Politecnico di Torino, *Department of Mathematical Sciences*
Eindhoven University of Technology, *Department of Mathematics and Computer Science*

Statistical Inference based on the Empirical Identity Process

Master's thesis

Author:
Ivo Stoepker

Supervisors:
E. Bibbona, *Politecnico di Torino*
M. Gasparini, *Politecnico di Torino*
R. M. Pires da Silva Castro, *Eindhoven University of Technology*

July 2018

Abstract

The empirical identity process (Enrico Bibbona, Giovanni Pistone, and Mauro Gasparini. The Empirical Identity Process: asymptotics and applications. *The Canadian Journal of Statistics*, 2017) gives rise to a test statistic d_n which can be used for statistical inference. In this thesis, a new statistic o_n based on the process is devised, and two inference settings are studied: goodness-of-fit and parameter estimation. For the latter, minimum-distance estimators are constructed. In the goodness-of-fit setting, we show that the new statistics are powerful in uniform settings with alternatives containing clusters. However, they are outperformed in other cases if the parameters of the null distribution are estimated from the data. In the parameter estimation setting, the minimum-distance estimator based on d_n is shown to have excellent performance, having similar performance as the maximum-likelihood estimator with the expectation-maximization algorithm in normal mixtures with high component overlap. Moreover, our minimum-distance estimators have excellent robustness properties in the normal mixture setting, especially compared with the maximum-likelihood estimator which is shown not to be robust. However, the minimum-distance estimator based on d_n is more sensitive to initialization.

Keywords: empirical identity process; goodness-of-fit; minimum-distance estimation; normal mixture distribution; robustness

Contents

List of Abbreviations

AD	Anderson-Darling (statistic).
CDF	Cumulative Distribution Function.
CvM	Cramèr-Von Mises (statistic).
ECDF	Empirical Cumulative Distribution Function.
EM	Expectation-Maximization (algorithm).
IEIP	Integrated Empirical Identity Process.
iid	Independent and Identically Distributed.
KS	Kolmogorov-Smirnov (statistic).
MDE	Minimum Distance Estimator.
MDE-Dist	Minimum Distance Estimator based on distance measure Dist.
MLE	Maximum Likelihood Estimator.
MSE	Mean-Squared Error.
MSP	Maximum Spacings (Estimator).
SW	Shapiro-Wilk (statistic).

List of Notations

$\xrightarrow{\text{a.s.}}$	Almost sure convergence.
$\xrightarrow{\text{P}}$	Convergence in probability.
$\Longrightarrow$	Weak convergence.
B_t	Standard Brownian bridge, i.e. $B_t = W_t - tW_1$, where W_t is standard Brownian motion.
F_n	Empirical Cumulative Distribution Function.
F_θ	Cumulative Distribution Function with parameters θ.
$F_n^{(\theta)}$	Empirical Cumulative Distribution Function of the probability integral transform $F_\theta(X)$, conditionally on θ.
$g_{\text{Dist}}(\theta)$	Optimization surface of the MDE based on Dist, such as d_n and o_n.
I_n	(Lower) Integrated Empirical Identity Process.
$I_n^{(\theta)}$	(Lower) Integrated Empirical Identity Process of the probability integral transform $F_\theta(X)$, conditionally on θ.
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2.
N	Number of samples used in a simulation.
n	Sample size used in a simulation.
W_t	Standard Brownian motion.

List of Figures

List of Tables

Chapter 1

Introduction

The empirical cumulative distribution function is a well-known non-parametric tool to estimate the distribution function of some sample data. The empirical quantile function is defined as its left-continuous generalized inverse. A recent paper [4] studies the process resulting from the composition of these two empirical functions.

By definition, the composition of the cumulative distribution function with the quantile function results in the identity. However, for the empirical versions, this is only asymptotically true, and the resulting function is named the empirical identity function. The second-order characteristics can then be studied by subtracting the identity function. This results in the *empirical identity process*.

The empirical identity process gives rise to a statistic, introduced in [4]. We also introduce our own statistic based on the empirical identity process. We examine their use for statistical inference in two cases: goodness-of-fit testing and parameter estimation.

