0 \in \Theta$ such that $\ell_{obs}(\boldsymbol{\theta}_0) > -\infty$.
- $\ell_{obs}(\boldsymbol{\theta})$ is continuous in Θ and differentiable in the interior of Θ.

With those assumptions, let us state without the proof the convergence theorem of the EM algorithm.

Theorem 2. *Let the function* $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ *defined in* (1.6) *be continuous both in* $\boldsymbol{\theta}$ *and* $\hat{\boldsymbol{\theta}}$. *Then all the limit points of any instance* $\left\{\hat{\boldsymbol{\theta}}^{(k)}\right\}$ *are stationary points of* $\ell_{obs}(\boldsymbol{\theta})$. *Further* $\left\{\ell_{obs}(\hat{\boldsymbol{\theta}}^{(k)})\right\}$ *converges monotonically to some value* $\ell^* = \ell_{obs}(\boldsymbol{\theta}^*)$, *where* $\boldsymbol{\theta}^*$ *is a stationary point of* $\ell_{obs}(\boldsymbol{\theta})$.

Proof. See Wu [1983].

□

Theorem 2 tells us that the EM algorithm gives us $\left\{\hat{\boldsymbol{\theta}}^{(k)}\right\}$ which monotonically converges to some local maximum of $\ell_{obs}(\boldsymbol{\theta})$. Now, we are interested in the speed of such convergence. Let $\hat{\boldsymbol{\theta}}^{(k+1)}$ be the updated parameter from the M-step, i.e.

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(k)})$$

where $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(k)})$ is defined at (1.11). Denote this mapping as $\boldsymbol{M} : \hat{\boldsymbol{\theta}}^{(k)} \longmapsto \hat{\boldsymbol{\theta}}^{(k+1)}$. It can be shown that for $\boldsymbol{M}$ sufficiently smooth, the following holds

$$\hat{\boldsymbol{\theta}}^{(k+1)} - \boldsymbol{\theta}^* = \frac{\partial \boldsymbol{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*) + o(\|\hat{\boldsymbol{\theta}}^{(k)} - \boldsymbol{\theta}^*\|) \tag{1.12}$$

holds and the Jacobi matrix can be expressed as

$$\frac{\partial \boldsymbol{M}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = \left[-\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ell_C(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\,\middle|\,\boldsymbol{Y}\right]\right]^{-1}\left(-\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \ln f(\boldsymbol{Z}|\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\,\middle|\,\boldsymbol{Y}\right]\right)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}.$$

From (1.12) we get the linear rate of convergence.

1.3 EM algorithm for Gaussian mixtures

In Section 1.2 we introduced the EM algorithm for the general finite mixture model and obtained the expression for the update of the mixing weights π_k. Now it is a question whether we can obtain a closed form expression for the updates of the parameters of the component densities $\boldsymbol{\theta}_k$. As we stated before, we have to ensure that the logarithm of the component density f_k is differentiable in $\boldsymbol{\theta}_k$. Moreover, we want this derivative to have a closed-form expression.
Now let us assume that we have a finite mixture where each component has a multivariate Gaussian distribution, a Gaussian mixture. The complete-data

log-likelihood function is

$$\begin{aligned}\ell_C(\boldsymbol{\theta}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln\left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)}\right) \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k}\ln\pi_k \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k}\left[-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_k) - \frac{1}{2}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)\right] \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k}\ln\pi_k.\end{aligned}$$

Let $\hat{\boldsymbol{\theta}}^{old}$ be the estimate of $\boldsymbol{\theta}$ from the previous iteration (or the initial estimate). We can proceed to the expectation step.

E-step (Expectation step):

$$\begin{aligned}\mathbb{E}_{\hat{\theta}^{old}}[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}] &= \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\hat{\theta}^{old}}[Z_{n,k}|\boldsymbol{Y}_n] \\ &\quad\cdot\left[-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_k) - \frac{1}{2}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)\right] \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\ln\pi_k\mathbb{E}_{\hat{\theta}^{old}}[Z_{n,k}|\boldsymbol{Y}_n]\end{aligned}$$

From (1.7) we have

$$\langle z_{n,k}\rangle = \mathbb{E}_{\hat{\theta}^{old}}[Z_{n,k}|\boldsymbol{Y}_n] = \frac{\pi_k^{old} f_k\left(\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}_k^{old}\right)}{\sum_{l=1}^{K}\pi_l^{old} f_l\left(\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}_l^{old}\right)}.$$

Then we find the updated value $\hat{\boldsymbol{\theta}}^{new}$ in the maximization step.

M-step (Maximization step):

$$\begin{aligned}\hat{\theta}^{new} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left[\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\ln\left(f_k(\boldsymbol{Y}_n;\boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\ln\pi_k\right] \\ &= \arg\max_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \\ &\quad\cdot\left[-\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln\det(\boldsymbol{\Sigma}_k) - \frac{1}{2}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)\right] \\ &\quad + \arg\max_{\pi_1,\ldots,\pi_K}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\ln\pi_k\end{aligned}$$

such that $\sum_{k=1}^{K} \pi_k = 1$. This again leads us to the Lagrange multiplier method. The Lagrange function in this case is

$$\begin{aligned} L(\boldsymbol{\theta}, \lambda) &= \mathbb{E}_{\hat{\theta}^{old}} \left[\ell_C(\boldsymbol{\theta}) \, | \boldsymbol{Y} \right] - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{n,k} \rangle \\ &\quad \cdot \left[-\frac{d}{2} \ln(2\pi) - \frac{1}{2} \ln \det(\boldsymbol{\Sigma}_k) - \frac{1}{2} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\ &\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{n,k} \rangle \ln \pi_k - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right). \end{aligned}$$

In the same way as we obtained (1.9), we get

$$\hat{\pi}_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \langle z_{n,k} \rangle, \; k = 1, \ldots, K.$$

Let us calculate the partial derivatives of the Lagrange function with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, set them equal to 0 (or zero vector/matrix) and find the updated values $\hat{\boldsymbol{\theta}}^{new}$, $\hat{\boldsymbol{\Sigma}}^{new}$. We will use the matrix differential calculus, see for example [Magnus, 2019, Part Three].

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}_k} L(\boldsymbol{\theta}, \lambda) &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[(\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \left[-2 \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\ &= \sum_{n=1}^{N} \langle z_{n,k} \rangle \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \stackrel{!}{=} \mathbf{0} \end{aligned}$$

Solving the equality yields

$$\begin{aligned} \sum_{n=1}^{N} \langle z_{n,k} \rangle (\hat{\boldsymbol{\Sigma}}_k^{new})^{-1} \hat{\boldsymbol{\mu}}_k^{new} &= \sum_{n=1}^{N} \langle z_{n,k} \rangle (\hat{\boldsymbol{\Sigma}}_k^{new})^{-1} \boldsymbol{Y}_n \\ \hat{\boldsymbol{\mu}}_k^{new} &= \frac{\sum_{n=1}^{N} \langle z_{n,k} \rangle \boldsymbol{Y}_n}{\sum_{n=1}^{N} \langle z_{n,k} \rangle}. \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} L(\boldsymbol{\theta}, \lambda) &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left[\ln \det \boldsymbol{\Sigma}_k + (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right] \stackrel{!}{=} \mathbf{0} \end{aligned}$$

Again, solving the equality yields

$$\begin{aligned} \sum_{n=1}^{N} \langle z_{n,k} \rangle (\hat{\boldsymbol{\Sigma}}_k^{new})^{-1} &= \sum_{n=1}^{N} \langle z_{n,k} \rangle (\hat{\boldsymbol{\Sigma}}_k^{new})^{-1} (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}) (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new})^T (\hat{\boldsymbol{\Sigma}}_k^{new})^{-1} \\ \hat{\boldsymbol{\Sigma}}_k^{new} &= \frac{\sum_{n=1}^{N} \langle z_{n,k} \rangle (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}) (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new})^T}{\sum_{n=1}^{N} \langle z_{n,k} \rangle}. \end{aligned}$$

We summarize the updated parameters.

$$\hat{\pi}_k^{new} = \frac{1}{N}\sum_{n=1}^{N}\langle z_{n,k}\rangle \tag{1.13}$$

$$\hat{\boldsymbol{\mu}}_k^{new} = \frac{\sum_{n=1}^{N}\langle z_{n,k}\rangle \boldsymbol{Y}_n}{\sum_{n=1}^{N}\langle z_{n,k}\rangle} \tag{1.14}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{new} = \frac{\sum_{n=1}^{N}\langle z_{n,k}\rangle \left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)\left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)^T}{\sum_{n=1}^{N}\langle z_{n,k}\rangle}$$

1.3.1 Homoscedastic components

Until now we considered the case where each Gaussian cluster has its own covariance matrix. However in practice we often set an assumption for them to being the same, i.e.

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma},\ k = 1, \ldots, K,$$

for some positive definite covariance matrix $\boldsymbol{\Sigma}$. The updated parameters $\hat{\pi}_k^{new}$ and $\hat{\boldsymbol{\mu}}_k^{new}$ will remain the same as for general Gaussian mixture, see (1.13) and (1.14). Let us calculate the partial derivative of the Lagrange function with respect to $\boldsymbol{\Sigma}$ and set it to zero matrix in order to obtain the updated parameter $\hat{\boldsymbol{\Sigma}}^{new}$.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} L\left(\boldsymbol{\theta}, \lambda\right) &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \frac{\partial}{\partial \boldsymbol{\Sigma}}\left[\ln\det\boldsymbol{\Sigma} + \left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}^{-1}\right] \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\langle z_{n,k}\rangle\boldsymbol{\Sigma}^{-1} - \langle z_{n,k}\rangle\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}^{-1}\right] \\
&= -\frac{N}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}^{-1} \stackrel{!}{=} \mathbf{0}
\end{aligned}$$

We obtained

$$N(\hat{\boldsymbol{\Sigma}}^{new})^{-1} = \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle(\hat{\boldsymbol{\Sigma}}^{new})^{-1}\left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)\left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)^T(\hat{\boldsymbol{\Sigma}}^{new})^{-1}$$

which leads to

$$\hat{\boldsymbol{\Sigma}}^{new} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle\left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)\left(\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}\right)^T. \tag{1.15}$$

1.3.2 Isotropic components

Let us consider another special case where each Gaussian cluster has Gaussian distribution with covariance matrix equal to

$$\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}_d,\ k = 1, \ldots, K,$$

for some $\sigma_k^2 > 0,\ k = 1, \ldots, K$. The updated parameters $\hat{\pi}_k^{new}$ and $\hat{\boldsymbol{\mu}}_k^{new}$ will remain the same as for general Gaussian mixture, see (1.13) and (1.14). Let us calculate the partial derivative of the Lagrange function with respect to $\boldsymbol{\sigma}_k$ and set it to a zero matrix in order to obtain the updated parameter $\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}^{new}$.

$$\begin{aligned}\frac{\partial}{\partial \sigma_k^2} L(\boldsymbol{\theta}, \lambda) &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \sigma_k^2} \left[\ln \det \sigma_k^2 \boldsymbol{I}_d - \frac{1}{\sigma_k^2} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{I}_d (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\ &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \left[\frac{d}{\sigma_k^2} - \frac{1}{\sigma_k^4} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \stackrel{!}{=} \boldsymbol{0}\end{aligned}$$

From there we have

$$\hat{\sigma}_k^{new} = \frac{1}{d \sum_{n=1}^{N} \langle z_{n,k} \rangle} \sum_{n=1}^{N} \langle z_{n,k} \rangle (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new})^T (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}). \tag{1.16}$$

1.3.3 Isotropic homoscedastic components

Let us consider another special case where each Gaussian cluster has Gaussian distribution with covariance matrix equal to

$$\boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I}_d,\ k = 1, \ldots, K,$$

for some $\sigma^2 > 0$. The updated parameters $\hat{\pi}_k^{new}$ and $\hat{\boldsymbol{\mu}}_k^{new}$ will remain the same as for general Gaussian mixture, see (1.13) and (1.14). From (1.15) and (1.16) we get

$$\hat{\sigma}^{new} = \frac{1}{Nd} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{n,k} \rangle (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new})^T (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}).$$

2. EM algorithm for truncated Gaussian mixture model

In this chapter we would like to apply the introduced algorithm to Gaussian mixture when dealing with truncated data. The truncation often occurs in spatial processes when the observation window is restricted to some size and we are not able to observe data points outside this window.

2.1 Truncated Gaussian distribution

Assume we have a sample from a Gaussian distribution in $\mathbb{R}^d$ with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ but our observation window is restricted to a bounded rectangle $[\boldsymbol{s}, \boldsymbol{t}]$, $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$. In order to take such truncation into account we often use a truncated Gaussian distribution instead. The probability density function of the truncated Gaussian distribution is

$$g(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}}, \quad \boldsymbol{y} \in \mathbb{R}^d$$

where $f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a probability density function of a normal random variable with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Let us now calculate the first two moments for the truncated Gaussian variable $\boldsymbol{Y}$. The gradient of a function f is denoted as ∇f.

$$\begin{aligned}
\mathbb{E}\boldsymbol{Y} &= \int_{\boldsymbol{s}}^{\boldsymbol{t}} \boldsymbol{y}\, g(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} = \int_{\boldsymbol{s}}^{\boldsymbol{t}} \boldsymbol{y} \frac{f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}}\, d\boldsymbol{y} \\
&= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} (\boldsymbol{y} - \boldsymbol{\mu} + \boldsymbol{\mu}) f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} \\
&= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \left[\boldsymbol{\Sigma} \int_{\boldsymbol{s}}^{\boldsymbol{t}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} + \boldsymbol{\mu} \int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x} \right] \\
&= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \boldsymbol{\Sigma} \int_{\boldsymbol{s}}^{\boldsymbol{t}} \nabla f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} + \boldsymbol{\mu} =: \mathcal{M}^1(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}])
\end{aligned} \tag{2.1}$$

The integral part of the formula above is

$$\int_{\boldsymbol{s}}^{\boldsymbol{t}} \nabla f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} = \begin{bmatrix} \int_{\boldsymbol{s}}^{\boldsymbol{t}} \frac{\partial}{\partial y_1} f(\boldsymbol{y})\, d\boldsymbol{y} \\ \int_{\boldsymbol{s}}^{\boldsymbol{t}} \frac{\partial}{\partial y_2} f(\boldsymbol{y})\, d\boldsymbol{y} \\ \vdots \\ \int_{\boldsymbol{s}}^{\boldsymbol{t}} \frac{\partial}{\partial y_d} f(\boldsymbol{y})\, d\boldsymbol{y} \end{bmatrix}. \tag{2.2}$$

Now, take the ith element of (2.2).

$$\int_{\boldsymbol{s}}^{\boldsymbol{t}} \frac{\partial}{\partial y_i} f(\boldsymbol{y})\, d\boldsymbol{y} = \int_{s_1}^{t_1} \cdots \int_{s_d}^{t_d} \frac{\partial}{\partial y_i} f(\boldsymbol{y})\, dy_1 \cdots dy_d = \int_{\boldsymbol{s}_{-i}}^{\boldsymbol{t}_{-i}} f(\boldsymbol{y})\, d\boldsymbol{y}_{-i} =: F_i(y_i)$$

The notation $\boldsymbol{x}_{-k}$ denotes the $(d-1)$th dimensional vector with elements x_i for $i \in \{1, \cdots d\} \setminus \{k\}$. Let $\phi_i(\boldsymbol{x}'; \boldsymbol{\mu}', \boldsymbol{\Sigma}')$, $\Phi_i(\boldsymbol{x}'; \boldsymbol{\mu}', \boldsymbol{\Sigma}')$ denote the probability density function and the cumulative distribution function, respectively, for i-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}'$ and covariance matrix $\boldsymbol{\Sigma}'$. Since f is Gaussian density, the conditional densities of $\boldsymbol{Y}_{-i}|Y_i$, where Y_i is the ith element of vector $\boldsymbol{Y}$, are also Gaussian with mean $\boldsymbol{\mu}_{-i|i}(y_i) = \boldsymbol{\Sigma}_{-i|i}\boldsymbol{\Sigma}_{i,i}^{-1} y_i$ and covariance matrix $\boldsymbol{\Sigma}_{-i|i} = \boldsymbol{\Sigma}_{-i,-i} - \boldsymbol{\Sigma}_{-i,i}\boldsymbol{\Sigma}_{i,i}^{-1}\boldsymbol{\Sigma}_{i,-i}$.

We can now express $F_i(y_i)$,

$$\begin{aligned} F_i(y_i) &= \int_{\boldsymbol{s}_{-i}}^{\boldsymbol{t}_{-i}} f(\boldsymbol{y})\, d\boldsymbol{y}_{-i} = \phi_1(x; \mu_i, \sigma_i^2) \int_{\boldsymbol{s}_{-i}}^{\boldsymbol{t}_{-i}} \phi_{d-1}(\boldsymbol{x}_{-i}; \boldsymbol{\mu}_{-i|i}(y_i), \boldsymbol{\Sigma}_{-i|i}) d\boldsymbol{y}_{-i} \\ &= \phi_1(x; \mu_i, \sigma_i^2) \left(\Phi_{d-1}(\boldsymbol{t}_{-i}; \boldsymbol{\mu}_{-i|i}(y_i), \boldsymbol{\Sigma}_{-i|i}) - \Phi_{d-1}(\boldsymbol{s}_{-i}; \boldsymbol{\mu}_{-i|i}(y_i), \boldsymbol{\Sigma}_{-i|i}) \right) \end{aligned} \tag{2.3}$$

By combination of (2.1) and (2.3) we get

$$\mathcal{M}^1(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}]) = \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \boldsymbol{\Sigma} \cdot \begin{pmatrix} F_1(Y_1) \\ F_2(Y_2) \\ \vdots \\ F_d(Y_d) \end{pmatrix} + \boldsymbol{\mu}.$$

Similarly we calculate the second moment of the truncated Gaussian variable $\boldsymbol{Y}$.

$$\begin{aligned} \mathbb{E}\boldsymbol{Y}\boldsymbol{Y}^T &= \int_{\boldsymbol{s}}^{\boldsymbol{t}} \boldsymbol{y}\boldsymbol{y}^T\, g(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} = \int_{\boldsymbol{s}}^{\boldsymbol{t}} \boldsymbol{y}\boldsymbol{y}^T \frac{f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}}\, d\boldsymbol{y} \\ &= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} \\ &\quad + \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} (-\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{y}^T + \boldsymbol{y}\boldsymbol{y}^T) f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y} \\ &= \underbrace{\frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} (\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})^T f(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\, d\boldsymbol{y}}_{\text{denoted by } N} \\ &\quad - \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\mathcal{M}^1(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}])^T + \mathcal{M}^1(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}])\boldsymbol{\mu}^T \\ &=: \mathcal{M}^2(\boldsymbol{\mu}, \boldsymbol{\Sigma}; [\boldsymbol{s}, \boldsymbol{t}]) \end{aligned}$$

Let us calculate N.

$$\begin{aligned} N &= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} (\boldsymbol{y}-\boldsymbol{\mu})(\boldsymbol{y}-\boldsymbol{\mu})^T f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\,d\boldsymbol{y} \\ &= \frac{1}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})\,d\boldsymbol{x}} \int_{\boldsymbol{s}}^{\boldsymbol{t}} J\left(\nabla f(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma})\right)\,d\boldsymbol{y} \end{aligned}$$

Here, J denotes the Jacobian matrix. So in order to express N, we need to calculate partial derivatives of the terms F_i defined in (2.3). It can be shown that

$$\begin{aligned} \frac{\partial F_i(y_i)}{\partial y_j} &= \frac{\partial}{\partial x_j} \int_{s_1}^{t_1} \cdots \int_{s_{i-1}}^{t_{i-1}} \int_{s_{i+1}}^{t_{i+1}} \cdots \int_{s_d}^{t_d} f(\boldsymbol{y})\,d\boldsymbol{y}_{-i} \\ &= \frac{\sigma_{j,i} y_i F_i(y_i)}{\sigma_{i,i}} + \sum_{q\neq i} \left(\sigma_{j,q} - \frac{\sigma_{i,q}\sigma_{j,i}}{\sigma_{i,i}}\right) \left(F_{i,q}(y_i, s_q) - F_{i,q}(x_i, t_q)\right) \end{aligned}$$

where

$$\begin{aligned} F_{i,q}(y_i, y_q) &= \int_{\boldsymbol{s}_{-\{i,q\}}}^{\boldsymbol{t}_{-\{i,q\}}} f(\boldsymbol{y})\,d\boldsymbol{y}_{-\{i,q\}} \\ &= \phi_2(y_i, y_q; \mu\{i,q\}, \Sigma_{\{i,q\},\{i,q\}}) \\ &\quad \cdot \int_{\boldsymbol{s}_{-\{i,q\}}}^{\boldsymbol{t}_{-\{i,q\}}} \phi_{d-2}(\boldsymbol{x}_{-\{i,q\}}; \boldsymbol{\mu}_{-\{i,q\}|\{i,q\}}(y_i, y_q), \boldsymbol{\Sigma}_{-\{i,q\}|\{i,q\}})\,d\boldsymbol{y}_{-\{i,q\}} \\ &= \phi_2(y_i, y_q; \mu\{i,q\}, \Sigma_{\{i,q\},\{i,q\}}) \\ &\quad \cdot \Big(\Phi_{d-2}(\boldsymbol{t}_{-\{i,q\}}; \boldsymbol{\mu}_{-\{i,q\}|\{i,q\}}(y_i, y_q), \boldsymbol{\Sigma}_{-\{i,q\}|\{i,q\}}) \\ &\qquad - \Phi_{d-2}(\boldsymbol{s}_{-\{i,q\}}; \boldsymbol{\mu}_{-\{i,q\}|\{i,q\}}(y_i, y_q), \boldsymbol{\Sigma}_{-\{i,q\}|\{i,q\}})\Big) \end{aligned} \tag{2.4}$$

holds, see Lee [1979]. Altogether, after further steps described in mentioned source, we would obtain the following expression

$$\begin{aligned} \mathbb{E}(Y_i, Y_j) &= \sum_{k=1}^{d} \sigma_{i,k} \frac{\sigma_{j,k}\left(s_k F_k(s_k) - t_k F_k(t_k)\right)}{\sigma_{k,k}} \\ &\quad + \sum_{k=1}^{d} \sigma_{i,k} \sum_{q\neq k} \left(\sigma_{j,q} - \frac{\sigma_{k,q}\sigma_{j,k}}{\sigma_{k,k}}\right) \Big[\left(F_{k,q}(s_k, s_q) - F_{k,q}(s_k, t_q)\right) \\ &\quad - \left(F_{k,q}(t_k, s_q) - F_{k,q}(t_k, t_q)\right) \Big] \end{aligned}$$

2.2 Truncated Gaussian mixture distribution

We would like to apply the EM algorithm for a Gaussian mixture data observed in a bounded rectangle window $[\boldsymbol{s}, \boldsymbol{t}]$, $\boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d$. The density of such process is then

$$g(\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{y};\boldsymbol{\theta})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x};\boldsymbol{\theta})\,d\boldsymbol{x}}, \quad \boldsymbol{y} \in \mathbb{R}^d \tag{2.5}$$

where

$$f(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, f_k(\boldsymbol{y};\boldsymbol{\theta}_k), \quad \boldsymbol{y} \in \mathbb{R}^d \tag{2.6}$$

is the probability density of a Gaussian mixture without truncation, see Definition 1. We can rewrite such density as

$$g(\boldsymbol{y};\boldsymbol{\theta}) = \sum_{k=1}^{K} \eta_k \, g_k(\boldsymbol{y};\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k \frac{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f_k(\boldsymbol{x};\boldsymbol{\theta}_k)\, d\boldsymbol{x}}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x}} \, \frac{f_k(\boldsymbol{y};\boldsymbol{\theta}_k)}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f_k(\boldsymbol{x};\boldsymbol{\theta}_k)\, d\boldsymbol{x}}$$

with mixing weights

$$\eta_k = \pi_k \frac{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f_k(\boldsymbol{x};\boldsymbol{\theta}_k)\, d\boldsymbol{x}}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f(\boldsymbol{x};\boldsymbol{\theta})\, d\boldsymbol{x}}, \quad k = 1, \ldots, K \tag{2.7}$$

and component densities

$$g_k(\boldsymbol{y};\boldsymbol{\theta}_k) = \frac{f_k(\boldsymbol{y};\boldsymbol{\theta}_k)}{\int_{\boldsymbol{s}}^{\boldsymbol{t}} f_k(\boldsymbol{x};\boldsymbol{\theta}_k)\, d\boldsymbol{x}}, \quad k = 1, \ldots, K. \tag{2.8}$$

We can see that component densities are actually truncated Gaussian densities. Hence, when dealing with Gaussian mixture data observed on a bounded rectangle window, we can approach it as **a truncated Gaussian mixture** with weights defined by (2.7).