In the goodness-of-fit setting, we study the statistics in two settings. In the first setting we specify the null distribution completely, including its parameter values. We then move to the more practical setting where we specify the null distribution without parameters. These parameters are subsequently estimated from the data. This results in a test where only the model, and not the parameters, need to be specified.

In the parameter estimation setting, we propose minimum-distance estimators. Since the statistics are heavily dependent on the spacings, they can be sensitive to clusters present in the data. Therefore, special attention is given

to the setting of normal mixtures, where clusters are expected. Furthermore, maximum-likelihood estimation is not directly applicable and powerful estimation techniques would be welcome in this setting. Additionally, since minimum-distance estimators are known to be robust, we study the robustness of our minimum-distance estimators.

The thesis is structured as follows. In Chapter 2, we survey relevant literature, starting with the paper on the empirical identity process. In Chapter 3, we devise our own statistic based on the empirical identity process. In Chapter 4, the statistics are used in a goodness-of-fit setting. In Chapter 5, we apply the statistics in a minimum-distance parameter estimation method. We conclude in Chapter 6. Some mathematical derivations, extra simulation results and code are deferred to the Appendix.

All computations and simulations are implemented using the statistical software R [29].

Chapter 2

Literature Study

In this chapter, we survey relevant literature. First off, we study the empirical identity process, which is the foundation for our statistical inference methods introduced later. Next, we study existing goodness-of-fit procedures, which can pose good alternatives for our own goodness-of-fit tests. We then review parameter estimator properties. In particular, we review some multidimensional parameter estimator performance measures - which prove useful when comparing estimators under models with many parameters, such as the normal mixture model. Then, we study existing parameter estimation methods, which might be used for comparison for our own methods. Special attention is given to the estimation methods in the setting of normal mixtures, as we apply our own methods on these distributions. Finally, we study the Expectation-Maximization algorithm, which can be used to apply maximum likelihood estimation effectively in the setting of normal mixtures.

2.1 The Empirical Identity Process

The empirical identity process is defined in [4] resulting from the composition of the empirical quantile function and the empirical distribution function.

It is constructed as follows. Let $U_1, \ldots, U_n$ be ordered uniform random variables, and define $U_0 = 0$ and $U_{n+1} = 1$. Denote $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(U_i \leq t)$, the empirical cumulative distribution function (ECDF). Denote $Q_n(u) = \inf\{t \in [0,1] : F_n(t) \geq u\}$ the empirical quantile function. Note that Q_n is the left-continuous generalized inverse of F_n.

The empirical identity function R_n is then defined as the mean of the lower

and upper empirical identity functions, given by:

$$R_n^L(t) := \begin{cases} 0 & \text{if } 0 \leq t < U_1 \\ Q_n(F_n(t)) & \text{if } U_1 \leq t \leq 1 \end{cases},$$
$$R_n^U(t) := \begin{cases} Q_n(F_n(t) + \frac{1}{n}) & \text{if } 0 \leq t < U_n \\ 1 & \text{if } U_n \leq t \leq 1 \end{cases},$$
$$R_n(t) := \frac{R_n^L(t) + R_n^U(t)}{2}.$$

The lower and upper empirical identity process are then obtained by subtracting the identity from the empirical identity functions. The empirical identity process $Y_n(t)$ for $t \in [0, 1]$ is then defined as the mean of the lower and upper identity process, given by:

$$\begin{aligned} Y_n^L(t) &:= (n+1)(R_n^L(t) - t), \\ Y_n^U(t) &:= (n+1)(R_n^U(t) - t), \\ Y_n(t) &:= (n+1)(R_n(t) - t) = \tfrac{1}{2}(Y_n^L(t) + Y_n^U(t)). \end{aligned}$$

As proven in [4], this empirical identity process converges to a highly irregular process, referred to as a white noise.