2.3 EM algorithm for truncated Gaussian mixture

In the previous section we introduced the truncated Gaussian mixture. Now, we can move on to using the EM algorithm itself on such mixture. As in Section 1.3, we are following steps described in Section 1.2. The established notation also remains the same.

The complete-data log-likelihood function for the truncated Gaussian mixture is

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln\left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)}\right) \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \eta_k \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}\right) d\boldsymbol{x} \\
&= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k}\left[d\ln(2\pi) + \ln\det(\boldsymbol{\Sigma}_k) + (\boldsymbol{Y}_n-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{Y}_n-\boldsymbol{\mu}_k)\right] \\
&+ \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \eta_k \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K} Z_{n,k} \ln \int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}\right) d\boldsymbol{x}.
\end{aligned}$$

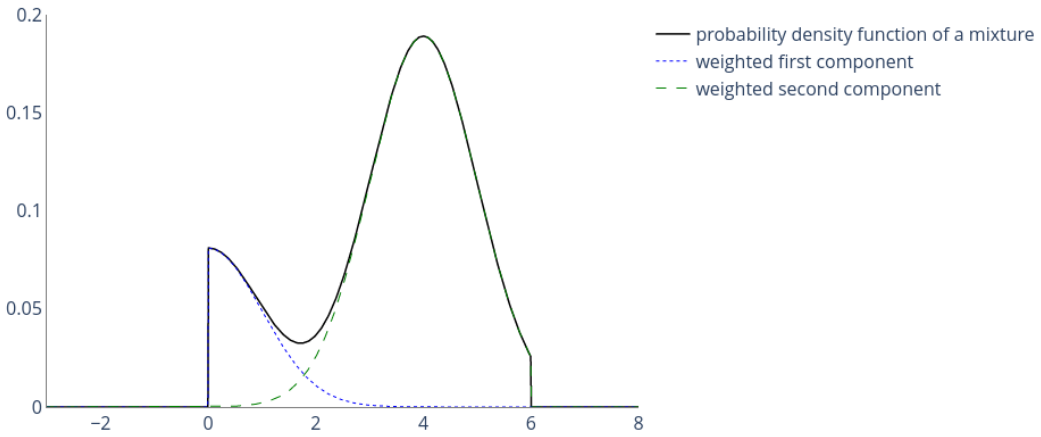

Figure 2.1: Example of a truncated Gaussian mixture with two clusters with means $\mu_1 = 0$ and $\mu_2 = 4$, common variance $\sigma_1^2 = \sigma_2^2 = 1$ and weights $\pi_1 = 0.3$, $\pi_2 = 0.7$, $[s, t] = [0, 6]$. Note that the aforementioned weights π_1, π_2 are weights of a Gaussian mixture without truncation, see (2.6). In order to calculate the mixing weights, we use formula (2.7).

Let $\hat{\boldsymbol{\theta}}^{old}$ be the estimate of $\boldsymbol{\theta}$ from the previous iteration (or the initial estimate). We can proceed to the expectation step.

E-step (Expectation step):

$$\begin{aligned}
\mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C(\boldsymbol{\theta})|\boldsymbol{Y}\right] = &-\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] \\
&\cdot\left[d\ln(2\pi) + \ln\det\left(\boldsymbol{\Sigma}_k\right) + \left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)\right] \\
&+\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right]\ln\eta_k \\
&-\sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] \\
&\cdot\left[\ln\int_{\boldsymbol{s}}^{\boldsymbol{t}}\left((2\pi)^{-\frac{d}{2}}\det\left(\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)}\right)d\boldsymbol{x}\right]
\end{aligned}$$

From (1.7) we have

$$\langle z_{n,k}\rangle = \mathbb{E}_{\hat{\theta}^{old}}\left[Z_{n,k}|\boldsymbol{Y}_n\right] = \frac{\eta_k^{old}g_k\left(\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}^{old}\right)}{\sum_{l=1}^{K}\eta_l^{old}g_l\left(\boldsymbol{Y}_n;\boldsymbol{\theta}_l\right)} = \frac{\eta_k^{old}\dfrac{f_k(\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}_k^{old})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}}f_k(\boldsymbol{x};\hat{\boldsymbol{\theta}}_k^{old})\,d\boldsymbol{x}}}{\sum_{l=1}^{K}\eta_l^{old}\dfrac{f_l(\boldsymbol{Y}_n;\hat{\boldsymbol{\theta}}_l^{old})}{\int_{\boldsymbol{s}}^{\boldsymbol{t}}f_l(\boldsymbol{x};\hat{\boldsymbol{\theta}}_l^{old})\,d\boldsymbol{x}}}$$

Then we find the updated value $\hat{\boldsymbol{\theta}}^{new}$ in the maximization step.

M-step (Maximization step):

$$\hat{\theta}^{new} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{n,k}\rangle \ln\left(g_k(\boldsymbol{Y}_n;\boldsymbol{\theta}_k)\right) + \sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{n,k}\rangle \ln \eta_k \right] \tag{2.9}$$

$$= \arg\max_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_K} \Bigg\{ -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{n,k}\rangle \tag{2.10}$$

$$\cdot \left[d\ln(2\pi) + \ln\det\left(\boldsymbol{\Sigma}_k\right) + \left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right) \right] \tag{2.11}$$

$$+ \sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{n,k}\rangle \ln \eta_k \tag{2.12}$$

$$- \sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{n,k}\rangle \tag{2.13}$$

$$\cdot \left[\ln \int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det\left(\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)} \right) d\boldsymbol{x} \right] \Bigg\} \tag{2.14}$$

such that $\sum_{k=1}^{K} \eta_k = 1$. This again leads us to the Lagrange multiplier method. The Lagrange function in this case is

$$\begin{aligned} L(\boldsymbol{\theta},\lambda) &= \mathbb{E}_{\hat{\theta}^{old}}\left[\ell_C(\boldsymbol{\theta})\,|\boldsymbol{Y}\right] - \lambda\left(\sum_{k=1}^{K}\eta_k - 1\right) \\ &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \\ &\quad \cdot \left[d\ln(2\pi) + \ln\det\left(\boldsymbol{\Sigma}_k\right) + \left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1}\left(\boldsymbol{Y}_n - \boldsymbol{\mu}_k\right) \right] \\ &\quad + \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \ln\eta_k \\ &\quad - \sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{n,k}\rangle \\ &\quad \cdot \left[\ln\int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det\left(\boldsymbol{\Sigma}_k\right)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)} \right) d\boldsymbol{x} \right] \\ &\quad - \lambda\left(\sum_{k=1}^{K}\eta_k - 1\right). \end{aligned} \tag{2.15}$$

$$\tag{2.16}$$

In the same way as we obtained (1.9), we get

$$\hat{\eta_k}^{new} = \frac{1}{N}\sum_{n=1}^{N}\langle z_{n,k}\rangle,\ k = 1,\ldots,K.$$

Let us calculate the partial derivatives of the Lagrange function with respect to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, set them equal to 0 (or zero vector/matrix) and find the updated

values $\hat{\boldsymbol{\theta}}_k^{new}$, $\hat{\boldsymbol{\Sigma}}_k^{new}$.

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} L(\boldsymbol{\theta}, \lambda) &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[(\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\
&\quad - \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[\ln \int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \right) d\boldsymbol{x} \right] \\
&= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \left[-2\boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \right] \\
&\quad - \sum_{n=1}^{N} \langle z_{n,k} \rangle \boldsymbol{\Sigma}_k^{-1} \left[\mathcal{M}^1(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k; [\boldsymbol{s}, \boldsymbol{t}]) - \boldsymbol{\mu}_k \right] \\
&= \boldsymbol{\Sigma}_k^{-1} \sum_{n=1}^{N} \langle z_{n,k} \rangle \boldsymbol{Y}_n - \boldsymbol{\Sigma}_k^{-1} \mathcal{M}^1(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k; [\boldsymbol{s}, \boldsymbol{t}]) \sum_{n=1}^{N} \langle z_{n,k} \rangle \stackrel{!}{=} \boldsymbol{0}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k} L(\boldsymbol{\theta}, \lambda) &= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \Big[\ln \det \boldsymbol{\Sigma}_k + (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) \\
&\quad - \ln \int_{\boldsymbol{s}}^{\boldsymbol{t}} \left((2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma}_k)^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}-\boldsymbol{\mu}_k)} \right) d\boldsymbol{x} \Big] \\
&= -\frac{1}{2} \sum_{n=1}^{N} \langle z_{n,k} \rangle \Big[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{Y}_n - \boldsymbol{\mu}_k) (\boldsymbol{Y}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
&\quad + \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} \mathcal{M}^2(\boldsymbol{0}, \boldsymbol{\Sigma}_k; [\boldsymbol{s} - \boldsymbol{\mu}_k, \boldsymbol{t} - \boldsymbol{\mu}_k]) \boldsymbol{\Sigma}_k^{-1} \Big] \stackrel{!}{=} \boldsymbol{0}
\end{aligned}$$

As the unknown parameters are part of the integral, it is not possible to obtain closed formulas for the updated values of $\hat{\boldsymbol{\mu}}_k^{new}$ and $\hat{\boldsymbol{\Sigma}}_k^{new}$. To obtain those estimates, the numerical optimization will be used. In order to guarantee that the covariance matrix will always be positive semi-definite, we perform Cholesky matrix decomposition

$$\boldsymbol{\Sigma}_k = \boldsymbol{Q}_k^T \boldsymbol{Q}_k, \; k = 1, \ldots, K. \tag{2.17}$$

This decomposition always exists and is unique as $\boldsymbol{\Sigma}_k$ is symmetric positive definite matrix, see for example Griffel [1989]. Then our aim is to find the unknown matrix $\boldsymbol{Q}_k$ and from the expression (2.17), $\boldsymbol{\Sigma}_k$ is computed.

2.3.1 Method used in the article Lee and Scott [2012]

In this diploma thesis we are dealing with truncated Gaussian mixtures and the application of the EM algorithm to these mixtures. The article EM algorithm for multivariate Gaussian mixtures models with truncated and censored data, Lee and Scott [2012], derived formulas for the updated parameters in M step as follows

$$\begin{aligned}
\hat{\pi}_k^{new} &= \frac{1}{N} \sum_{n=1}^{N} \langle z_{n,k} \rangle \\
\hat{\boldsymbol{\mu}}_k^{new} &= \frac{\sum_{n=1}^{N} \langle z_{n,k} \rangle \boldsymbol{Y}_n}{\sum_{n=1}^{N} \langle z_{n,k} \rangle} - \boldsymbol{m}_k \\
\hat{\boldsymbol{\Sigma}}_k^{new} &= \frac{\sum_{n=1}^{N} \langle z_{n,k} \rangle (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new}) (\boldsymbol{Y}_n - \hat{\boldsymbol{\mu}}_k^{new})^T}{\sum_{n=1}^{N} \langle z_{n,k} \rangle} + \boldsymbol{H}_k
\end{aligned}$$

where $\boldsymbol{m}_k$ and $\boldsymbol{H}_k$ are

$$\begin{aligned}\boldsymbol{m}_k &= \mathcal{M}^1(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k; [\boldsymbol{s} - \hat{\boldsymbol{\mu}}_k, \boldsymbol{t} - \hat{\boldsymbol{\mu}}_k]) \\ \boldsymbol{H}_k &= \hat{\boldsymbol{\Sigma}}_k - \mathcal{M}^2(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k; [\boldsymbol{s} - \hat{\boldsymbol{\mu}}_k, \boldsymbol{t} - \hat{\boldsymbol{\mu}}_k])\end{aligned} \tag{2.18}$$

and $\hat{\boldsymbol{\mu}}_k$, $\hat{\boldsymbol{\Sigma}}_k$ are the updated parameters from the previous iteration step and $\mathcal{M}^1(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k; [\boldsymbol{s} - \hat{\boldsymbol{\mu}}_k, \boldsymbol{t} - \hat{\boldsymbol{\mu}}_k])$, $\mathcal{M}^2(\mathbf{0}, \hat{\boldsymbol{\Sigma}}_k; [\boldsymbol{s} - \hat{\boldsymbol{\mu}}_k, \boldsymbol{t} - \hat{\boldsymbol{\mu}}_k])$ are the first, respective the second, moment of a centred Gaussian distribution with covariance $\hat{\boldsymbol{\Sigma}}_k$ while truncation to a bounded rectangle window $[\boldsymbol{s} - \hat{\boldsymbol{\mu}}_k, \boldsymbol{t} - \hat{\boldsymbol{\mu}}_k]$.
The simplification from the more correct approach described in this diploma thesis lies in the expressions $\boldsymbol{m}_k$ and $\boldsymbol{H}_k$ as they use the updated parameters from the previous step. However, there should be the unknown parameters $\hat{\boldsymbol{\mu}}_k^{new}$, $\hat{\boldsymbol{\Sigma}}_k^{new}$ respectively. It is understandable simplification as this way we can get rid of the unknown parameters in integrals and the whole expressions for the unknown parameters is reduced to a simple formula. If such simplification is made, we do not have guaranteed that the basic theory behind the EM algorithm is applicable. No justification was made in this article. We could struggle with the convergence speed and even convergence itself. Moreover, we can obtain singular covariance matrix during the M step. In one-dimensional case it would occur when

$$\frac{\sum_{n=1}^{N} \langle z_{n,k} \rangle \left(Y_n - \mu_k^{new}\right)^2}{\sum_{n=1}^{N} \langle z_{n,k} \rangle} \leq \mathcal{M}^2(0, \hat{\sigma}_k; [s - \hat{\mu}_k, t - \hat{\mu}_k]) - \hat{\sigma}_k^2.$$

2.4 Effect of truncation on EM algorithm

Applying the EM algorithm for truncated Gaussian mixtures causes non-existence of the explicit formulas for the updated parameters in M-step. Hence, we have to use the numerical optimization for obtaining those parameters. the question is if it is necessary to take into account the truncation. We will compare the standard algorithm with Gaussian distributions and the version where truncation is considered.

In Figure 2.2 of one-dimensional Gaussian mixture with two components such that $\mu_1 = -3$ and $\mu_2 = 20$ and the variance is for both $\sigma^2 = 10$. The truncation interval is $[0, 40]$ so the first mean μ_1 lies outside the observation window. We can see that both algorithms, the one described in this Chapter and its standard version described in Chapter 1 where truncation is not considered, produce a good estimate of μ_2. On the other hand, the estimate of mean μ_1 which lies outside the observation window with standard version of the EM algorithm is not good. The truncated version of EM algorithm performs much better, it correctly detects that the mean of first cluster lies outside the window. Moreover, the estimated variance using standard EM algorithm is much lower for the cluster centred outside the window, approximately 2.6, as it is estimated only based on observed data. the same variance estimated using truncated version of EM algorithm is close to true value, approximately 11.6. The conclusion of this example would be that it is not recommended to use the standard version of the EM algorithm when we have a priori information about some not negligible truncation.

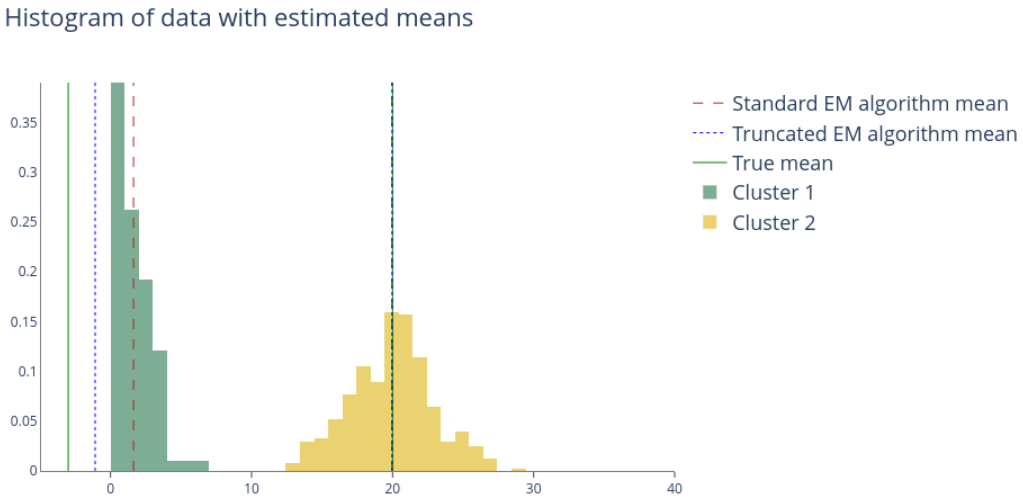

Figure 2.2: The experiment with one-dimensional synthetic data. Data comes from Gaussian mixture with two components with means -3 and 20 and common variance 10. In total, 1000 data points were simulated. Then the truncation at interval $[0, 40]$ was performed. The histogram (top) of truncated data is shown together with real mean and estimated means with the standard EM algorithm and using the algorithm described in this Chapter and its standard version described in Chapter 1.

2.5 Generalisation

Until now we consider Gaussian mixtures truncated on a bounded rectangle window $[\boldsymbol{s}, \boldsymbol{t}]$, $s, t \in \mathbb{R}^d$, however we could easily extend the presented theory to cases where the truncation window is general arbitrary window W. Then the integral in denominator of expression (2.8) is replaced by $\int_W f_k(\boldsymbol{x}; \boldsymbol{\theta}_k)\, d\boldsymbol{x}$ which is an integral of Gaussian density over a window W. As long as we can calculate this integral, we are able to proceed with the EM algorithm as described for a bounded rectangular window.

3. Miscellaneous topics

In the previous chapters, we introduced the EM algorithm, however in order to perform this algorithm in practice, we need to deal with several issues. In the first section, we address the problem of the unknown number of Gaussian components. Next, we look into the initialization of the EM algorithm. In the last section, we mention several possibilities on how to define the stopping criterion for the EM algorithm.

Topics in this chapter are often discussed when dealing with the EM algorithm, however not specifically for the truncated Gaussian mixtures. Our aim is to adopt known methods and try to use them for our issue. Then evaluate them and decide which approach to each subtopic is the most convenient.

3.1 Unknown number of components

In the previous chapter, we described the EM algorithm for truncated Gaussian mixture. We assumed that we know the number of clusters (K in Definition 1). However, as stated before, this number is often unknown and before proceeding to the algorithm itself, we need to estimate it. Denote $\mathcal{M}_K$ the class of all possible K-component truncated Gaussian mixtures. We are not able to estimate the unknown value of K via the EM algorithm itself as the classes $\mathcal{M}_{K+1}$ and $\mathcal{M}_K$ are nested, that is $\mathcal{M}_K \subseteq \mathcal{M}_{K+1}$. Then the function $h(K) = \sup_{\boldsymbol{\Theta}\in\Theta} \ell_{obs}(\boldsymbol{\theta}; K)$ is a non-decreasing function of K.

3.1.1 Likelihood ratio test statistics

One way how to estimate the number of components is to start with some reasonably small number of components, run the EM algorithm, then add one component and again run the EM algorithm. With those two runs we perform the likelihood ratio test, see more at McLachlan and Rathnayake [2014]. We test the null hypothesis $H_0 : k = k_0$ against the alternative hypothesis $H_A : k = k_0+1$.

The likelihood ratio λ is defined as

$$\lambda = \frac{\mathcal{L}(\hat{\boldsymbol{\theta}}(k_0 + 1))}{\mathcal{L}(\hat{\boldsymbol{\theta}}(k_0))}$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}}(k_i))$ is the observed likelihood function assuming k_i Gaussian components evaluated at the EM algorithm estimate of the mixture parameters $\hat{\boldsymbol{\theta}}(k_i)$. The likelihood ratio test statistic is $-2\ln\lambda$ which can be rewritten as

$$LRTS = -2\ln\lambda = 2\left[\ell_{obs}(\hat{\boldsymbol{\theta}}(k_0)) - \ell_{obs}(\hat{\boldsymbol{\theta}}(k_0 + 1))\right].$$

The higher $LRTS$ the higher the evidence against the null hypothesis. We keep adding one component until the increase in the log-likelihood starts to fall or the evidence against the null hypothesis is not evident.

3.1.2 Information criteria

Often, a deterministic method for determining the number of components is used. We select set of possible number of components $\mathcal{K}$. The estimated number of components is then obtained as

$$\widehat{K} = \arg\min_{K} \left\{ \mathcal{C}(\hat{\boldsymbol{\theta}}(K); K); K \in \mathcal{K} \right\}$$

where $\mathcal{C}(\hat{\boldsymbol{\theta}}(K); K)$ is some criterion as a function of the number of components K and the EM algorithm estimate of the mixture parameters $\hat{\boldsymbol{\theta}}$ while assuming K components. The criterion $\mathcal{C}$ should penalize higher values of k in order to avoid the over-fitting while taking into account the value of the observed log-likelihood $\ell_{obs}(\hat{\boldsymbol{\theta}}(K))$.

First criterion to mention would be Akaike's information criterion (AIC for short), see the original paper Akaike [1974]. It is defined as

$$AIC(\hat{\boldsymbol{\theta}}(K); K) = -2\ell_{obs}(\hat{\boldsymbol{\theta}}(K)) + 2p$$

where p stands for the number of unknown parameters. It can be shown that by minimizing AIC, we minimize the estimate of the Kullback-Leibler divergence, see for example Cavanaugh and Neath [2019].

Second criterion is Bayesian information criterion (BIC for short) defined as

$$BIC(\hat{\boldsymbol{\theta}}(K); K) = -2\ell_{obs}(\hat{\boldsymbol{\theta}}(K)) + p \ln N$$

where p is again the number of unknown parameters and N is the sample size. As the name suggests, it is a criterion based on Bayesian statistics. More on this criteria can be found in the original paper Schwarz [1978] where this criteria has been introduced.

In both cases, we will select model with lower value. As stated in Panić et al. [2020] or Biernacki et al. [2003] we can observe that AIC favours complex models with a higher number of components due to the small penalization term. BIC on the other hand penalizes a higher number of components more heavily. So when the number of observations is large, the penalization term in AIC is negligible and the criterion favours model with high number of components which can lead to over-fitting. So we would like to avoid AIC criterion in such situations. On the other hand, the penalization term in BIC increases with increasing number of observations so it can be used with large number of observations without fear of over-fitting.

For small sample size, Hurvich and Tsai proposed the corrected AIC (AICc) which is defined as

$$AICc(\hat{\boldsymbol{\theta}}(K); K) = -2\ell_{obs}(\hat{\boldsymbol{\theta}}(K)) + 2p + \frac{2p(p+1)}{N-p-1}$$

where p and N are defined as above.

3.1.3 Adaptive EM algorithm

Another way how to select the number of Gaussian components is to implement the selection process into EM algorithm itself via the adaptive EM algorithm. Instead of using the classical log-likelihood function and respective Lagrange function, we subtract from the defined Lagrange function (2.16) during M-step the penalization

$$\frac{d}{2}\ln N + \frac{T}{2}\sum_{k=1}^{K}\ln \eta_k$$

where d is the total number of unknown parameters and $T = \frac{D(D+3)}{2}$ with D being the dimension. By taking the partial derivative of such function with respect to the mixing weights, we obtain

$$\hat{\eta}_k^{new} = \frac{\sum_{n=1}^{N}\langle z_{n,k}\rangle - \frac{T}{2}}{N - \frac{TK}{2}},\ k = 1, \ldots, K.$$

In the case when $\sum_{n=1}^{N}\langle z_{n,k}\rangle \leq \frac{T}{2}$, it indicates that the k-th component should be killed. The whole algorithm is as follows.

Initialization: Choose the initial number of components K together with the vector of all unknown parameters $\boldsymbol{\theta}^{init}$ and then proceed to E step.

E-step: Calculate $\mathsf{E}_{\hat{\theta}^{old}}\left[\ell(\boldsymbol{\theta})|\boldsymbol{Y}\right]$.

M-step: Calculate the updated weights as

$$\hat{\eta}_k^{new*} = max\left(\frac{\sum_{n=1}^{N}\langle z_{n,k}\rangle - \frac{T}{2}}{N - \frac{TK}{2}}, 0\right),\ k = 1, \ldots, K.$$

If $\hat{\eta}_k^{new*} = 0$ then the k-th component should be dropped, K is decreased by one and $\hat{\boldsymbol{\mu}}_k = 0$ and $\hat{\boldsymbol{\Sigma}}_k = \mathbf{0}$, otherwise the updated values are found as in the standard EM algorithm procedure. Then the component's weight should be renormalized as

$$\hat{\eta}_k^{new} = \frac{\hat{\eta}_k^{new*}}{\sum_{k=1}^{K}\hat{\eta}_k^{new*}}.$$

Stopping criterion: Check if the stopping criterion is satisfied. If not, continue with the updated value to the E-step where $\hat{\theta}^{new}$ from the current iteration becomes $\hat{\theta}^{old}$ in the next iteration. If yes, the algorithm ends. Because of the penalization term, the moments of the observed data are not conserved so after the stopping criterion is satisfied, we perform one standard EM iteration.

3.2 Initialization

Now, let us assume we know the number of Gaussian components (or at least we have a reasonable estimate, see Section 3.1), denote it K. The EM algorithm is very sensitive to the choice of the initial distribution, especially when dealing with truncation as we need to use the optimization to find the local maximum likelihood because we have no analytic solution to the system of equations. The convergence of this non-linear optimization problem is not guaranteed for arbitrary initial parameters so it is important to find a reasonable initial value for the unknown vector of parameters. Moreover, only the convergence of the log-likelihood function to a local maxima is guaranteed so the initial values should be close enough to the desired solution in order to avoid the convergence to the wrong local maxima. In some practical situations it could be reasonable to run the EM algorithm with set of different initial values and then select the solution which makes most sense in given case.

3.2.1 Random initialization

The simplest way to initialize the EM algorithm is to randomly choose K component means located reasonably in the observed window or not far away from it. Then each point from our random sample assign to the closest component mean. Then for each component mean calculate the sample variance from assigned data points. The weights are calculated as the ratio of the number of assigned data points and the total number of data points. The simplicity of this approach has its fly in the ointment. There is a risk that the initial choice of parameters will lead to a local maximum which is not the global maximum.

3.2.2 Previous short runs of EM

More sophisticated method would be to select some set of possible initial parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. Then perform predefined small number n_1 of iterations and compare the values of the observed log-likelihood function evaluated at the n_1th updated parameter estimates. Select such initial parameter $\boldsymbol{\theta}_{selected} \in \boldsymbol{\Theta}_0$ which leads to the highest value of the log-likelihood function.

3.2.3 K-Means

K-means clustering is well known unsupervised machine learning algorithm. We briefly introduce this algorithm based on Sun et al. [1994]. The beauty of this algorithm lies in its simplicity, yet it is an extremely powerful algorithm in many cases. The algorithm runs as follows.

1. Choose the number of clusters K.
2. Select K random data points as centroids.
3. Assign all other points to the closest centroid which will form clusters.
4. Select the new centroids as the middle point of clusters.
5. Repeat steps 3. and 4. until predefined stopping criteria is satisfied.

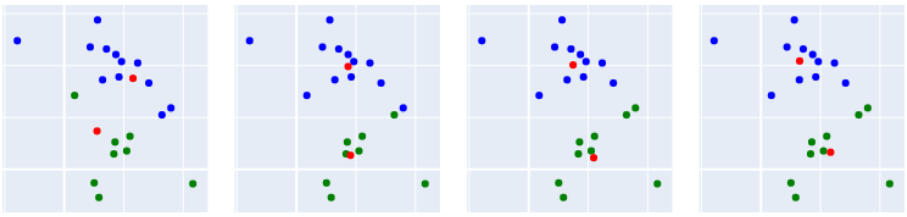

Figure 3.1: Example of K-means clustering algorithm. In each step, the red points representing cluster means were selected. Then the rest of points were coloured (blue or green) based on the closest cluster point. In total, 4 iterations are shown.

The weights are calculated as for the random initialization, it is as the ratio of the number of assigned data points and the total number of data points. An example of K-Mean algorithm is in Figure 3.1.

3.3 Stopping criteria

We discussed how to initialize the EM algorithm and now it is time to talk about how to stop such algorithm. This section is in addition to those listed next based on article Abbi et al. [2008] In theory, the log-likelihood function is guaranteed not to decrease each iteration until complete convergence. In order to reduce the computational time, we often stop the algorithm before its complete convergence using some heuristic approach. Usually, the relative change of the log-likelihood function or estimated parameters is used together with setting a fixed lower bound and when the relative change is smaller than this bound, the EM algorithm stops. When the EM algorithm is stopped using this criteria, it tells us that the progress is negligible.

3.3.1 Number of iterations

One approach is to set a fixed number of iterations and stop the EM algorithm after this number of steps is reached. It is necessary to select a reasonable large number of steps in order to guarantee a solution close to the local maxima. The disadvantage of such approach is that it is highly dependable on the given dataset. We can have a dataset where complete convergence occurs after 10 steps and another dataset where this convergence occurs after 10^3 steps. It would be time consuming to select one large number of steps and apply it to all cases.

3.3.2 Difference in log-likelihood function

Another quite straightforward approach is to watch the change of the observed log-likelihood function after each step.

We can consider the absolute change

$$\left(\ell_{obs}\left(\boldsymbol{\theta}^{new}\right) - \ell_{obs}\left(\boldsymbol{\theta}^{old}\right)\right) < \delta$$

or the relative change

$$\frac{\ell_{obs}\left(\boldsymbol{\theta}^{new}\right) - \ell_{obs}\left(\boldsymbol{\theta}^{old}\right)}{|\ell_{obs}\left(\boldsymbol{\theta}^{old}\right)|} < \delta$$

for some predefined small δ. If the above condition is met, the algorithm is stopped.

3.3.3 Difference in estimated parameters

Next we can observe the change in the estimated parameters. Here we have several options however it seems to be the best approach to consider the change in the covariance matrix (or the variance in the one-dimensional case).

3.3.4 Aitken acceleration-based stopping criterion

Both the change in the log-likelihood function or in the estimated parameters are measures of lack of progress rather than of actual convergence, see Lindstrom and Bates [1988]. Here we introduce a criterion based on the estimate of the log-likelihood limit point. the following is mainly based on Geoffrey J. McLachlan [2000]. From Theorem 2 we know that there exists ℓ^* such that the sequence of the observed log-likelihood $\left\{\ell(\hat{\boldsymbol{\theta}}^{(k)})\right\}$ converges monotonically to this value. For the simplicity of notation we denote

$$\ell(\hat{\boldsymbol{\theta}}^{(k)}) = \ell^{(k)}.$$

Moreover, the rate of convergence is linear, see Section 1.2.1, so we can write

$$\ell^{(k+1)} - \ell^* \approx c\,(\ell^{(k)} - \ell^*)$$

for all k and some c, $0 < c < 1$. We can rewrite such relationship as

$$\ell^{(k+1)} - \ell^{(k)} \approx (1-c)(\ell^* - \ell^{(k)}).$$

From this we can see that the small increment in the log-likelihood does not have to mean the same distance of the log-likelihood in the kth step and the limit l^*. From here, we express the limit

$$\ell^* \approx \ell^{(k)} + \frac{1}{1-c}(\ell^{(k+1)} - \ell^{(k)}).$$

Since the parameter c is unknown, we use its estimate

$$c^{(k)} = \frac{\ell^{(k+1)} - \ell^{(k)}}{\ell^{(k)} - \ell^{(k-1)}}.$$

Altogether, Aitken accelerated estimate of ℓ^* is

$$\ell_A^{(k+1)} = \ell^{(k)} + \frac{1}{1 - c^{(k)}}(\ell^{(k+1)} - \ell^{(k)}).$$

The algorithm is stopped when

$$\left|\ell_A^{(k+1)} - \ell_A^{(k)}\right| < \delta$$

for some predefined small δ.

3.4 Implementation

The whole EM algorithm for truncated Gaussian mixtures was implemented in `Python`, as of now compatible with the version `3.10.4`. It can be found on author's personal Github repository, `https://github.com/NguyenAdela/EM-algorithm-for-truncated-Gaussian-mixtures`.

We are also using `R`, version `4.1.3`, and its library `truncnorm`. The most important packages used in `Python` are `scikit-learn` version `1.0.2`, `scipy` version `1.8.0`, `numpy` version `1.22.3` and for the visualisation mainly `matplotlib` version `3.5.1` and `plotly` version `5.6.0`. We use the Python interface to `R` lanugage with the help of library `PypeR` version `1.1.2`.

In order to solve the find the updated parameters in M step, we use the Nelder–Mead method implemented in `scipy`.

For more details on the implementation, please visit the provided link to the Github repository.

4. Application

4.1 Synthetic data

We are going to apply the EM algorithm to synthetic data generated from a known distribution in order to verify whether the proposed algorithm in this diploma thesis (hereinafter *thesis algorithm*) will perform better than in the mentioned article Lee and Scott [2012] (hereinafter *article algorithm*) where the used algorithm is simplified by heuristic arguments but not mathematically justified.

To generate data, we use the algorithm implemented in `R` package `truncnorm`. We run the calculation on a standard personal computer with processor Intel® Core™ i5. For both approaches, we use the same initialization, K-means initialization with the same seed used by the random number generator (which leads to the identical initial parameters), and the same stopping criteria based on the difference in the observed log-likelihood function. Our selected threshold for stopping criteria is $\delta = 10^{-8}$.

In subsection 4.1.2 we define the Kullback-Leibler divergence score which will be one of the criteria helping us evaluate the models' performance. We will also be interested in the number of iterations until the stopping criterion is met together with the time of the whole calculation.

4.1.1 Generating data from GMM

One way how to generate a sample from a truncated Gaussian mixture is to generate data from a Gaussian mixture and then perform the truncation. This is often the case when dealing with a real dataset when it follows a Gaussian mixture distribution but the observation window is restricted. The only problem with such an approach is that we are not able to generate a given number of points in a predefined number of steps.

The second approach would be the mixed rejection algorithm for univariate sampling and the Gibbs algorithm for multivariate sampling. Those algorithms are well described in the article Geweke [1998] and implemented in `R` package `truncnorm` which we will use in our experiments.

Another possibility is to generate a sample from the multinomial distribution with n trials and with probabilities given by the cluster weights which provides us with a number of data points for each cluster and then generate a respective number of points for each cluster.

4.1.2 Evaluation

In order to evaluate how the estimate produced by the EM algorithm performs we will calculate Kullback–Leibler divergence (KL) score which in layman's terms

quantifies the difference between two probability distributions. The theory behind KL score can be found in Thomas M. Cover and Thomas [2006], among others. The formal definition from Taboga [2021] follows.

Definition 2. *Let X and Y be two continuous random variables with supports R_X and R_Y and probability density function f_X and f_Y such that*

$$\int_A f_X(\boldsymbol{x})\, d\boldsymbol{x} \neq 0 \Rightarrow \int_A f_Y(\boldsymbol{x})\, d\boldsymbol{x} \neq 0$$

for any measurable set $A \subseteq R_X$. Then the Kullback-Leibler (KL) divergence of f_Y from f_X is defined as

$$D_{KL}(f_X||f_Y) = -\int_{\boldsymbol{x}\in R_X} f_X(\boldsymbol{x}) \ln \frac{f_Y(\boldsymbol{x})}{f_X(\boldsymbol{x})}\, d\boldsymbol{x}.$$

In practice f_X would be the known true distribution and f_Y its estimate. In order to calculate KL score, we generate N_C data points drawn from the known true distribution f_X and use the following approximation

$$D_{KL}(p||q) = \mathbb{E}_p\left[\ln f_X - \ln f_Y\right] \approx \frac{1}{N_C}\sum_{n=1}^{N_C}\left[\ln f_X(\boldsymbol{x^n}) - \ln f_Y(\boldsymbol{x^n})\right] \tag{4.1}$$

where $\boldsymbol{x^n}$ are the generated data points. As KL score quantifies the difference between the true distribution and its estimate, the lower the value of KL score, the better.

4.1.3 One-dimensional case

Let us consider one-dimensional data. We fix the observation window for all examples in this subsection to the interval $[s, t] = [0, 40]$. For most of the experiments, we calculate the KL score where we fix $N_C = 2000$ in equation (4.1). In addition to the overall estimate of the distribution, we will be interested in the estimates of parameters themselves, especially in the cluster means.

We start with the simplest examples with one cluster in the middle of the observation window (Experiment 1), then shift the cluster mean close to the border of the window (Experiment 2) and finally outside the observation window (Experiment 3).

Third experiment leads us to the next two experiments. In Experiment 4 we explore the behaviour of estimates depending on the number of simulated data points. Single-cluster experiments in one dimension end with a simulation study, Experiment 5, where we generate 500 datasets drawn from the same distribution and look into distributions of estimates.

In the last one-dimensional experiment, Experiment 6, we explore the behaviour for two clusters based on the size of their overlap.

$N = 150$, $K = 1$, $\mu_1 = 20$, $\sigma_1^2 = 25$, $s = 0$, $t = 40$

Thesis algorithm $\hat{\mu}_1 = 19.8391$, $\hat{\sigma}_1^2 = 21.3388$
Article algorithm $\hat{\mu}_1 = 19.8391$, $\hat{\sigma}_1^2 = 21.3388$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	1.8652 s	-442.3556	0.0077
Article algorithm	3	1.5296 s	-442.3556	0.0077

Table 4.1: Summary of Experiment 1.

Experiment 1.

One cluster with mean in the middle of the observation window

We start with the simplest example, one cluster centered in the middle of the observation window, with variance low enough for the truncation to be negligible. We expect both algorithms perform similarly well since the integral in (2.5) will be close to 1, and so we can use the standard EM algorithm for Gaussian mixtures because $\boldsymbol{m_k}$ and $\boldsymbol{H_k}$ defined in (2.18) will be negligible.

Figure 4.1 shows the results of such experiment. We have one cluster with mean $\mu = 20$. The variance is $\sigma^2 = 25$. In total, 150 data points were simulated. At the top of the figure, we have a histogram of the simulated dataset together with the true and the estimated means. At the bottom, the observed log-likelihood function calculated after each iteration for both approaches is shown. We can see that both approaches perform really well in terms of finding an accurate mean estimate. The change in the log-likelihood function is minimal and the algorithm is stopped after 2 iterations in case of the thesis algorithm, respectively 3 iterations for the article algorithm.

In Table 4.1, we summarized the experiment's results. We compare the time of the calculation, as well as the number of iterations. Even with one iteration more, the article algorithm is 0.3 seconds faster than the thesis algorithm. Nevertheless, in terms of complete-data log-likelihood or KL score, we see no noticeable difference.

Such results are not surprising as we mentioned at the beginning of the experiment. Moreover, even K-means algorithm itself performs quite well, its results (i.e. the initial parameters for the EM algorithm) are

$$\hat{\mu}_{1,init} = 19.8392, \quad \hat{\sigma}^2_{1,init} = 21.4756.$$

It is thus safe to use the article algorithm or even K-means algorithm, which in the case of a single cluster reduces to just sample mean and sample variance estimation, when the truncation is clearly negligible.

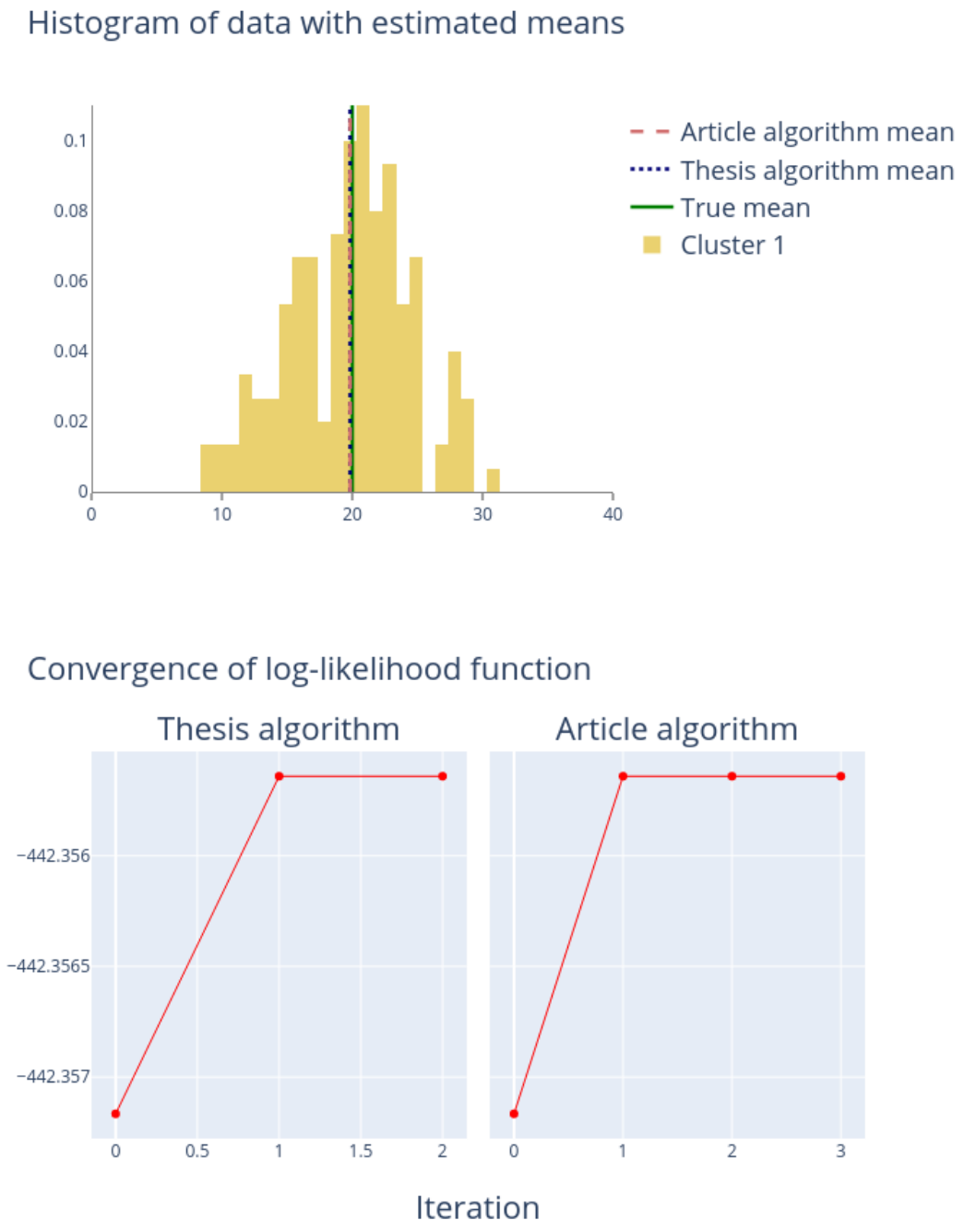

Figure 4.1: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with one component with mean $\mu_1 = 20$ and variance $\sigma_1^2 = 25$. In total, $N = 150$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

$N = 150$, $K = 1$, $\mu_1 = 3$, $\sigma_1^2 = 25$, $s = 0$, $t = 40$

Thesis algorithm $\hat{\mu}_1 = 4.3738$, $\hat{\sigma}_1^2 = 11.9524$
Article algorithm $\hat{\mu}_1 = 4.3738$, $\hat{\sigma}_1^2 = 11.9522$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	0.5758 s	-363.6647	0.0493
Article algorithm	25	3.5965 s	-363.6647	0.0493

Table 4.2: Summary of Experiment 2.

Experiment 2.

One cluster with mean close to the border of the observation window

A more interesting example is in Figure 4.2 where we have the same case as in Experiment 1 with the only difference in the position of the cluster mean, here set as $\mu = 3$. In such case, the truncation plays a more significant role, we are facing a problem of missing information. The summary of this experiment is shown in Table 4.2.

For the thesis algorithm, the observed log-likelihood again converges to the local maximum in only 2 iterations. The article algorithm requires 25 iterations to converge. In terms of time, the thesis algorithm is more than three times faster.

Nevertheless, both approaches arrived at the same results. Both underestimate the amount of truncation, they estimate the mean closer to the center than the true mean and less than half of the true variance.

The initial parameters, obtained by K-means, clearly cannot capture the truncation, we have

$$\hat{\mu}_{1,init} = 5.0644, \quad \hat{\sigma}^2_{1,init} = 8.5114.$$

The initial and the final probability density functions (considering only the thesis algorithm, however for the article algorithm it would be very similar) compared to the true density is captured in Figure 4.3. Loosely speaking, the final estimate is somewhere between the initial and the true distribution. However, we can see that the left tail of the final estimate is quite underestimated.

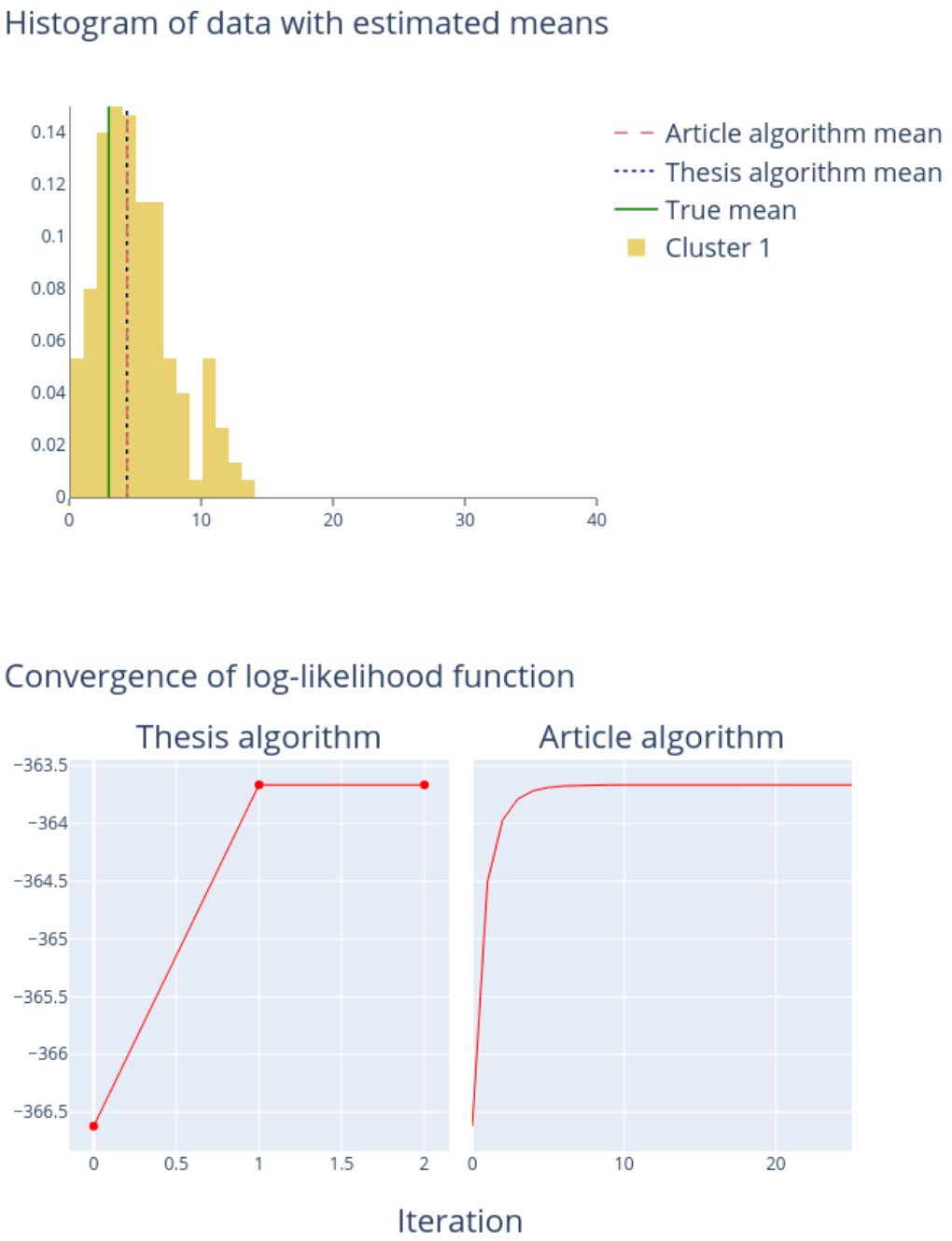

Figure 4.2: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with one component with mean $\mu_1 = 3$ and variance $\sigma_1^2 = 25$. In total, $N = 150$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

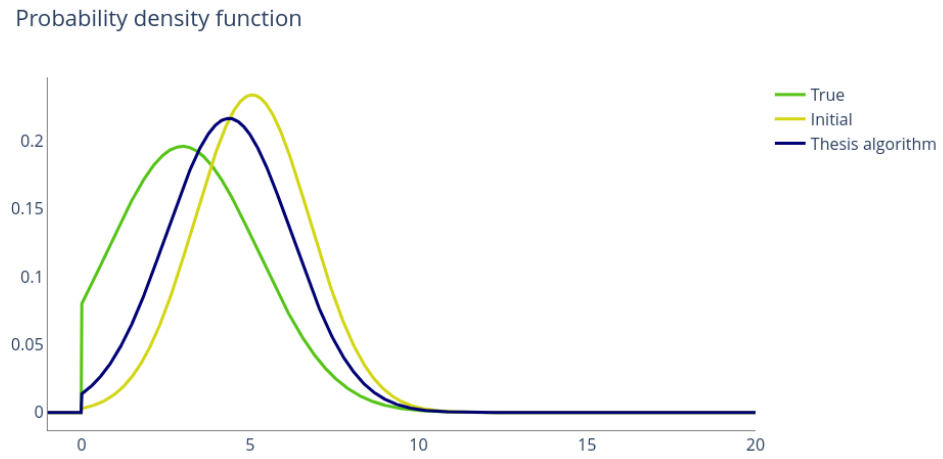

Figure 4.3: The initial and the final estimated probability density function (considering the thesis algorithm) for Experiment 2 compared to the true density which is a density of the truncated normal distribution with mean 3 and variance 25, truncated at the interval [0.40].