To find a more regular process, the integral of the process is studied. Noting that the lower and upper empirical identity functions are asymptotically equivalent, the integrated empirical identity process is only studied using the lower empirical identity process. We show in Appendix A.1 that for certain applications, the choice of lower or upper identity process is irrelevant. The Integrated Empirical Identity Process (IEIP) is now defined by:

$$I_n(t) := -\int_0^t Y_n^L(u)du = (n+1)\int_0^t \big(u - R_n^L(u)\big)du, \quad t \in [0, 1].$$

Note that between consecutive datapoints (U_{i-1}, U_i), the function $R_n^L(u)$ is constant and equal to U_{i-1}. Therefore, I_n can be expressed as:

$$\begin{aligned} I_n(t) &= (n+1)\int_0^t \big(u - R_n^L(u)\big)du \\ &= (n+1)\sum_{i=1}^{nF_n(t)} \int_{U_{i-1}}^{U_i} (u - U_{i-1})du + (n+1)\int_{nF_n(t)}^{t} (u - U_{nF_n(t)})du \\ &= \frac{n+1}{2}\sum_{i=1}^{nF_n(t)} (u_i - u_{i-1})^2 + \frac{n+1}{2}(t - R_n^L(t))^2. \end{aligned}$$

In Appendix A.1 we also derive simple expressions for the upper and central IEIP.

The following weak convergence is then proven:

$$2\sqrt{n+1}(I_n(t) - F_n(t)) \Longrightarrow V(t), \tag{2.1}$$

where $V(t)$ is a process given by:

$$V(t) = 2\sqrt{5}W_t - 2(\sqrt{5}-1)tW_1, \quad 0 \leq t \leq 1.$$

Here, W_t denotes standard Brownian motion. Note that we can interpret $V(t)$ as a scaled Brownian bridge, pinned at 0 at $t = 0$ and pinned at a random endpoint $\frac{W_1}{\sqrt{5}}$ at $t = 1$. Now, using the continuous mapping theorem for the weak convergence (2.1) yields the d_n statistic:

$$d_n := 2 \sup_{t\in[0,1]} \sqrt{n+1}|I_n(t) - F_n(t)| \Longrightarrow \sup_{t\in[0,1]} |V(t)|.$$

The authors demonstrate the use of this statistic in a goodness-of-fit procedure.

2.2 Goodness-of-fit Procedures

Goodness-of-fit testing is a useful tool in statistical modeling to check the compatibility of a statistical model with the data observed. Given data $X_1, \ldots, X_n$, the goodness-of-fit test decides between the following two competing hypotheses concerning the model F for our data:

$$H_0 : F \in \mathcal{F} \qquad \text{vs.} \qquad H_1 : F \notin \mathcal{F},$$

where $\mathcal{F}$ is some class of models. In the case of $\mathcal{F} = F_0$, the goodness-of-fit test becomes *simple*, otherwise it is known as *composite*. The latter can be more useful, but tests for composite nulls are generally harder to calibrate.

If we have a simple null hypothesis, then under the null, the probability integral transform $F_0(X)$ is uniformly distributed. If we know the distribution of a test statistic $T(U)$ when U is uniform, we can reject the null hypothesis if the value of $T(F_0(X))$ is significant. This approach has the attractive property of being distribution-free under the null hypothesis; the distribution of the test statistic $T(F_0(X))$ is independent of F_0 if the null hypothesis is true. For the goodness-of-fit statistics, we are therefore mainly

interested in their distribution when the data is uniform.

In Chapter 4, we study new goodness-of-fit procedures. We therefore now examine existing procedures which can be used for comparison. We focus on goodness-of-fit testing in case of a simple null hypothesis and continuous data.

2.2.1 ECDF Statistics

Goodness-of-fit tests based on the ECDF are some of the most well-known tests. We first discuss related convergence theorems, and then study the statistics that follow from them.

Convergence Results of the ECDF.

Let $X_1, \ldots, X_n$ be iid data from distribution F and let F_n be their ECDF. From the law of large numbers, we obtain the almost sure convergence:

$$F_n(t) \xrightarrow{\text{a.s.}} F(t)", \quad \forall_t. \tag{2.2}$$

Furthermore, using the central limit theorem, we obtain the weak convergence:

$$\sqrt{n}(F_n(t) - F(t)) \Longrightarrow \mathcal{N}(0, F(t)(1 - F(t))), \quad \forall_t. \tag{2.3}$$

Both of these results can be extended to more powerful and useful results. The pointwise convergence (2.2) can be extended to uniform convergence. This result is known as the Glivenko-Cantelli theorem:

$$\sup_t |F_n(t) - F(t)| \xrightarrow{\text{a.s.}} 0.$$

The weak convergence (2.3) can be extended to a functional version. This was done by Donsker [12][13], who obtained:

$$\sqrt{n}(F_n - F) \Longrightarrow G_n, \tag{2.4}$$

with G_n having zero mean and covariance function:

$$\text{Cov}(G_n(t_i), G_n(t_j)) = F(\min(t_i, t_j)) - F(t_i)F(t_j).$$

Note that G_n is a time-changed Brownian bridge[†].