The article algorithm takes 23 iterations more untill the stopping criterion is met compared to the thesis algorithm. We are interested in the estimate of the mean using the article algorithm in case it is stopped after 2 iterations as in the thesis algorithm. Figure 4.4 shows the estimates of the mean for each iteration with the value at 2 iterations marked by a horizontal line. We can see that even after only 2 iterations, the article algorithm estimates the mean quite well, however we still need the value to move by 0.3 in order to get the final estimate.

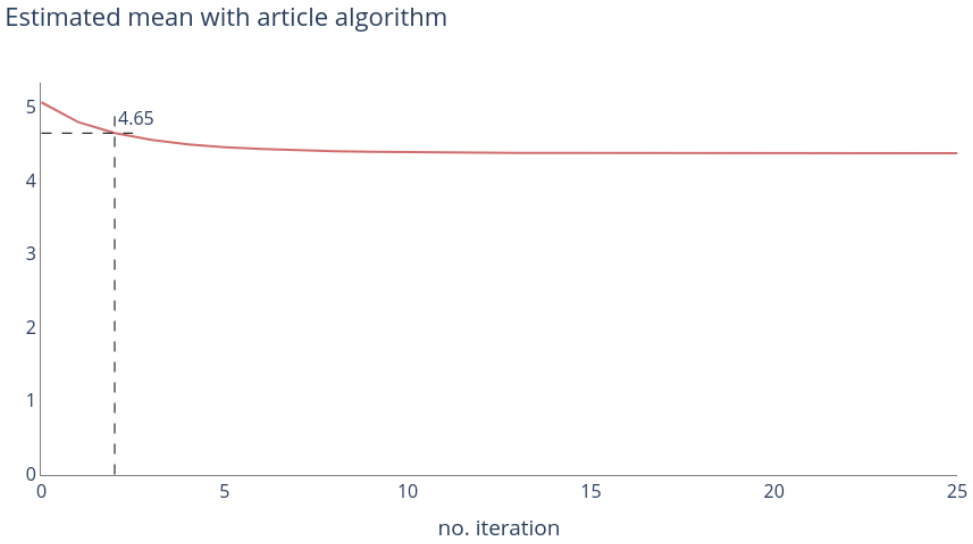

Figure 4.4: The estimate of the mean for each iteration using the article algorithm. The value at 2 iterations is highlighted for comparison with the thesis algorithm, which takes 2 iterations to converge.

$N = 150$, $K = 1$, $\mu_1 = -8$, $\sigma_1^2 = 25$, $s = 0$, $t = 40$

Thesis algorithm $\hat{\mu}_1 = -14.6803$, $\hat{\sigma}_1^2 = 38.7838$
Article algorithm $\hat{\mu}_1 = -14.6568$, $\hat{\sigma}_1^2 = 38.7368$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	0.4452 s	-260.1273	0.0012
Article algorithm	1150	101.3453 s	-260.1273	0.0012

Table 4.3: Summary of Experiment 3.

Experiment 3.

One cluster with mean outside the observation window

The last set of parameters with one cluster in one dimension is in Figure 4.5 where again we have the same situation as in the previous experiments, except the mean which is located outside the observation window, specifically we have $\mu = -8$.

The results are in Table 4.3. Both approaches estimated that the mean is outside the observation window (which was not the case for the standard algorithm, see Figure 2.2). But what is more remarkable is that we again only need 2 iterations until the thesis algorithm is stopped. Article approach is stopped after more than 1000 iterations and it is more than a hundred times slower.

In Figure 4.6, we have plots of the squared Euclidean distances between the true mean and its estimates calculated at each iteration of the EM algorithm. For the thesis algorithm, it behaves expected, that is the distance is decreasing between the initial and the first iteration and then remains the same (and the algorithm is stopped). This is not the case for the article algorithm. The distance is decreasing up to iteration 86 where it almost reaches zero and then starts to increase and reaches the same distance as the thesis algorithm. It could seem that the article algorithm is able to achieve better results at some point, it estimates the true mean almost perfectly. However, for a real problem, we do not know that such situation occurs so we are unable to decide to stop the algorithm at that point. Furthermore, we consider only one realisation of the respective random sample, for different realisations this does not have to be the case.
As we see in the bottom right plot in Figure 4.5, the complete data log-likelihood does not change a lot from iteration 86 to the final iteration, more accurately the complete data log-likelihood after iteration 86 is 260.3031 which is about 0.18 less than the final log-likelihood. In this specific case, we could choose higher threshold for stopping criteria, for example $\delta = 0.01$, and then the algorithm would stop with the estimate of the mean closer to the true mean, however we need some general rule which is applicable for real problems which would be really hard to obtain. Moreover, it is not guaranteed that the algorithm would not stop even earlier with the estimate even further from the true mean. Therefore, this does not seem reasonable to do.

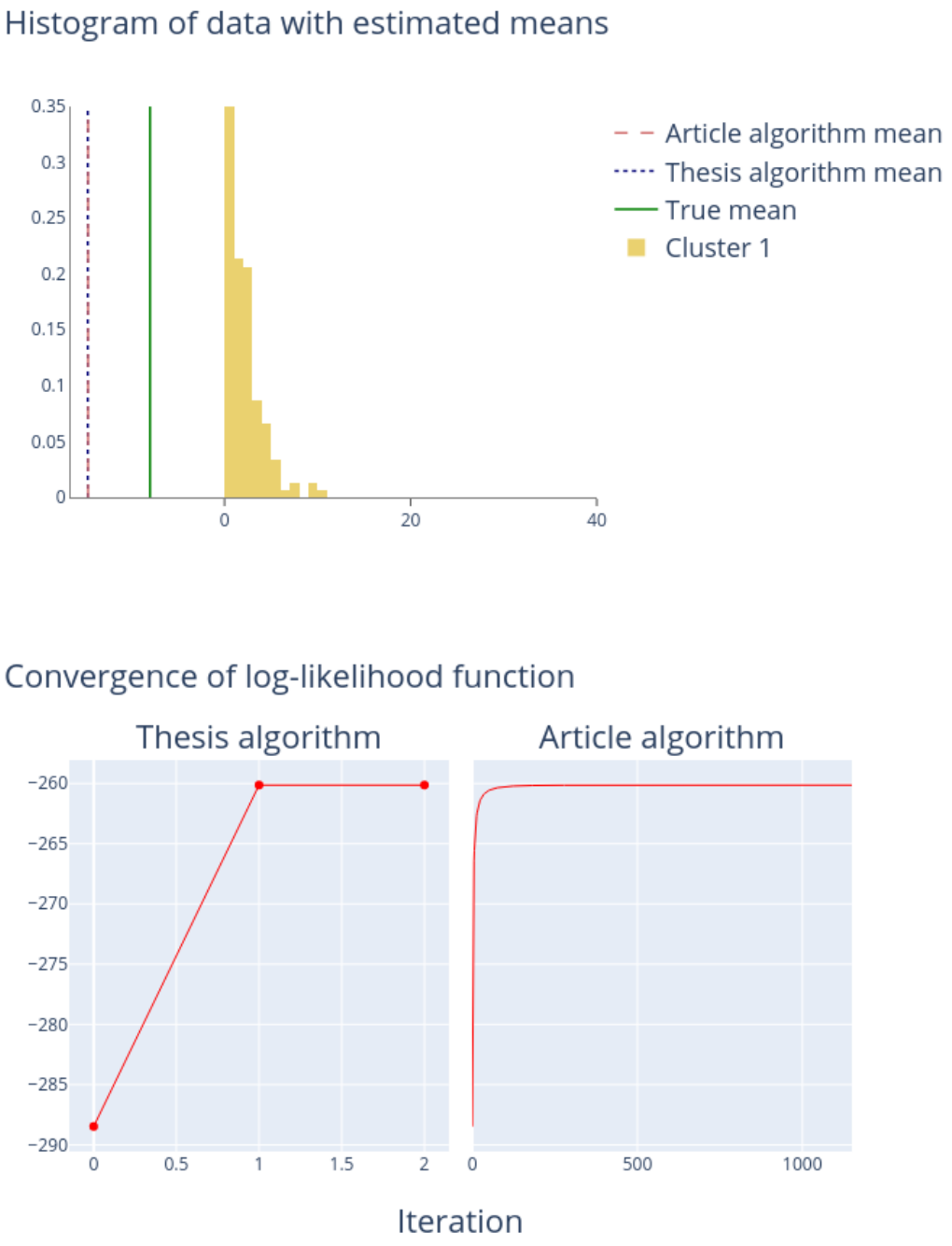

Figure 4.5: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with one component with mean $\mu_1 = -8$ and variance $\sigma_1^2 = 25$. In total, $N = 150$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

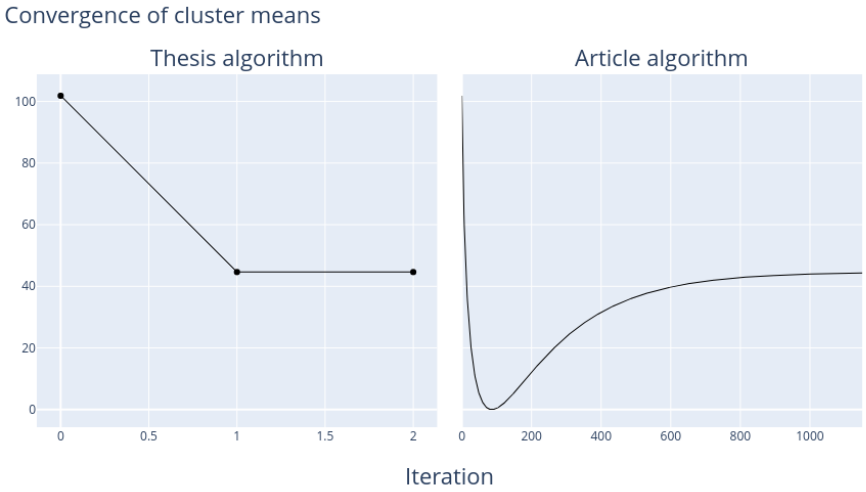

Figure 4.6: Convergence of the estimates of the mean for Experiment 3.

As in the previous experiments, the initial parameters, obtained by K-means, cannot capture the truncation and the fact that the cluster mean is outside the observation window. K-mean algorithm in general tries to estimate given distribution with the density concentrated in the observation window, we have

$$\hat{\mu}_{1,init} = 2.0918, \quad \hat{\sigma}^2_{1,init} = 3.7241.$$

We again plot the initial and the final probability density functions (considering only the thesis algorithm) compared to the true density, see Figure 4.7. In addition, the estimate using the standard EM algorithm is shown. Unlike the previous experiment, when we had the center of the cluster inside the observation window, now the density and especially its shape has changed radically from the initial estimate to the final estimate. Let us look at the estimate using the standard EM algorithm. It is obvious that in cases when the cluster mean is outside the observation window, it would be really inappropriate to use the standard EM algorithm, as the estimated density differs from the true density significantly.

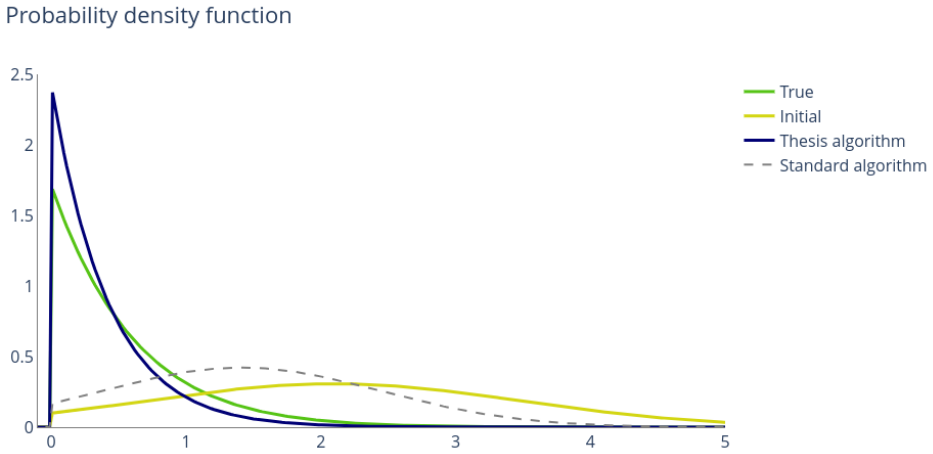

Figure 4.7: The initial and the final estimate of the probability density function (considering the thesis algorithm) for Experiment 3 compared to the true density which is a density of the truncated normal distribution with mean 8 and variance 25, truncated at the interval [0.40]. The grey dashed line shows the estimate when standard EM algorithm is used.

Experiment 4.

One cluster with mean outside the observation window depending on the number of simulated data points

From the previous three experiments, we can see that the more significant truncation is, the worse the estimates we get. This leads us to the question if the size of the dataset can improve those estimates. We take the parameters from Experiment 3 with the difference that we will increase the number of generated data points. We take $N = 400$ as the starting size of our dataset and then increase N by 100 up to 2000. For each N we generate one dataset so the datasets are mutually independent.

We are interested in how the estimates of a mean and a variance change with increasing N. The results are summarized in Table 4.4. The first interesting observation is that the number of iterations for the thesis algorithm remains the same for all N, we only need 2 iterations until the stopping criterion is met, whereas the article algorithm needs more than twice as many iterations for $N = 400$ compared to the case with $N = 2000$.

Figures 4.8 and 4.9 shows the plots of the estimates depending on the number of data points. We can observe that for a small dataset (up to 600 data points), both algorithms tend to overestimate the variance and place the mean below the true value. For more data points, the estimate of the mean remains above the true mean. However it seems that it stays at a reasonable distance from the true value. The same holds for the estimate of the variance (it stays below the true variance but reasonably close).

$K = 1$, $\mu_1 = -8$, $\sigma_1^2 = 25$, $s = 0$, $t = 40$

	Thesis algorithm			Article algorithm		
N	No. iterations	$\hat{\mu}_1$	$\hat{\sigma}_1^2$	No. iterations	$\hat{\mu}_1$	$\hat{\sigma}_1^2$
400	2	−15.66	42.97	1279	−15.65	42.94
500	2	−12.04	33.91	1001	−12.03	33.9
600	2	−8.20	25.81	681	−8.20	25.80
700	2	−5.83	20.28	509	−5.82	20.27
800	2	−6.68	21.82	585	−6.68	21.81
900	2	−6.08	20.51	543	−6.08	20.51
1000	2	−5.26	19.07	477	−5.25	19.07
1100	2	−5.36	19.34	487	−5.36	19.34
1200	2	−4.51	17.65	423	−4.50	17.65
1300	2	−5.22	18.99	482	−5.22	18.99
1400	2	−5.76	20.15	526	−5.76	20.15
1500	2	−6.15	21.28	556	−6.15	21.27
1600	2	−5.81	20.35	533	−5.80	20.35
1700	2	−5.93	20.44	549	−5.93	20.43
1800	2	−5.54	19.62	518	−5.53	19.61
1900	2	−5.27	19.07	498	−5.27	19.07
2000	2	−4.64	17.82	449	−4.64	17.82

Table 4.4: Summary of Experiment 4.

We have to keep in mind that for each selected dataset with given number of data points, we present only one realisation of the respective random sample, for different realisations it could turn out differently. We will address this variability for different realisations in Experiment 5.

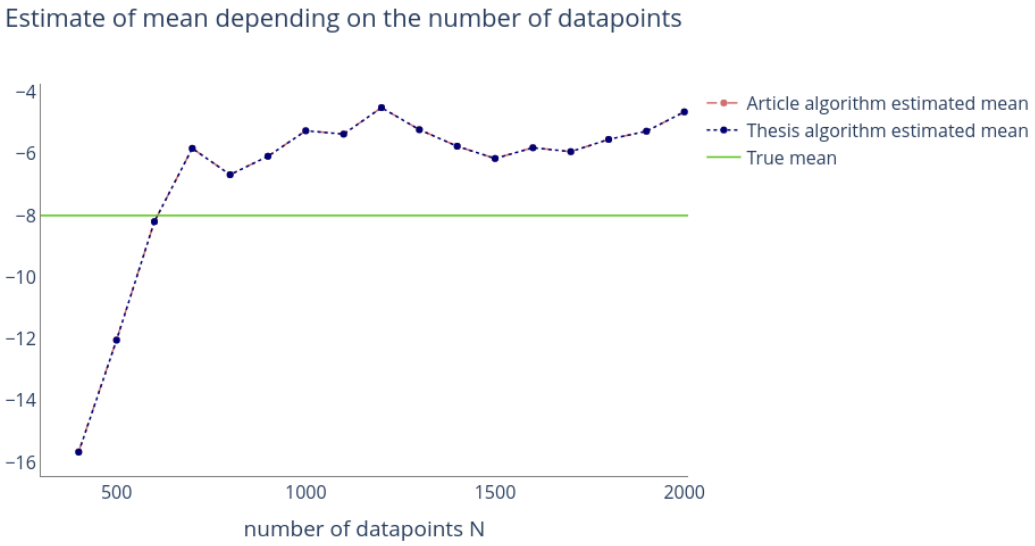

Figure 4.8: Estimates of the mean for Experiment 4 depending on the number of simulated data points considering the article algorithm and the thesis algorithm. The true value of the mean is also highlighted.

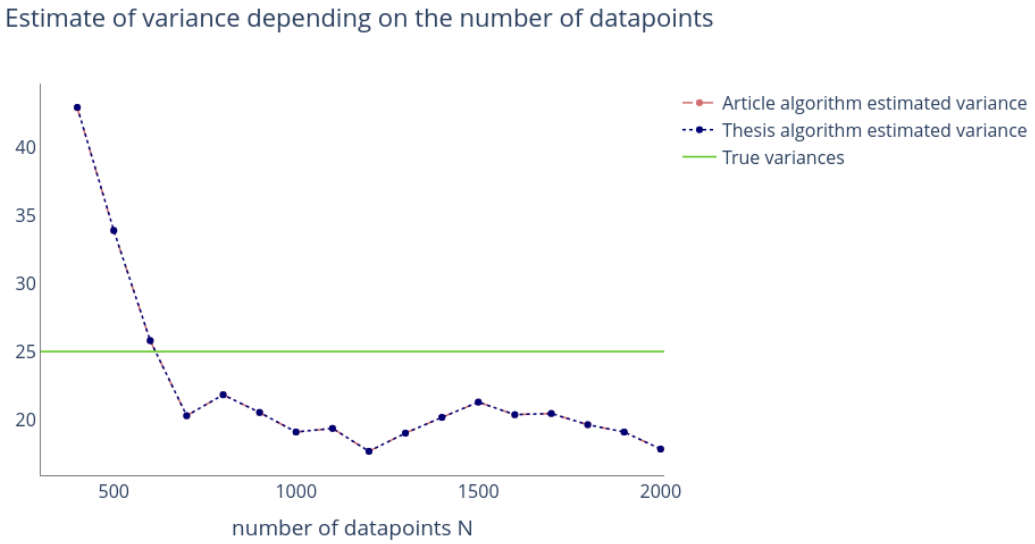

Figure 4.9: Estimates of the variance for Experiment 4 depending on the number of simulated data points considering the article algorithm and the thesis algorithm. The true value of the variance is also highlighted.

Experiment 5.

A simulation study of one cluster with mean outside the observation window

We generate 500 datasets drawn from the same distribution, a Gaussian mixture with one cluster on a bounded window $[0, 40]$ with mean $\mu_1 = -8$ and variance $\sigma_1^2 = 25$. Each dataset contains 150 points. We will be interested in the behaviour of the thesis and the article algorithms. In the previous two experiments, we saw datasets where the results are reasonable and no issues occur. However this is not always the case as we will see.

For 25 datasets, the thesis algorithm reached the point where it was unable to find the solution for the updated parameters in M step (see (2.14)) after 10000 iterations of selected optimization algorithm. In 22 cases out of those 25, the optimization algorithm failed in the first iteration. The problem is that the estimated mean converges to some negative number (smaller than -120) with the estimated standard deviation between 14 and 25 and so the integral in the denominator of density function $\int_0^{40} f(x; \boldsymbol{\theta})\, dx$ is numerically zero. The histogram of estimated means with thesis algorithm if we omit mentioned 25 cases where the optimization failed is in Figure 4.10. The corresponding descriptive statistics are in Table 4.5. While some estimates of the mean are unreasonably small, the average value of the estimated mean is -10.4336 which is fairly close to -8, given that we only observe points above 0.

Let us take a look at the article algorithm. As mentioned in subsection 2.3.1, the article algorithm cannot guarantee the estimate of variance to be a positive value (or a positive semi-definite matrix in case of higher dimensions). We can observe such pathological occurrences in this experiment too, for 8 datasets, the final estimate of the variance is negative. For another 14 cases, the algorithm was not able to deliver the final results, as it failed when the updated parameters after M step are such that the integral in the denominator in equation (2.5) is numerically zero. Figure 4.11 deptics the histogram of the estimated mean, excluding the 25 cases where the optimization failed. The corresponding descriptive statistics are in Table 4.6. We observe very similar results as for the thesis algorithm.

As a summary of Experiment 5, we cannot say that one approach performs better than the other one if we take pathological cases (whre we do not obtain the final parameters or obtain unreasonable estimates) into account. In the cases where the article algorithm delivered negative variances, the thesis algorithm failed to find the solution when performing the optimization in M step so the outcome is the same, we are unable to estimate the final distribution via the EM algorithm. In conclusion, both approaches are not 100% reliable, however in most cases, the results are good enough.

count	mean	std	min	25%	50%	75%	max
475	-10.4336	13.5871	-124.5635	-11.9126	-6.4007	-3.2731	0.9529

Table 4.5: Experiment 5: Descriptive statistics of the estimated mean using the thesis algorithm.

count	mean	std	min	25%	50%	75%	max
478	-10.5526	12.9910	-81.2337	-11.9777	-6.3977	-3.2782	0.9531

Table 4.6: Experiment 5: Descriptive statistics of estimated mean using the article algorithm.

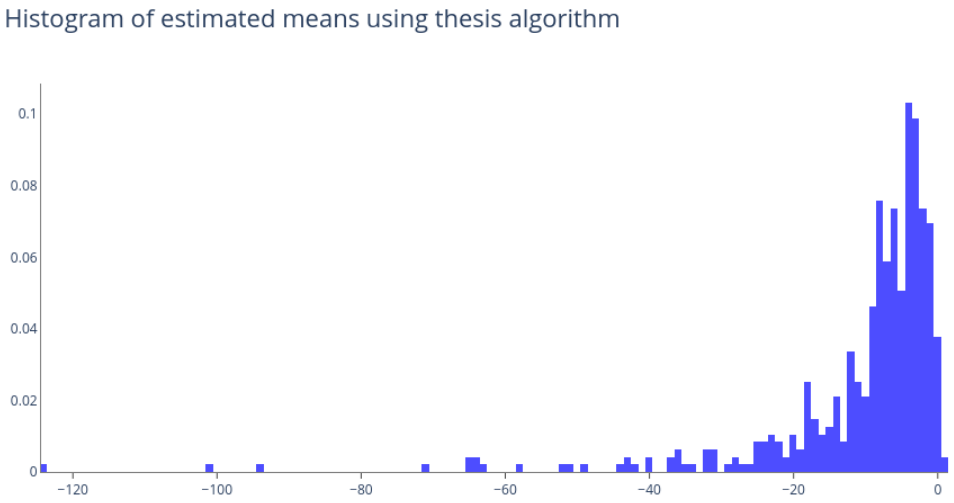

Figure 4.10: Histogram of the estimated means using thesis algorithm.

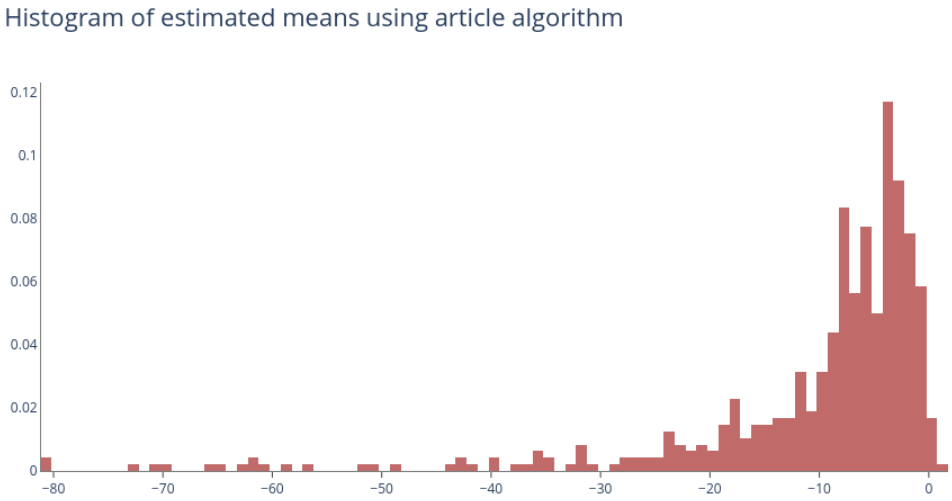

Figure 4.11: Histogram of the estimated means using the article algorithm.

Experiment 6.

Two clusters with means inside the observation window depending on the size of overlap.

Now we consider two clusters with means inside the observation window. We investigate three possibilities depending on how much the probability densities are overlapping. For all three cases, we fix $\sigma_1^2 = \sigma_2^2 = 10$, $\mu_2 = 20$ and the weights for an untruncated mixture, $\pi_1 = 0.6$, $\pi_2 = 0.4$. Mixing weights η_k, $k = 1, 2$, for a respective truncated mixture are calculated using equation (2.7). For the first case, we choose $\mu_1 = 5$ which leads us to two clusters with an insignificant overlap, see Figure 4.12. Next, we put $\mu_1 = 10$ such that the two clusters have a visible overlap, see Figure 4.13. For the last case, we set $\mu_1 = 15$, the two clusters have a significant overlap, see Figure 4.14.

We expect the last case to be more complicated for both algorithms because we cannot easily separate clusters from each other. On the other hand, the first case is similar to single-cluster examples. Simply said, we could separate the clusters, i.e. assign each point to one of the cluster, and then perform single-cluster EM algorithm for each cluster separately. We will not do that, it is just for explanation, why it is not more interesting than Experiment 1.

The results are summarized in Table 4.7. Apparently, the more significant the overlap is, the more complicated the estimation is. As expected, both algorithms converge faster for the case with insignificant overlap. We can also see that estimates of the weights are more accurate for this case, it is easier for the EM algorithm to split the dataset more precisely. With increasing overlapping area, both algorithms tend to overestimate the weight of the first component which goes hand in hand with overestimating its mean and variance. The higher the component weight, the higher the probability that a point in the overlapping area belongs to that cluster. So these middle points in the overlapping area move the mean of the first component to the right. With that, the variance increases. Analogously, with decreasing weight of the second component, the estimate of the variance is more and more underestimated and the mean is moving to the right bound.

In general, we are not able to tell which of these two overlapping clusters will be preferred by the algorithm (i.e. assign the higher weight to it). In Experiment A1 listed in Appendix we have a similar situation with $\mu_1 = 15$ where both algorithms still favour the first component. However the opposite situation when the estimates of the means will be both shifted to the left can occur as well, see Experiment A2 also listed in Appendix, where we set with $\mu_1 = 18$.

We can see more examples with two clusters in Appendix.

$N = 500$, $K = 2$, $\mu_2 = 20$, $s = 0$, $t = 40$, $\sigma_1^2 = 10$, $\sigma_2^2 = 10$

$\mu_1 = 5$, $\eta_1 = 0.5859$, $\eta_2 = 0.4141$

Thesis algorithm $\hat{\mu}_1 = 5.2311$, $\hat{\sigma}_1^2 = 7.9037$, $\hat{\eta}_1 = 0.5941$
$\hat{\mu}_2 = 19.8622$, $\hat{\sigma}_2^2 = 8.4719$, $\hat{\eta}_2 = 0.4059$

Article algorithm $\hat{\mu}_1 = 5.2311$, $\hat{\sigma}_1^2 = 7.9037$, $\hat{\eta}_1 = 0.5941$
$\hat{\mu}_2 = 19.8622$, $\hat{\sigma}_2^2 = 8.4719$, $\hat{\eta}_2 = 0.4059$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	12	21.9929 s	-1533.4939	0.0086
Article algorithm	17	29.8616 s	-1533.4939	0.0086

$\mu_1 = 10$, $\eta_1 = 0.5998$, $\eta_2 = 0.4002$

Thesis algorithm $\hat{\mu}_1 = 10.3411$, $\hat{\sigma}_1^2 = 10.3826$, $\hat{\eta}_1 = 0.6479$
$\hat{\mu}_2 = 20.3790$, $\hat{\sigma}_2^2 = 6.5891$, $\hat{\eta}_2 = 0.3521$

Article algorithm $\hat{\mu}_1 = 10.3411$, $\hat{\sigma}_1^2 = 10.3826$, $\hat{\eta}_1 = 0.6479$
$\hat{\mu}_2 = 20.3791$, $\hat{\sigma}_2^2 = 6.5891$, $\hat{\eta}_2 = 0.3521$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	85	153.8340 s	-1523.7437	0.0104
Article algorithm	87	135.5842 s	-1523.7437	0.0104

$\mu_1 = 15$, $\eta_1 = 0.6$, $\eta_2 = 0.4$

Thesis algorithm $\hat{\mu}_1 = 15.5909$, $\hat{\sigma}_1^2 = 9.9131$, $\hat{\eta}_1 = 0.7577$

$\hat{\mu}_2 = 21.2226$, $\hat{\sigma}_2^2 = 4.6477$, $\hat{\eta}_2 = 0.2423$

Article algorithm $\hat{\mu}_1 = 15.5911$, $\hat{\sigma}_1^2 = 9.9137$, $\hat{\eta}_1 = 0.7577$

$\hat{\mu}_2 = 21.2228$, $\hat{\sigma}_2^2 = 4.6473$, $\hat{\eta}_2 = 0.2423$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	349	464.1091 s	-1373.2374	0.0156
Article algorithm	369	402.9779 s	-1373.2374	0.0156

Table 4.7: Summary of Experiment 6.

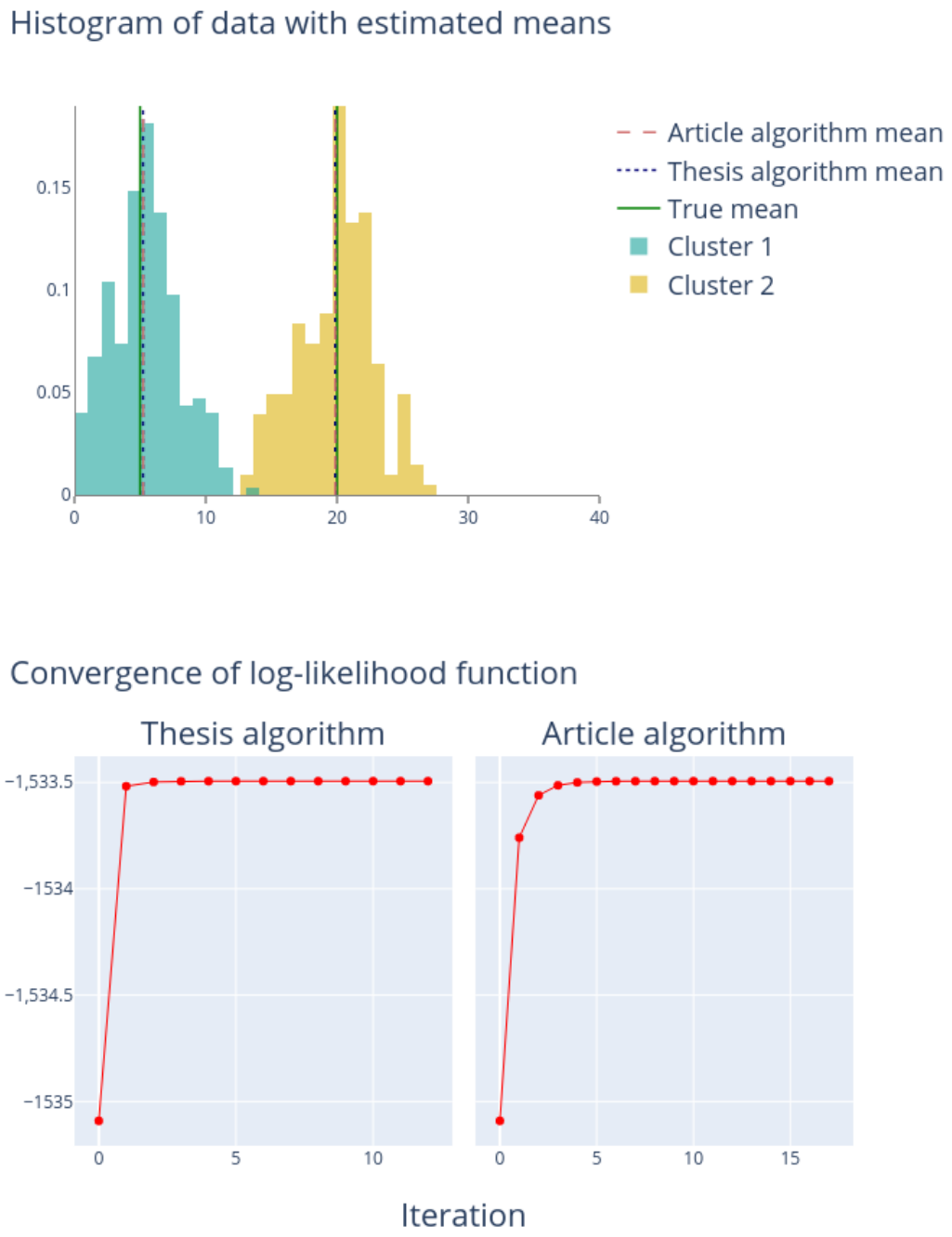

Figure 4.12: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 5$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

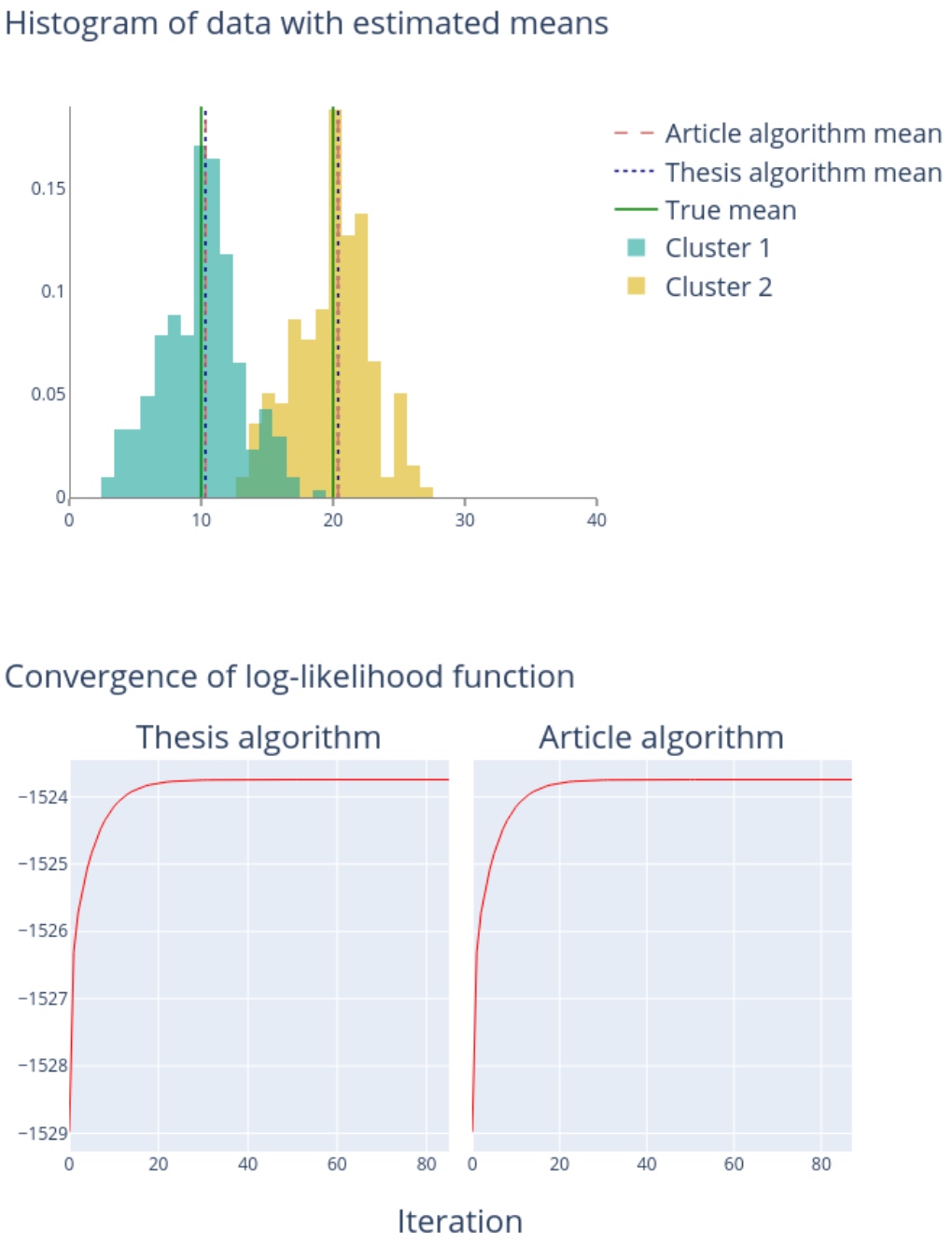

Figure 4.13: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 10$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

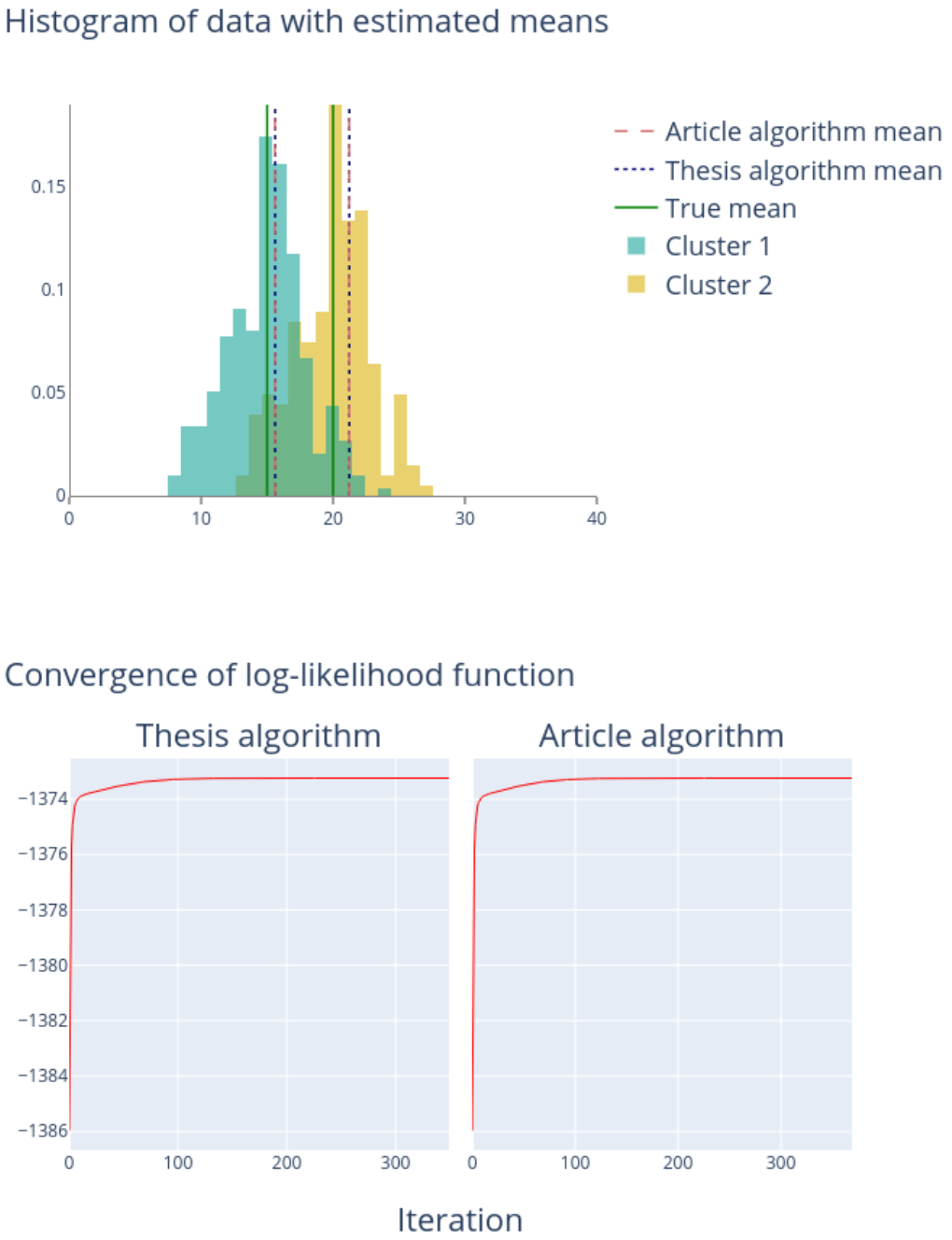

Figure 4.14: The experiment with one-dimensional synthetic data. Data comes from a Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 15$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

4.1.4 Two-dimensional case

Let us now consider two-dimensional data. We fix the observation window for all examples in this subsection to a bounded rectangle window $[\boldsymbol{s}, \boldsymbol{t}] = [(0,0),(25,25)]$. For each experiment, we calculate the KL score with $N_C = 2000$ in equation (4.1). As in the previous section concernin one-dimensional data, we will be interested not only in the estimate of the distribution itself, but in the estimates of parameters, focusing on the cluster means.

In Experiment 7, we have a simple example with the mean located in the middle of the observation window. We will compare the case with a spherical covariance matrix and the case where a correlation between the cluster components is assumed. Then we shift the cluster mean to the edge of the observation window in Experiment 8. We end the single-cluster examples with the cluster having the mean outside the observation window in Experiment 9. Finally, we show an example of a pathological case, see Experiment 10.

Experiment 7.

One cluster with mean in the middle of the observation window

We again start with the simplest example, one cluster with the mean located in the middle of the observation window, $\boldsymbol{\mu}_1^T = (12.5, 12.5)$. With the additional dimension, we have the possibility to consider, in addition to the variance, also the correlation between the variable's dimensions, which will change the shape of the distribution. For illustrative purposes, let us choose two different posibilities for the covariance matrix $\boldsymbol{\Sigma}_1$. The first one will be an isotropic (spherical) covariance matrix, $\begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$, the second will have correlated components, $\begin{pmatrix} 20 & -12 \\ -12 & 20 \end{pmatrix}$. Analogously as for one-dimensional case, see Experiment 1, we do not expect to see any significant differences between the thesis and the article algorithm, for the isotropic as well as the correlated covariance matrix, both algorithms should get equally good results as when using the standard EM algorithm.

Figure 4.15 shows the results for the isotropic case, in Figure 4.16 is its correlated version. For both simulations, 100 data points were simulated. Let us remind again that we generate the exact amout of data points located in the observation window using `R` library `truncnorm` with unknown number of data points outside the observation window, these data were truncated. At the top of the figures, we have scatter plots of our simulated data together with the true mean (left), the estimated mean using the thesis algorithm (middle) and using the article algorithm (right). In addition to the true and the estimated means, we show the true, respectively the estimated, 95% confidence covariance ellipses. At the bottom of both plots, the observed log-likelihood function calculated after each iteration in the thesis and in the article algorithm is plotted. What is worth to mention is that the log-likelihood decreases each iteration of the article algorithm. As we mentioned before, we do not have any theory behind the article algorithm as for the thesis algorithm so we are unable to ensure increasing (respective non-decreasing) log-likelihood in each iteration and this experiment is its example where the log-likelihood does not behave as in the EM algorithm theory, see Theorem 1. However, in terms of how the algorithm estimates the final parameters, this decrease in the log-likelihood does not cause any significant deviation from the parameters, both algorithms perform really well in finding the true distribution.

The summarized results are shown in Table 4.8 for uncorrelated case and in Table 4.9 for its correlated version. For both versions of covariance matrix, the article algorithm needs more steps in order to met the stopping criterion, namely 9 iterations in the isotropic version and 11 iterations in the correlated version. The thesis algorithm is stopped after 2, respectivelly 3 iterations. However, from the time point of view both algorithms spend only about 1 s with the computation.

Let us think about the reasons why the article algorithm worsen the log-likelihood.

As we saw in one-dimensional experiments, the decreasing in the log-likelihood did not happen so we assume this is some pathological example. Assume the isotropic case for now. The initial estimate of the mean, produced by the K-means algorithm, is

$$\hat{\mu}_{1,init}^T = (12.1143, 12.5392)$$

which is already a truely good estimate. The Euclidean distance between the initial value of the mean estimate and the true mean is 0.3877. The distance between the final estimate using the article algorithm and the true mean is 0.3991. So the article algorithm slightly worsen the initial estimate of the mean which goes hand in hand with decreasing log-likelihood. From the theory of the EM algorithm we know that we do not have ensured the convergence to the global maxima and it could happen that for some initial parameters, we can obtain estimates which do not correspond to the global maxima of the log-likelihood. However, this is obviously not the case as for the thesis algorithm we do not observe such behaviour of the log likelihood.The distance between the final estimate of the mean using the thesis algorithm and the true mean is 0.3776 so the thesis EM algorithm does help to improve the estimate of the mean. So in this case, the article algorithm demonstrates poor behaviour compared to the thesis algorithm.

$N = 100$, $K = 1$, $\boldsymbol{\mu}_1^T = (12.5, 12.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$

Thesis algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (12.1263, 12.5542)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 17.2416 & -1.6026 \\ -1.6026 & 16.7910 \end{pmatrix}$

Article algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (12.103, 12.5412)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 17.5729 & -1.6541 \\ -1.6541 & 17.0078 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	0.6643 s	-565.4234	0.0010
Article algorithm	9	1.1941 s	-565.4381	0.0009

Table 4.8: Summary of Experiment 7 with a spherical covariance matrix.

$N = 100$, $K = 1$, $\boldsymbol{\mu}_1^T = (12.5, 12.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & -12 \\ -12 & 20 \end{pmatrix}$

Thesis algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (12.1238, 12.6717)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 18.3338 & -12.0244 \\ -12.0244 & 17.9838 \end{pmatrix}$

Article algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (12.1053, 12.6724)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 18.8285 & -12.4279 \\ -12.4279 & 18.3852 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	3	0.8594 s	-543.0206	0.0027
Article algorithm	11	1.2627 s	-543.0410	0.0026

Table 4.9: Summary of Experiment 7 with correlated components.

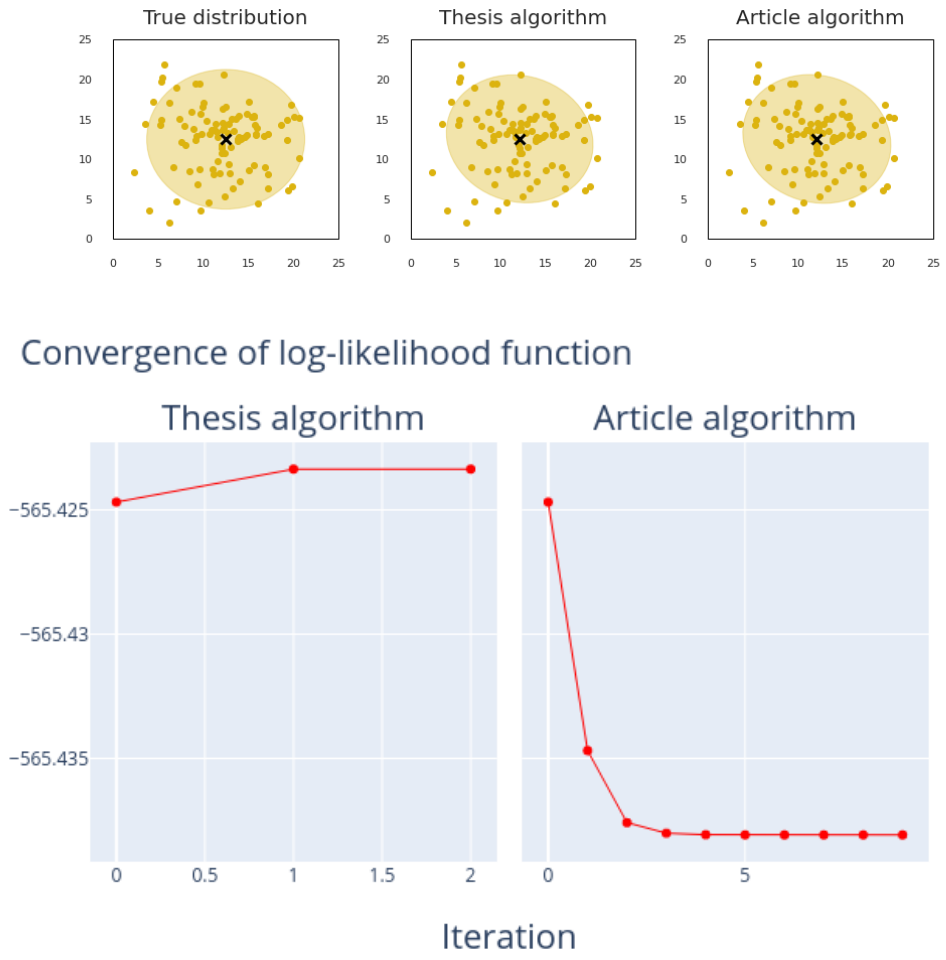

Figure 4.15: The experiment with two-dimensional synthetic data. Data comes from a Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (12.5, 12.5)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 100 data points were simulated. The scatter plot (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

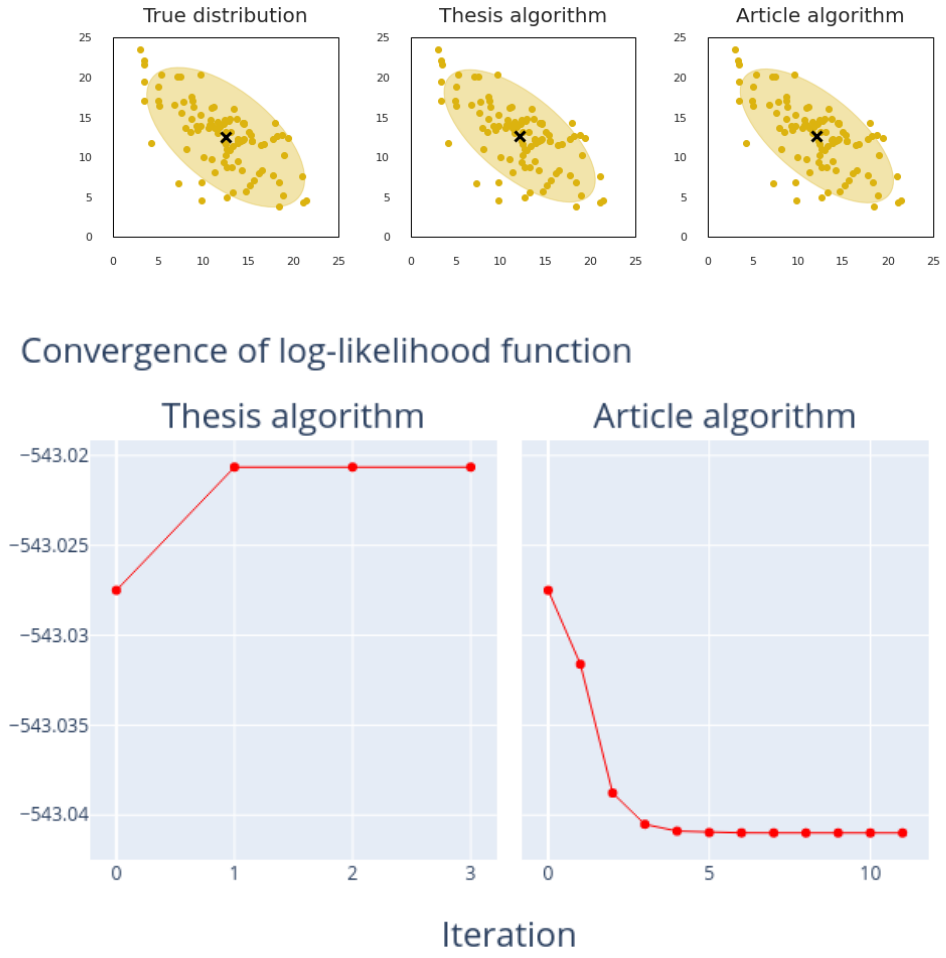

Figure 4.16: The experiment with two-dimensional synthetic data. Data comes from a Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (12.5, 12.5)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & -12 \\ -12 & 20 \end{pmatrix}$. In total, 100 data points were simulated. The scatter plot (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

$N = 200$, $K = 1$, $\boldsymbol{\mu}_1^T = (25, 23)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & -6 \\ -6 & 20 \end{pmatrix}$

Thesis algorithm $\hat{\boldsymbol{\mu}}_1^T = (26.6097, 21.9138)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 23.4305 & -8.9496 \\ -8.9496 & 23.767 \end{pmatrix}$

Article algorithm $\hat{\boldsymbol{\mu}}_1^T = (26.6085, 21.9145)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 23.4261 & -8.9476 \\ -8.9476 & 23.7668 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	1.2262 s	-881.334	-0.0005
Article algorithm	229	36.2994 s	-881.334	-0.0005

Table 4.10: Summary of Experiment 8.

Experiment 8.

One cluster with mean at the borders of the observation window

Now we consider an example with the mean located at the border of the observation window, specifically we set $\boldsymbol{\mu}_1^T = (25, 23)$. As we saw in Experiment 7, having a correlated covariance matrix does not change the course of the EM algorithm so from now on we consider only the correlated covariance matrix, the corresponding isotropic examples can be found in 4.2. We choose $\Sigma_1 = \begin{pmatrix} 20 & -12 \\ -12 & 20 \end{pmatrix}$ and generate 200 data points.

Analogously to the one-dimensional case (Experiment 2) the truncation makes it more difficult for the article algorithm to converge. However, we do not observe anything unexpected with one-dimensional example. The results are shown in Figure 4.17 and in Table 4.10.

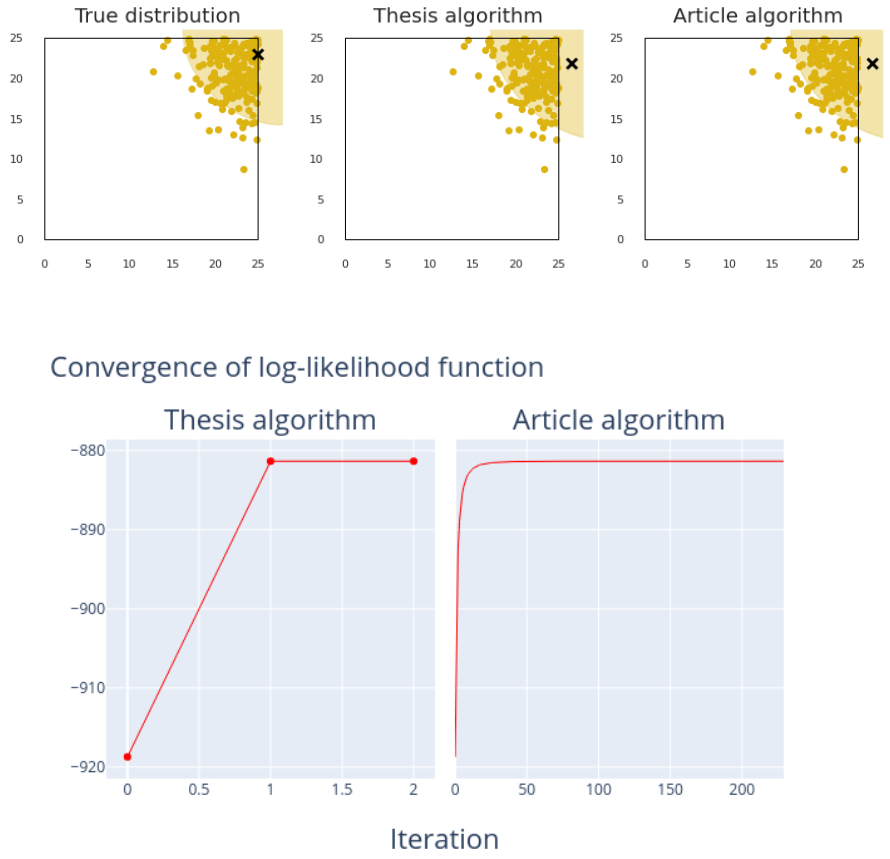

Figure 4.17: The experiment with two-dimensional synthetic data. Data comes from a Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (25, 23)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & -6 \\ -6 & 20 \end{pmatrix}$. In total, 200 data points were simulated. The scatter plot (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

$N = 250$, $K = 1$, $\boldsymbol{\mu}_1^T = (30, 12.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 12 \\ 12 & 20 \end{pmatrix}$

Thesis algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (32.3202, 12.6368)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 24.4978 & 10.9871 \\ 10.9871 & 16.8278 \end{pmatrix}$

Article algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (32.8516, 13.2679)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 26.1228 & 13.0533 \\ 13.0533 & 19.3531 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	2.0256 s	-1107.6211	-0.0014
Article algorithm	825	234.3416 s	-1108.2365	-0.0016

Table 4.11: Summary of Experiment 9.

Experiment 9.

One cluster with mean outside the observation window

Let us consider an example with the mean located outside the observation window, specifically we put $\boldsymbol{\mu}_1^T = (30, 12.5)$ with $\Sigma_1 = \begin{pmatrix} 20 & 12 \\ 12 & 20 \end{pmatrix}$ and generate 250 data points. The results are shown in Figure 4.18 and in Table 4.11.

$$\hat{\mu}_{1,init}^T = (22.8131, 8.3729)$$

Again, this experiment does not bring any new observations when compared to the one-dimensional case in Experiment 3.

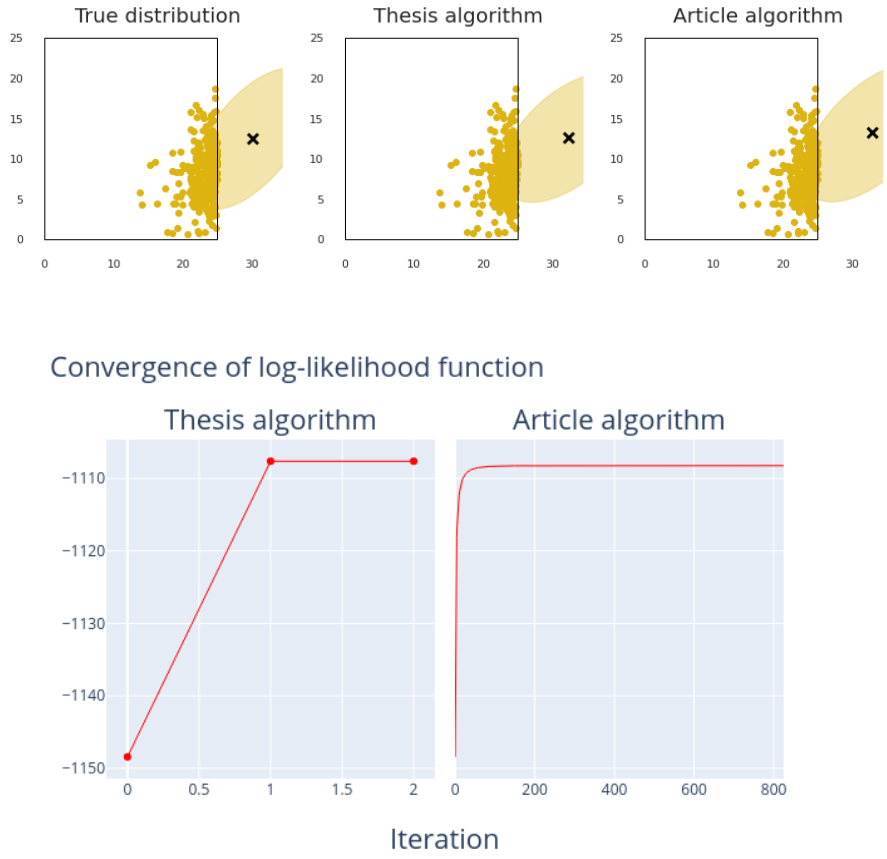

Figure 4.18: The experiment with two-dimensional synthetic data. Data comes from a Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (30, 12.5)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 12 \\ 12 & 20 \end{pmatrix}$. In total, 250 data points were simulated. The scatter plot (top) of truncated data is shown together with the true mean and the estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment 10

Annihilation of one of the two clusters

As our last commented experiment, we choose one pathological case that occured during the investigation of various examples in two dimensions. Let $\boldsymbol{\mu}_1^T = (5, 15)$, $\boldsymbol{\mu}_2^T = (10, 10)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}$ and generate 500 data points.

The results can be seen in Figure 4.19 and in Table 4.12. We choose this particular example because for the article algorithm, the estimate of the first weight η_1 converges to zero (and so the estimate of the second weight is one). The complete-data log-likelihood for the article algorithm is increasing at first, then it starts to decrease up to the iteration where the first cluster has vanished which leads to a small increase in the complete-data log-likelihood. Afterwards, it remains the same and the algorithm is stopped. This behavior does not happen with the thesis algorithm which provides reasonable results.

We showed this example in order to illustrate the weak point of the article algorithm: the complete-data log-likelihood is not guaranteed to be well-behaved as is the case for the thesis algorithm. Our conclusion would be that the complete-data log-likelihood might behave unexpectedly.

$N = 500,\ K = 2$	$\boldsymbol{\mu}_1^T = (5, 15),\ \boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix},\ \eta_1 = 0.6$
	$\boldsymbol{\mu}_2^T = (10, 10),\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix},\ \eta_2 = 0.4$
Thesis algorithm	$\hat{\boldsymbol{\mu}}_1^T = (5.8074, 14.8263),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 10.102 & -0.817 \\ -0.817 & 4.6529 \end{pmatrix}$
	$\hat{\eta}_1 = 0.6240$
	$\hat{\boldsymbol{\mu}}_2^T = (10.5196, 10.0108),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 4.3952 & 2.104 \\ 2.104 & 18.0839 \end{pmatrix}$
	$\hat{\eta}_2 = 0.3760$
Article algorithm	$\hat{\boldsymbol{\mu}}_1^T = (-6.2156, 14.9474),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 1.9461 & -0.0365 \\ -0.0365 & 0.0019 \end{pmatrix}$
	$\hat{\eta}_1 = 0.0000$
	$\hat{\boldsymbol{\mu}}_2^T = (7.2926, 13.1321),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 15.3476 & -5.9657 \\ -5.9657 & 15.7114 \end{pmatrix}$
	$\hat{\eta}_2 = 1.0000$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	97	214.9704 s	-2631.1575	0.0007
Article algorithm	282	290.8407 s	-2711.7988	-0.0016

Table 4.12: Summary of Experiment 10.

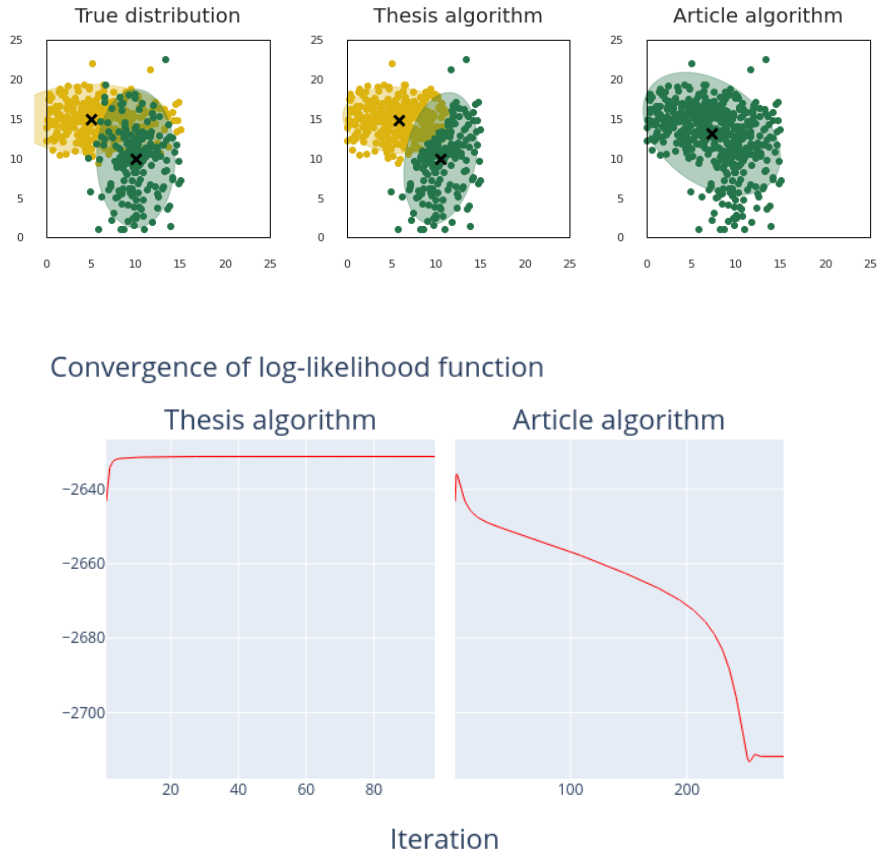

Figure 4.19: The experiment with two-dimensional synthetic data. Data comes from Gaussian mixture truncated in each dimension at interval $[0, 25]$ with two components with means $\boldsymbol{\mu}_1^T = (5, 15)$, $\boldsymbol{\mu}_2^T = (10, 10)$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and covariance matrices $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 500 data points were simulated. The scatter plot (top) of truncated data is shown together with true mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

4.2 Real dataset

In this sub-chapter we discuss the real dataset **California Redwoods Point Pattern (Ripley's Subset)** provided in `R` library `spatstat`. Its subset we will work with represents locations of 62 Californian redwood seedlings and saplings. The observation window is after rescaling a bounded unit square. The plot of this dataset is in Figure 4.20.

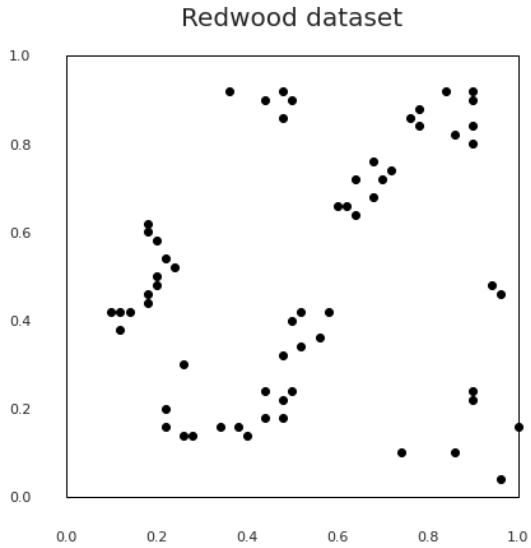

Figure 4.20: **Redwood dataset:** the location of 62 seedlings and samplings of California Giant Redwood.

Number of clusters: We compare two before-mentioned information criteria to determine the optimal number of clusters, it is $AICc$ and BIC, see section 3.1.2. We choose the corrected version of AIC as the sample size is small, we have only 62 data points. We run the algorithm with one up to seven clusters and calculate these two criteria. The reason we do not run the calculation for more than 7 clusters lies in its numerical instability. We have only 62 points, each with the precision of 2 decimals places. This could quite easily lead to a cluster with only two points for which will be the covariance matrix singular (or numericaly almost singular and thus useless as a covariance matrix estimate). In practice, we choose the number of components for which the one or the other criterion reached the minimum value and/or it does not decrease a lot (relatively to the rest of the values). It is not always easy to determine such number of components.

For the simulations for given number of clusters, we again choose the same initial parameters for each of the three approaches (standard EM algorithm, thesis EM algorithm and article algorithm) obtained by K-means algorithm. The final results are not sensitive to changing the initial random seed, we run the algorithms with different seeds and looked at the initial parameters, as well as the final parameters and basically there were no differences. We use the stopping criterion based on the difference between the observed log-likelihood function with threshold $\delta = 10^{-6}$.

The resulting values of both criteria for each approach are shown in Figure 4.2. Using the corrected AIC, we have an obvious match for both, the thesis algorithm and the article algorithm, in determining the number of clusters, we choose $K = 4$. For the standard EM algorithm, we would probably choose $K = 5$. For BIC, the decision is not that obvious. Presumably, for the thesis algorithm, we choose again $K = 4$, for the article and the standard EM algorithm we may choose $K = 5$.

In Figure 4.2 and Figure 4.2, we have the final estimates of cluster means and 95% confidence ellipses using the standard EM algorithm, the thesis algorithm and the article algorithm depending on the number of clusters.

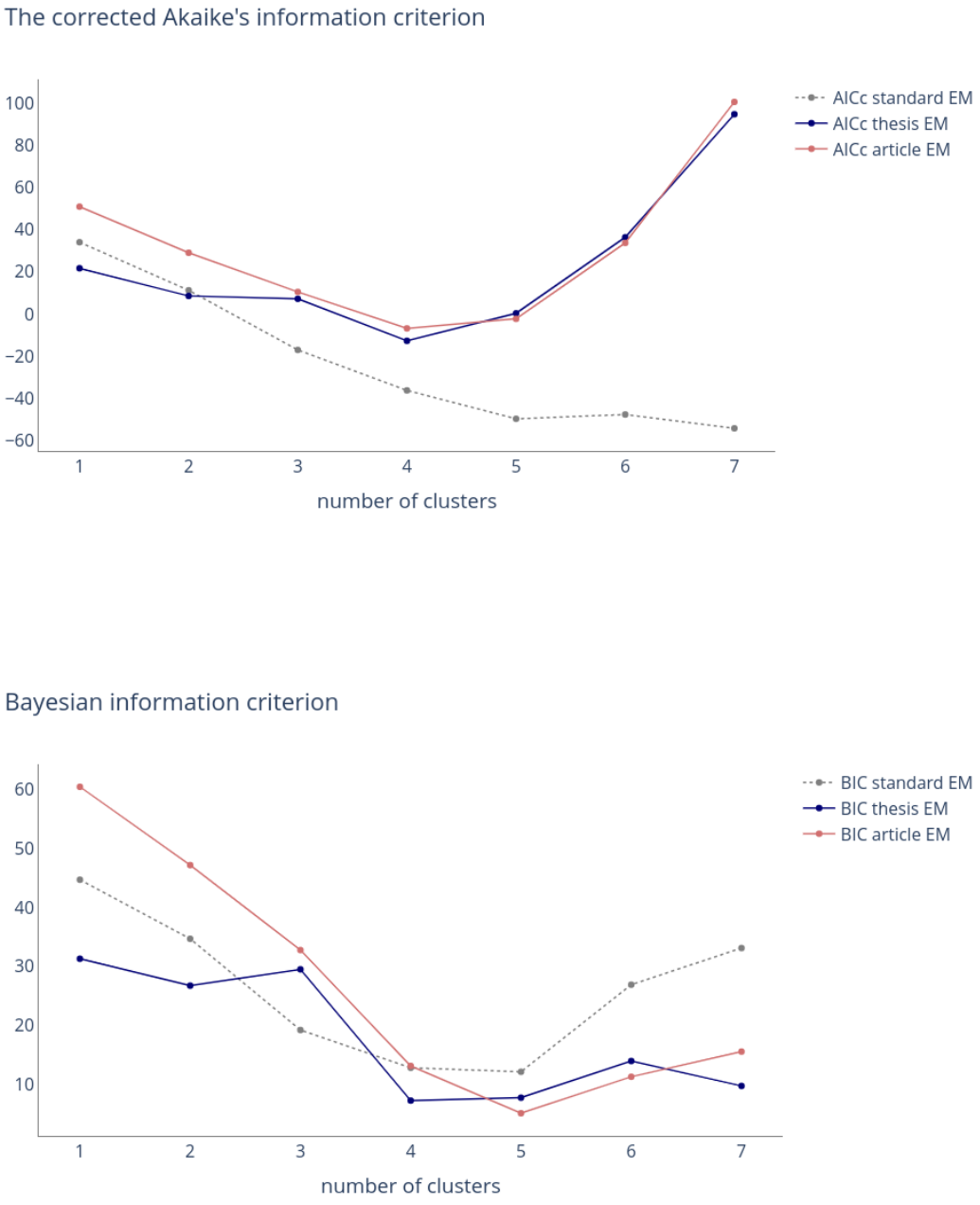

Figure 4.21: The values of AIC (top) and BIC (bottom) depending on the number of clusters for Redwood dataset using the standard EM algorithm, the algorithm described in this thesis and the algorithm described in the article Lee and Scott [2012].

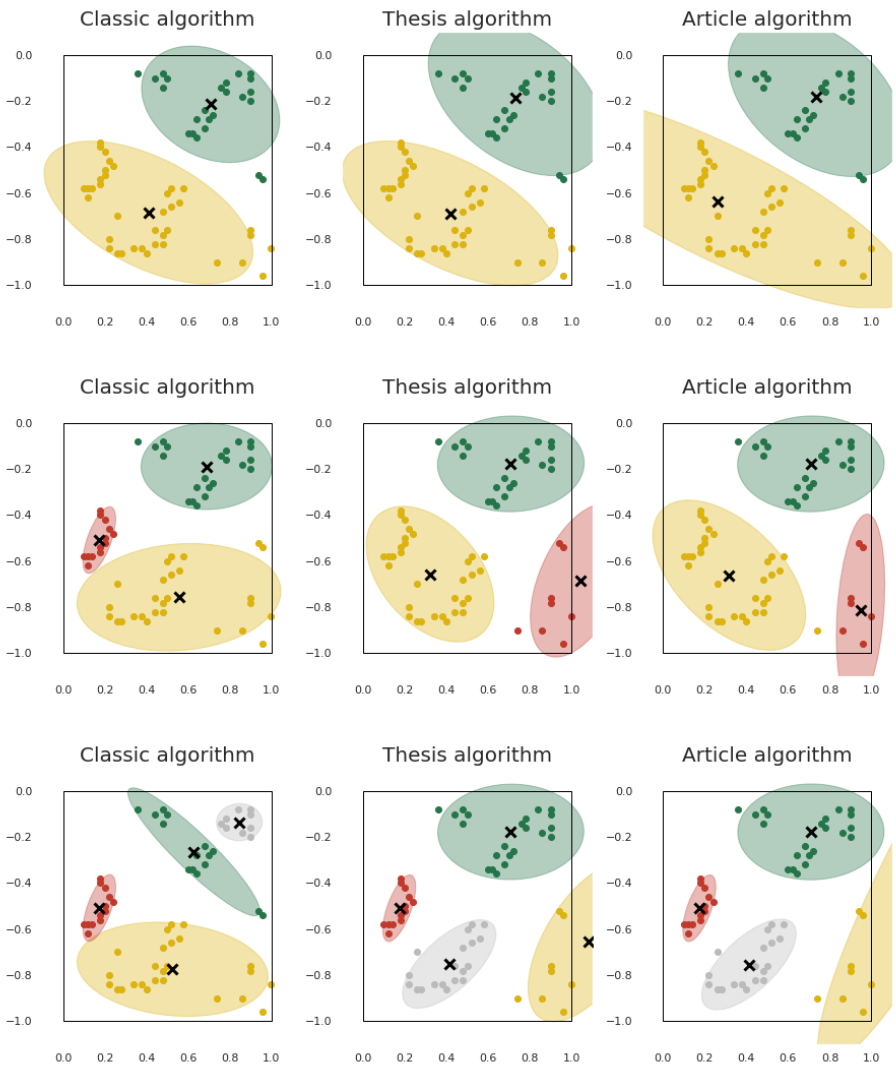

Figure 4.22: Real dataset: scatter plots together with estimates of means and 95% confidence ellipses using the standard EM algorithm, the thesis algorithm and the article algorithm for different number of clusters (upper: $K = 2$, middle: $K = 3$, bottom: $K = 4$).

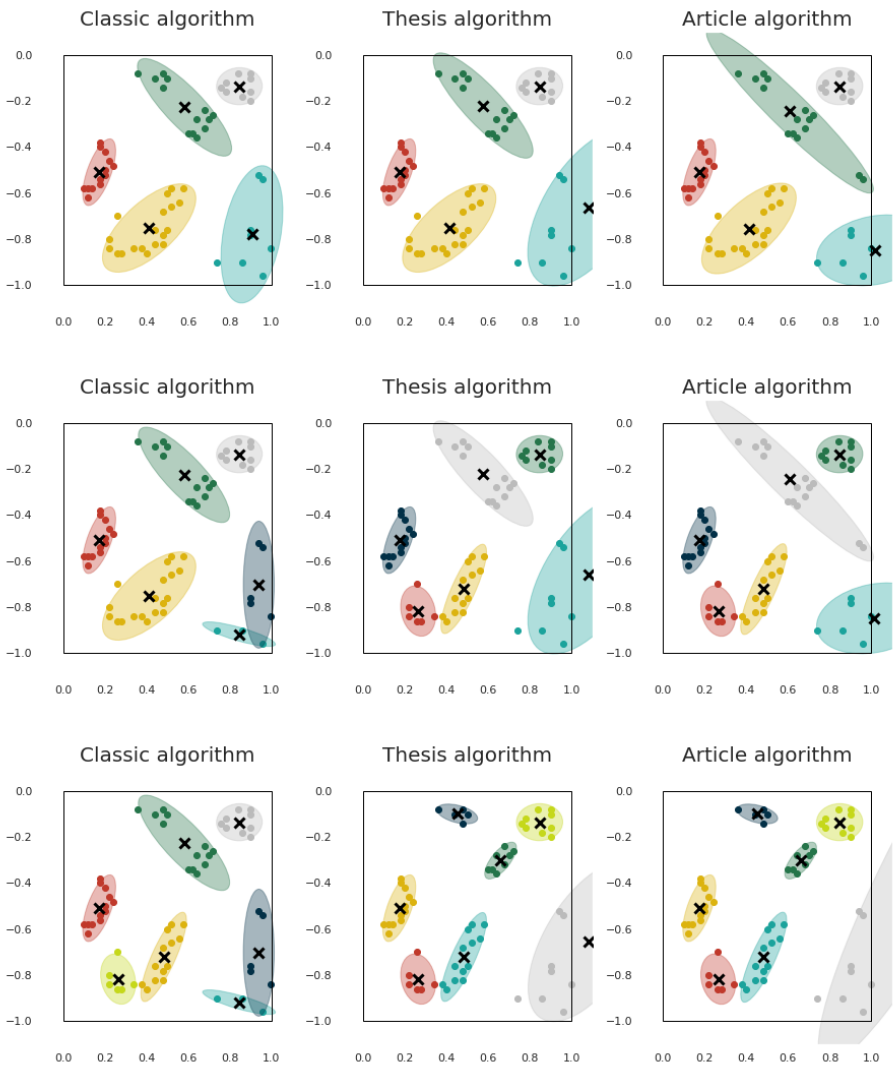

Figure 4.23: Real dataset: scatter plots together with estimates of means and 95% confidence ellipses using the standard EM algorithm, the thesis algorithm and the article algorithm for different number of clusters (upper: $K = 5$, middle: $K = 6$, bottom: $K = 7$).

Conclusion

This thesis introduced the application of the EM algorithm to the Gaussian mixture distribution when we are restricted to a bounded rectangle observation window. While there are numerous publications on the EM algorithm for Gaussian mixtures, this is not the case for its truncated version.

With the help of the EM algorithm, we can get an important insight on the process of interest. We have studied the article Lee and Scott [2012] in detail as it is one of the few resources on this topic available. However, the proposed method in the mentioned article has a weak point, it is not based on the general theory of the EM algorithm and thus we cannot rely on it. As we saw in the last chapter of this thesis, in the more complex and complicated examples, this simplified version of the EM algorithm does not give us faster convergence, nor better results than the proposed algorithm in this thesis. We therefore find it preferrable to use the algorithm based on the general EM-algorithm theory, described in this thesis.

We can extend the work done in this thesis in multiple ways. From the practical point of view, the algorithm derived in this thesis could be used in a real application, such as the flow cytometry dataset analysed in Lee and Scott [2012]. The library written for this thesis could also be released as as an open-source library, or utilized for a contribution into an existing library.

Bibliography

Revlin Abbi, Elia El-Darzi, Christos Vasilakis, and Peter Millard. *Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay.* IEEE, Varna, September 2008. ISBN 9781424417391.

Hirotugu Akaike. A new look at the statistical model identification. In *Springer Series in Statistics*, pages 215–222. Springer New York, 1974.

J Anděl. *Základy matematické statistiky.* Matfyzpress, Praha, 2007.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4): 561–575, jan 2003.

Pierre Brémaud. *Probability Theory and Stochastic Processes.* Springer International Publishing, April 2020. ISBN 3030401820.

Joseph E. Cavanaugh and Andrew A. Neath. The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs Computational Statistics*, 11(3), mar 2019.

A. Dempster. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

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

David Peel Geoffrey J. McLachlan. *Finite Mixture Models.* WILEY, October 2000. ISBN 0471006262.

John Geweke. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. *Comput. Sci. Statist.*, 23, 03 1998.

D H Griffel. *Linear algebra and its applications.* Ellis Horwood Series in Mathematics & Its Applications. Ellis Horwood Ltd, Publisher, Harlow, England, February 1989.

Debashis Kushary. The EM Algorithm and Extensions. *Technometrics*, 40(3): 260–260, aug 1998.

Gyemin Lee and Clayton Scott. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816–2829, sep 2012.

Lung-Fei Lee. On the first and second moments of the truncated multi-normal distribution and a simple estimator. *Economics Letters*, 3(2):165–169, jan 1979.

Mary J. Lindstrom and Douglas M. Bates. Newton-raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014, dec 1988.

Magnus. *Matrix Differential Calculus 3*. John Wiley & Sons, March 2019. ISBN 1119541204.

Geoffrey J. McLachlan and Suren Rathnayake. On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, sep 2014.

Volodymyr Melnykov and Igor Melnykov. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, 56(6):1381–1395, jun 2012.

Omelka. Modern statistical methods, nmst 434, course notes, 2021. URL `https://www2.karlin.mff.cuni.cz/~omelka/Soubory/nmst434/nmst434_course-notes_2020.pdf`.

Branislav Panić, Jernej Klemenc, and Marko Nagode. Improved Initialization of the EM Algorithm for Mixture Model Parameter Estimation. *Mathematics*, 8 (3):373, mar 2020.

Carl Edward Rasmussen. The Infinite Gaussian Mixture Model. *Proceedings of the 12th International Conference on Neural Information Processing Systems*, page 554–560, 1999.

Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), mar 1978.

Li-Xian Sun, Fen Xu, Yi-Zeng Liang, Yu-Long Xie, and Ru-Qin Yu. Cluster analysis by the k-means algorithm and simulated annealing. *Chemometrics and Intelligent Laboratory Systems*, 25(1):51–60, sep 1994.

Marco Taboga. Kullback-leibler divergence., 2021. URL `https://www.statlect.com/fundamentals-of-probability/Kullback-Leibler-divergence`.

Thomas M. Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, Nashville, TN, 2 edition, June 2006.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1), mar 1983.

List of Figures

List of Tables

Appendix

Theorem 3. *Assume that the random vector* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ *has* d*-dimensional normal distribution with mean* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$ *and covariance matrix* $\boldsymbol{\Sigma}$*. For* $K \subset \{1, \ldots, d\}$*, divide the vector* $\boldsymbol{X}$ *into* $\boldsymbol{X}_{\{k\in K\}}$ *and* $\boldsymbol{X}_{-\{k\in K\}}$ *where* $\boldsymbol{X}_{\{k\in K\}}$ *has elements* X_i*,* $i \in K$ *and* $\boldsymbol{X}_{-\{k\in K\}}$ *has elements* X_j *for* $j \in \{1, \ldots, d\} \setminus K$*. In the same way, divide the vector* $\boldsymbol{\mu}$ *into* $\boldsymbol{\mu}_{\{k\in K\}}$ *and* $\boldsymbol{X}_{-\{k\in K\}}$ *and the matrix* $\boldsymbol{\Sigma}$ *into four sub-matrices* $\boldsymbol{\Sigma}_{\{k\in K\},\{k\in K\}}$*,* $\boldsymbol{\Sigma}_{-\{k\in K\},-\{k\in K\}}$*,* $\boldsymbol{\Sigma}_{-\{k\in K\},\{k\in K\}}$*,* $\boldsymbol{\Sigma}_{\{k\in K\},-\{k\in K\}}$*. Then the conditional distribution of* $\boldsymbol{X}_{\{k\in K\}}$ *given* $\boldsymbol{X}_{-\{k\in K\}}$ *is the* k*-dimensional normal distribution with mean*

$$\boldsymbol{\mu}_{\{k\in K\}} + \boldsymbol{\Sigma}_{\{k\in K\},-\{k\in K\}}\boldsymbol{\Sigma}^{-1}_{-\{k\in K\},-\{k\in K\}}(\boldsymbol{X}_{-\{k\in K\}} - \boldsymbol{\mu}_{-\{k\in K\}})$$

and covariance matrix

$$\boldsymbol{\Sigma}_{\{k\in K\},\{k\in K\}} - \boldsymbol{\Sigma}_{\{k\in K\},-\{k\in K\}}\boldsymbol{\Sigma}^{-1}_{-\{k\in K\},-\{k\in K\}}\boldsymbol{\Sigma}_{-\{k\in K\},\{k\in K\}}.$$

Proof. See Anděl [2007].

□

Experiment A1.

Two clusters, with mean in the middle of the observation window

$N = 500$, $K = 2$, $s = 0$, $t = 40$	
	$\mu_1 = 15$, $\sigma_1^2 = 10$, $\eta_1 = 0.6$
	$\mu_2 = 20$, $\sigma_2^2 = 10$, $\eta_2 = 0.4$
Thesis algorithm	$\hat{\mu}_1 = 15.5909$, $\hat{\sigma}_1^2 = 9.9131$, $\hat{\eta}_1 = 0.7577$
	$\hat{\mu}_2 = 21.2226$, $\hat{\sigma}_2^2 = 4.6477$, $\hat{\eta}_2 = 0.2423$
Article algorithm	$\hat{\mu}_1 = 15.5911$, $\hat{\sigma}_1^2 = 9.9137$, $\hat{\eta}_1 = 0.7577$
	$\hat{\mu}_2 = 21.2228$, $\hat{\sigma}_2^2 = 4.6473$, $\hat{\eta}_2 = 0.2423$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	349	464.1091 s	-1373.2374	0.0151
Article algorithm	369	402.9779 s	-1373.2374	0.0151

Table 4.13: Summary of Experiment A1.

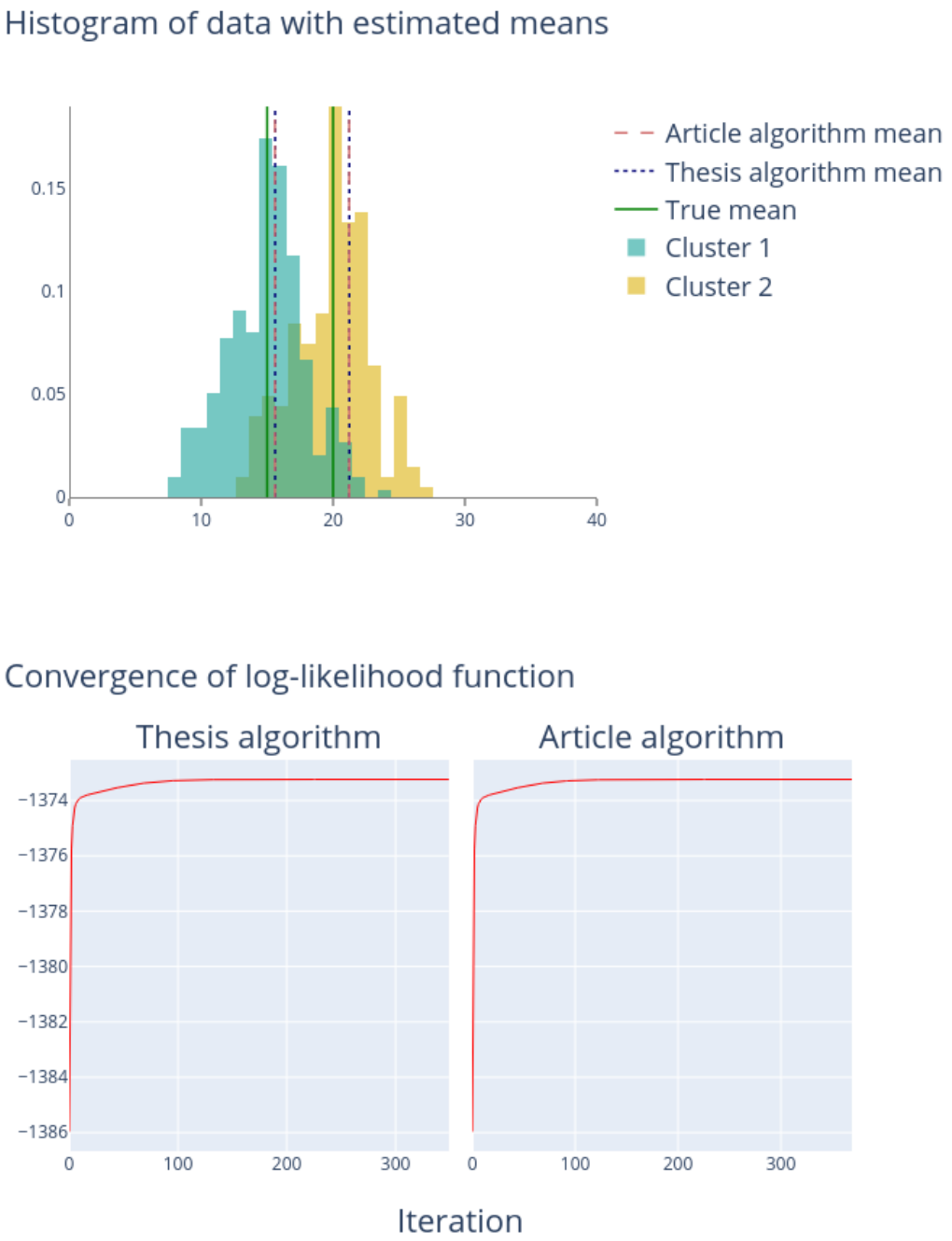

Figure 4.24: The experiment with one-dimensional synthetic data. Data comes from Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 15$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with real mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A2.

Two clusters, with mean in the middle of the observation window

$N = 500$, $K = 2$, $s = 0$, $t = 40$	
	$\mu_1 = 18$, $\sigma_1^2 = 10$, $\eta_1 = 0.6$
	$\mu_2 = 20$, $\sigma_2^2 = 10$, $\eta_2 = 0.4$
Thesis algorithm	$\hat{\mu}_1 = 13.7156$, $\hat{\sigma}_1^2 = 2.2846$, $\hat{\eta}_1 = 0.0825$
	$\hat{\mu}_2 = 19.1853$, $\hat{\sigma}_2^2 = 7.6779$, $\hat{\eta}_2 = 0.9175$
Article algorithm	$\hat{\mu}_1 = 13.7154$, $\hat{\sigma}_1^2 = 2.2843$, $\hat{\eta}_1 = 0.0825$
	$\hat{\mu}_2 = 19.1852$, $\hat{\sigma}_2^2 = 7.6782$, $\hat{\eta}_2 = 0.9175$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	669	893.8064 s	-1269.2279	0.0169
Article algorithm	683	789.0853 s	-1269.2279	0.0169

Table 4.14: Summary of Experiment A2.

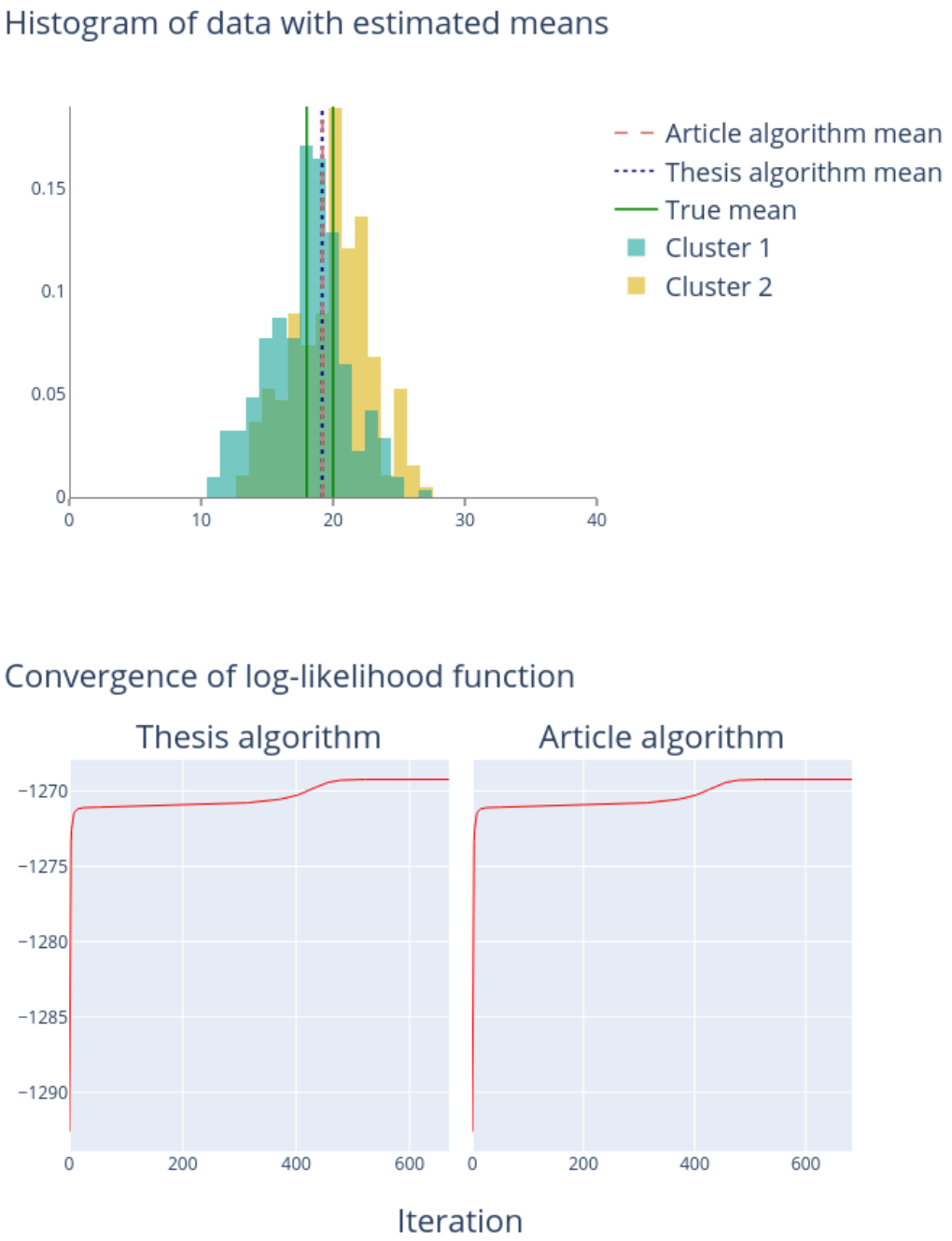

Figure 4.25: The experiment with one-dimensional synthetic data. Data comes from Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 18$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with real mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A3.

Two clusters, first one with mean at the border, second one with mean inside the observation window

$N = 500$, $K = 2$, $s = 0$, $t = 40$

$\mu_1 = 0$, $\sigma_1^2 = 10$, $\eta_1 = 0.4286$

$\mu_2 = 20$, $\sigma_2^2 = 10$, $\eta_2 = 0.5714$

Thesis algorithm $\hat{\mu}_1 = -0.9608$, $\hat{\sigma}_1^2 = 10.4582$, $\hat{\eta}_1 = 0.4199$

$\hat{\mu}_2 = 19.9374$, $\hat{\sigma}_2^2 = 8.9120$, $\hat{\eta}_2 = 0.5801$

Article algorithm $\hat{\mu}_1 = -0.9602$, $\hat{\sigma}_1^2 = 10.4569$, $\hat{\eta}_1 = 0.4199$

$\hat{\mu}_2 = 19.9374$, $\hat{\sigma}_2^2 = 8.912$, $\hat{\eta}_2 = 0.5801$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	5	12.4640 s	-1442.5232	0.0093
Article algorithm	170	261.8069 s	-1442.5232	0.0093

Table 4.15: Summary of Experiment A3.

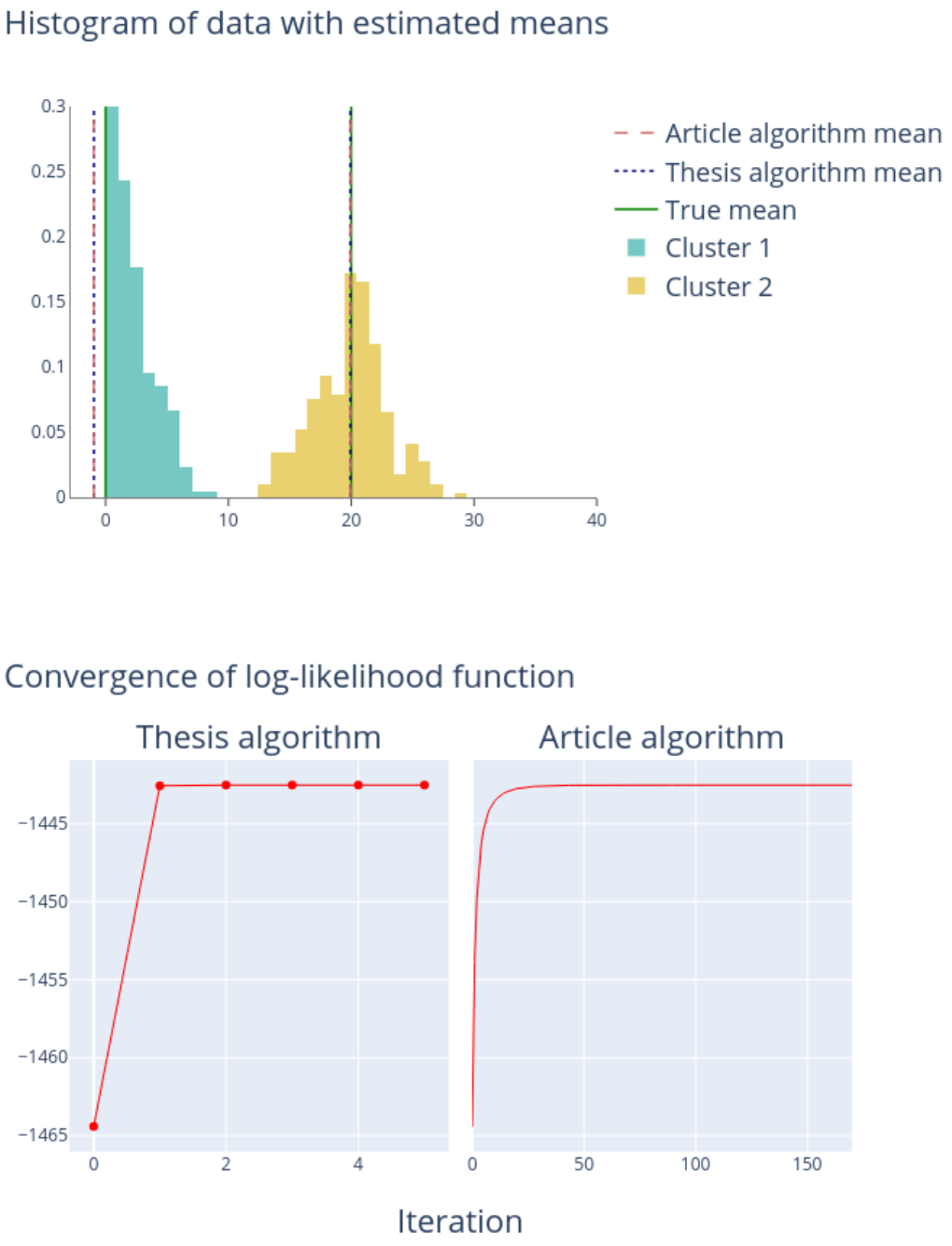

Figure 4.26: The experiment with one-dimensional synthetic data. Data comes from Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = 0$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with real mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A4.

Two clusters, first one with mean outside the observation window, second one with mean inside the observation window

$N = 500$, $K = 2$, $s = 0$, $t = 40$

$$\mu_1 = -3,\ \sigma_1^2 = 10,\ \eta_1 = 0.2045$$

$$\mu_2 = 20,\ \sigma_2^2 = 10,\ \eta_2 = 0.7955$$

Thesis algorithm $\hat{\mu}_1 = -1.0908,\ \hat{\sigma}_1^2 = 6.2728,\ \hat{\eta}_1 = 0.198$

$$\hat{\mu}_2 = 19.9627,\ \hat{\sigma}_2^2 = 8.7363,\ \hat{\eta}_2 = 0.802$$

Article algorithm $\hat{\mu}_1 = -1.0900,\ \hat{\sigma}_1^2 = 6.2715,\ \hat{\eta}_1 = 0.1980$

$$\hat{\mu}_2 = 19.9627,\ \hat{\sigma}_2^2 = 8.7363,\ \hat{\eta}_2 = 0.8020$$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	4	9.8523 s	-1397.9866	0.0049
Article algorithm	185	284.7992 s	-1397.9866	0.0049

Table 4.16: Summary of Experiment A4.

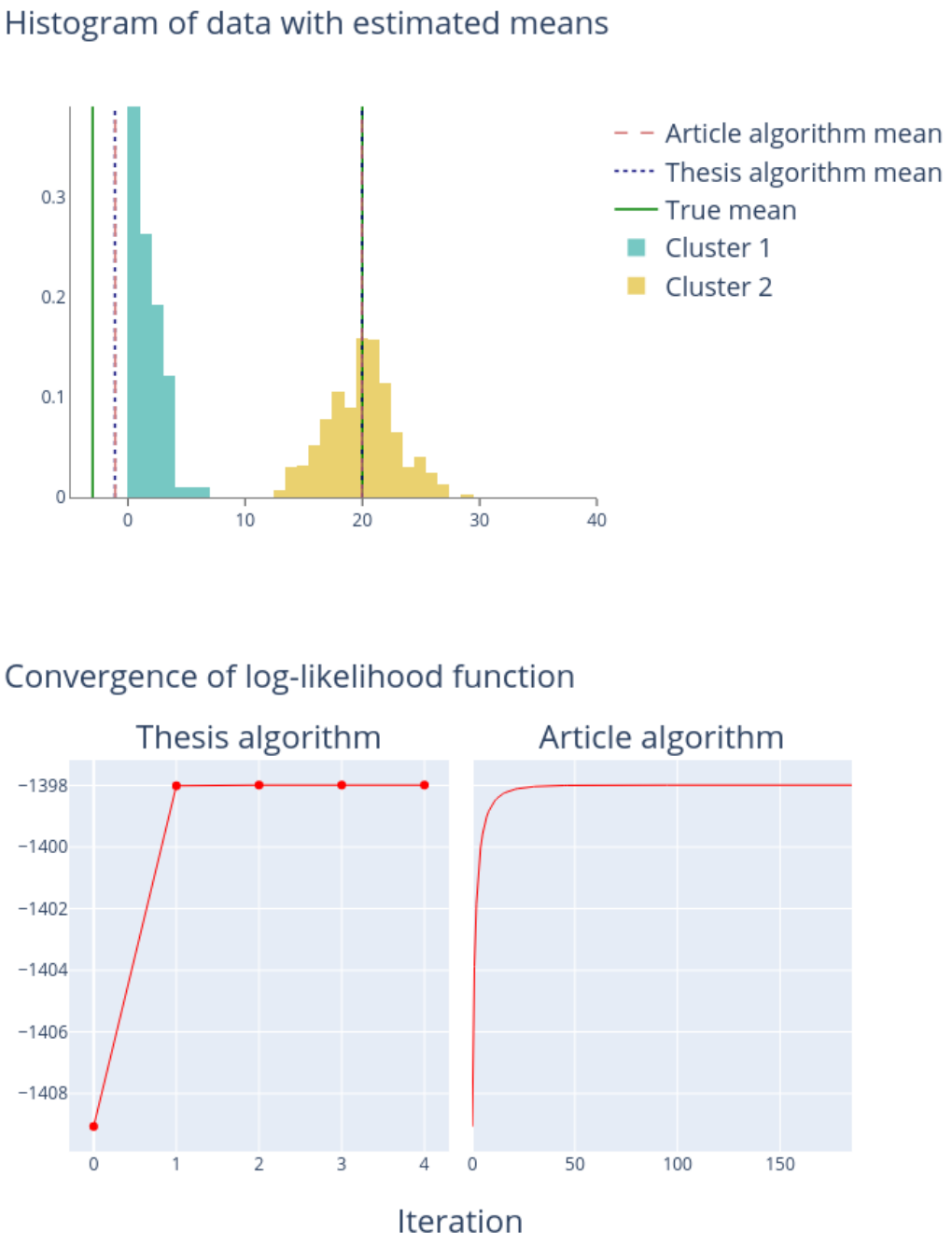

Figure 4.27: The experiment with one-dimensional synthetic data. Data comes from Gaussian mixture truncated at interval $[0, 40]$ with two components with means $\mu_1 = -3$, $\mu_2 = 20$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and common variance $\sigma^2 = 10$. In total, $N = 500$ data points were simulated. The histogram (top) of truncated data is shown together with real mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A5.

One cluster with mean at the borders of the observation window

$N = 200,\ K = 1,\ \boldsymbol{\mu}_1^T = (25, 23),\ \boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$

Thesis algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (24.9409, 23.7676),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 17.5992 & -0.0516 \\ -0.0516 & 24.7631 \end{pmatrix}$

Article algorithm $\quad \hat{\boldsymbol{\mu}}_1^T = (24.9404, 23.7682),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 17.5977 & -0.0515 \\ -0.0515 & 24.7666 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	3	1.3232 s	-918.58	0.0001
Article algorithm	115	18.0086 s	-918.58	0.0001

Table 4.17: Summary of Experiment A5.

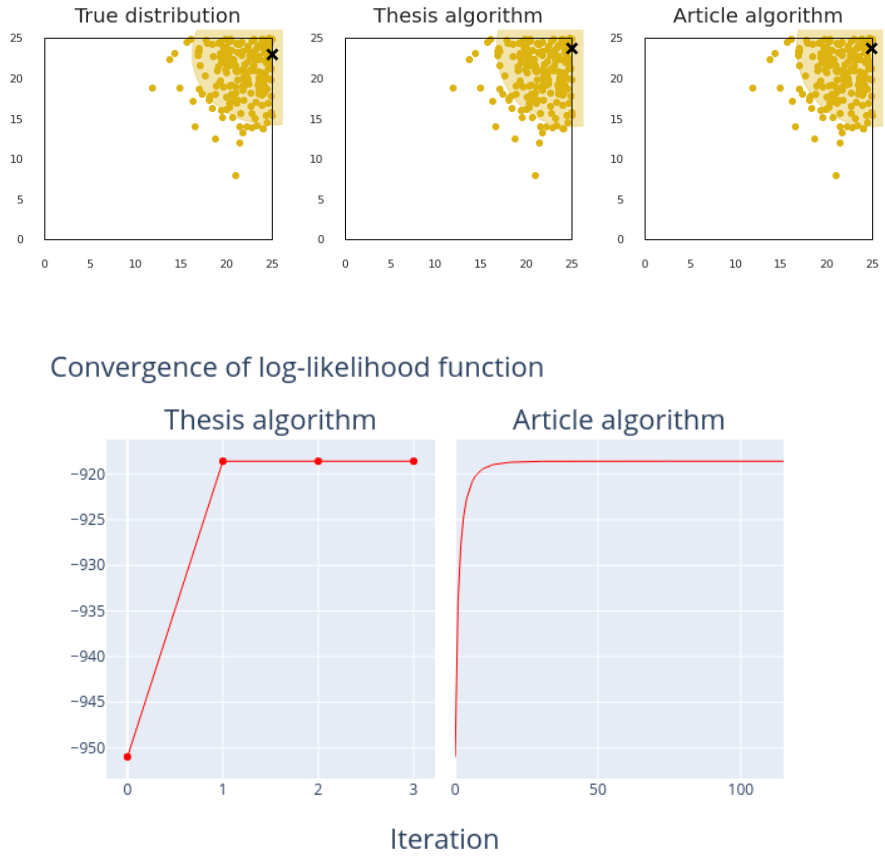

Figure 4.28: The experiment with two-dimensional synthetic data. Data comes from Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (25, 23)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 200 data points were simulated. The scatter plot (top) of truncated data is shown together with true mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A6.

One cluster with mean outside the observation window

$N = 250$, $K = 1$, $\boldsymbol{\mu}_1^T = (30, 12.5)$, $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$

Thesis algorithm $\hat{\boldsymbol{\mu}}_1^T = (32.3253, 11.7747)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 26.0491 & -2.1302 \\ -2.1302 & 17.7105 \end{pmatrix}$

Article algorithm $\hat{\boldsymbol{\mu}}_1^T = (32.3177, 11.7447)$, $\hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 26.0285 & -2.1596 \\ -2.1596 & 17.9893 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	2	1.3003 s	-1165.1878	-0.0005
Article algorithm	474	101.3228 s	-1165.2047	-0.0005

Table 4.18: Summary of Experiment A6.

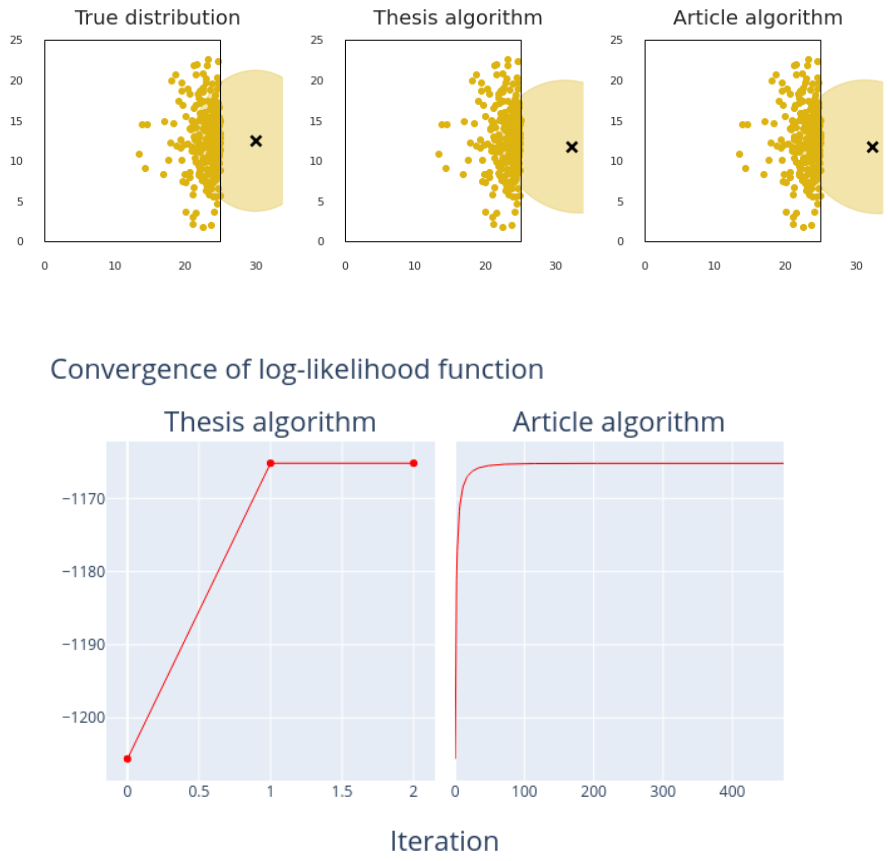

Figure 4.29: The experiment with two-dimensional synthetic data. Data comes from Gaussian mixture truncated in each dimension at interval $[0, 25]$ with one component with mean $\boldsymbol{\mu}_1^T = (30, 12.5)$ and variance matrix $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 250 data points were simulated. The scatter plot (top) of truncated data is shown together with true mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A7.

Two clusters with means inside the observation window with significant overlap

$N = 250,\ K = 2,$	$\boldsymbol{\mu}_1^T = (16, 16),\ \boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}, \pi_1 = 0.4$
	$\boldsymbol{\mu}_2^T = (12.5, 12.5),\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}, \pi_2 = 0.6$
Thesis algorithm	$\hat{\boldsymbol{\mu}}_1^T = (15.682, 15.8094),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 4.4594 & -1.0568 \\ -1.0568 & 19.9229 \end{pmatrix}$
	$\hat{\boldsymbol{\mu}}_2^T = (12.0637, 12.8596),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 15.6636 & -0.1046 \\ -0.1046 & 3.4416 \end{pmatrix}$
Article algorithm	$\hat{\boldsymbol{\mu}}_1^T = (15.6413, 15.7912),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 4.4362 & -1.1430 \\ -1.1430 & 20.6881 \end{pmatrix}$
	$\hat{\boldsymbol{\mu}}_2^T = (12.0464, 12.7972),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 15.5915 & -0.0502 \\ -0.0502 & 3.4728 \end{pmatrix}$

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	77	88.2472 s	-1302.9344	0.0027
Article algorithm	138	53.219 s	-1302.9615	0.0027

Table 4.19: Summary of Experiment A7.

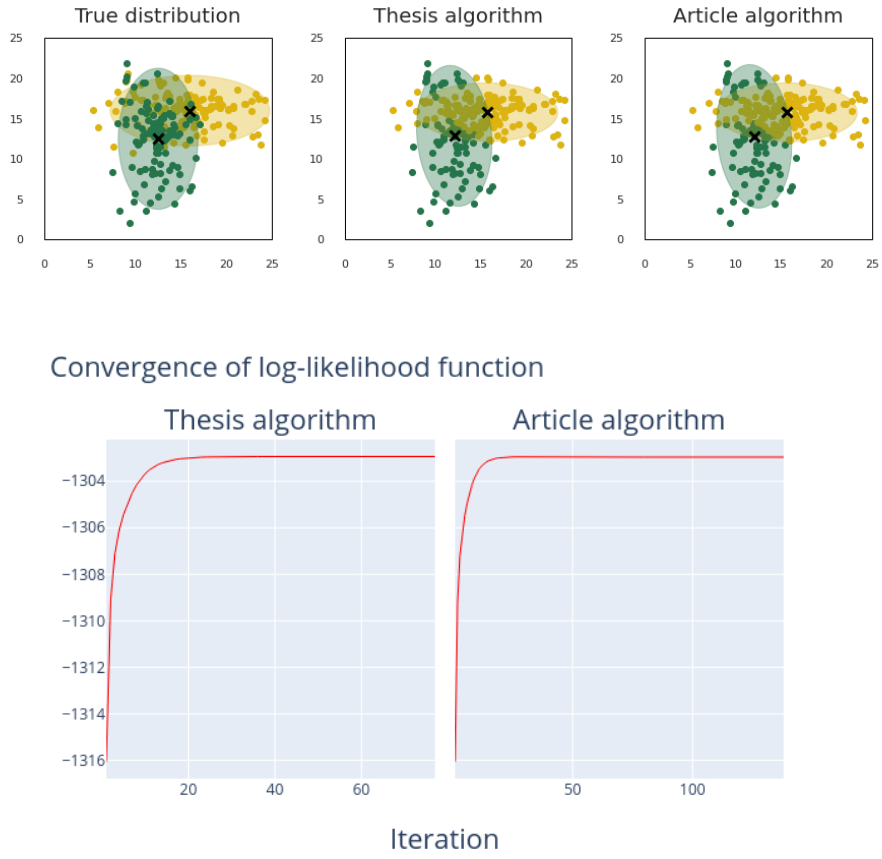

Figure 4.30: The experiment with two-dimensional synthetic data. Data comes from Gaussian mixture truncated in each dimension at interval $[0, 25]$ with two components with means $\boldsymbol{\mu}_1^T = (16, 16)$, $\boldsymbol{\mu}_2^T = (12.5, 12.5)$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and covariance matrices $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 250 data points were simulated. The scatter plot (top) of truncated data is shown together with true mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.

Experiment A8.

Two clusters with means inside the observation window with insignificant overlap

$N = 250,\ K = 2,$	$\boldsymbol{\mu}_1^T = (17, 16),\ \boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}$	
	$\boldsymbol{\mu}_2^T = (5, 10),\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}$	
Thesis algorithm	$\hat{\boldsymbol{\mu}}_1^T = (16.763, 16.1075),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 16.5389 & -0.7923 \\ -0.7923 & 4.4812 \end{pmatrix}$	
	$\hat{\boldsymbol{\mu}}_2^T = (4.9546, 10.2439),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 4.1754 & -0.6962 \\ -0.6962 & 15.2654 \end{pmatrix}$	
Article algorithm	$\hat{\boldsymbol{\mu}}_1^T = (16.7738, 16.111),\ \hat{\boldsymbol{\Sigma}}_1 = \begin{pmatrix} 16.4494 & -0.8194 \\ -0.8194 & 4.4660 \end{pmatrix}$	
	$\hat{\boldsymbol{\mu}}_2^T = (4.9076, 10.2015),\ \hat{\boldsymbol{\Sigma}}_2 = \begin{pmatrix} 4.5166 & -0.7421 \\ -0.7421 & 15.9203 \end{pmatrix}$	

	No. iterations	Time	Log-likelihood	KL
Thesis algorithm	20	79.0723 s	-1390.1911	0.0014
Article algorithm	42	49.762 s	-1390.4235	0.0010

Table 4.20: Summary of Experiment A8.

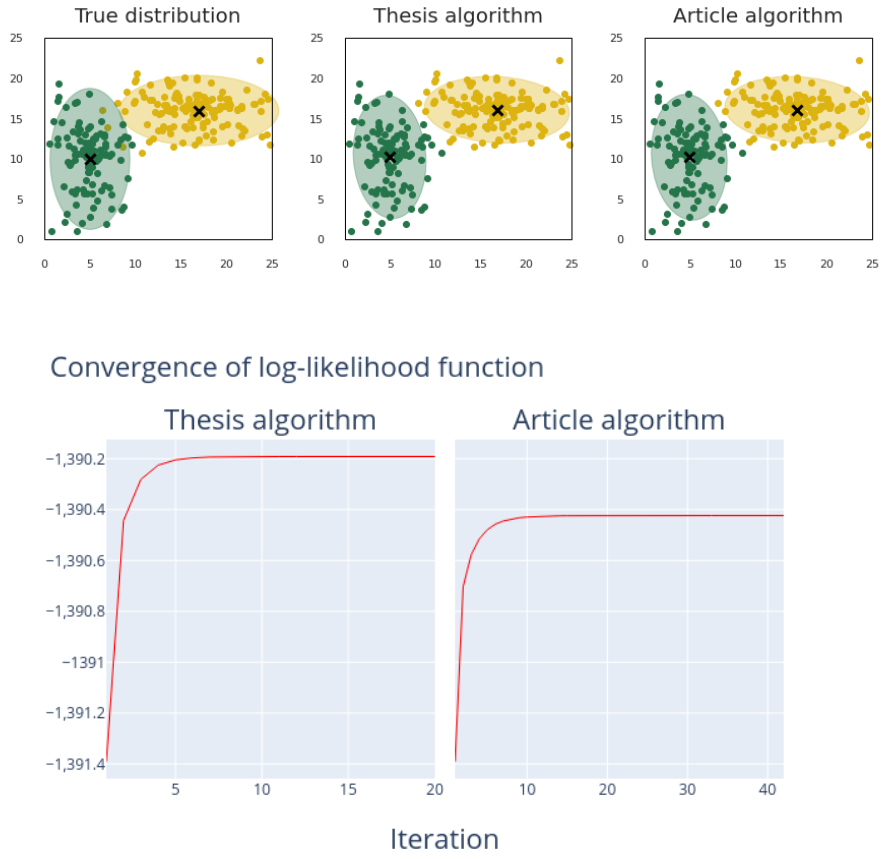

Figure 4.31: The experiment with two-dimensional synthetic data. Data comes from Gaussian mixture truncated in each dimension at interval $[0, 25]$ with two components with means $\boldsymbol{\mu}_1^T = (17, 16)$, $\boldsymbol{\mu}_2^T = (5, 10)$, weights $\pi_1 = 0.6$, $\pi_2 = 0.4$ and covariance matrices $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 20 & 0 \\ 0 & 5 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 5 & 0 \\ 0 & 20 \end{pmatrix}$. In total, 250 data points were simulated. The scatter plot (top) of truncated data is shown together with true mean and estimated means with the EM algorithm using the algorithm described in this thesis versus the algorithm described in the article Lee and Scott [2012]. At the bottom, we have graphs of the observed log-likelihood calculated at each iteration of the EM algorithm. The EM algorithm was stopped when the difference in the observed log-likelihood function was lower than 10^{-8}. The horizontal axis indicates the number of iterations until the stopping criterion was met.