[†]The standard Brownian bridge on $[0, 1]$ is given by $B_t = W_t - tW_t$, where W_t is standard Brownian motion, with mean 0 and variance 1. The time-changed Brownian bridge is obtained by substituting t with $F(t)$.

Statistics

The weak convergence (2.4) can be used with the continuous mapping theorem to obtain a variety of test statistics. The first statistic is known as the Kolmogorov-Smirnov (KS) statistic [21][40]†, and is given by:

$$D_n := \sup_t |F_n(t) - F(t)|.$$

Originally, the distribution of D_n was not found using the weak convergence (2.4), but derived directly by Kolmogorov, before the result by Donsker.

The second statistic was developed around the same time by Cramer [9] and Von Mises [41]. Referred to as the Cramèr-Von Mises (CvM) statistic, it is given by:

$$\omega^2 := \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 dF(x).$$

This statistic was improved by Anderson and Darling [1] by adding a weight function ϕ:

$$\omega_\phi^2 := \int_{-\infty}^{\infty} \phi(x)(F_n(x) - F(x))^2 dF(x).$$

In particular, the weight function $\phi(x) = (F(x)(1 - F(x))^{-1}$ was studied, resulting in the Anderson-Darling (AD) statistic:

$$A_n := \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1 - F(x)} dF(x).$$

Compared to the Cramèr-Von Mises statistic, the Anderson-Darling statistic places more weight on discrepancies at the tails of the distribution.

The performance of these three statistics when data is normal has been studied in [34]. The study claims that, out of these three statistics, the Kolmogorov-Smirnov statistic performs the poorest, while the Anderson-Darling statistic performs the best.

2.2.2 Spacing Statistics

There also exist goodness-of-fit tests based on spacings of uniform variables. Let $U_1, \ldots, U_n$ be ordered uniform random variables, and define $U_0 = 0$ and $U_{n+1} = 1$. The uniform spacings are defined as:

$$D_i := U_i - U_{i-1}, \quad i = 1, \ldots, n + 1. \tag{2.5}$$

†The original paper is in Italian. The second citation discusses the original article in English.

Statistics based on these spacings measure how "uniform" the spacings are. Note that the sum of spacings is, by definition, equal to 1, and the expected value is $\frac{1}{n+1}$ for each spacing. A statistic on the spacings can therefore measure the dispersion of the spacings from their expected value. This interpretation leads to many different statistics, as described in [18].

An early statistic based on the spacings is known as the Greenwood [17] statistic:

$$G_n := \sum_{i=1}^{n+1} D_i^2, \tag{2.6}$$

for which Moran [23]† developed a convergence law:

$$\sqrt{n+1}\left(\frac{n+1}{2}\sum_{i=1}^{n+1} D_i^2 - 1\right) \Longrightarrow \mathcal{N}(0,1).$$

Naturally, direct applications of the central limit theorem are not applicable, as the variables D_i are not independent.

The statistic (2.6) was extended to higher powers of spacings and studied by Kimball [20]. In [18], the dispersion of the spacings is measured using the Gini-index.

We can also consider spacings of higher order. In [18], the overlapping spacings of order m are defined as:

$$D_i^{(m)} = \begin{cases} U_{i+m} - U_i, & i = 0, \ldots, n+1-m, \\ 1 + U_{i+m-n-1} - U_i, & i = n+2-m, \ldots, n, \end{cases}$$

and the non-overlapping spacings of order m are defined as:

$$D_i^{(m')} = U_{(i+1)m} - U_{(i)m}, \quad i = 0, \ldots, \left\lfloor \frac{n}{m} \right\rfloor,$$

These higher-order spacings can be used to generalize the Greenwood statistic (2.6), as has been done in [33]. Results suggest that the higher-order spacings improve upon the one-step spacings, having higher efficiency.

†In the cited paper, the uniform data has length $n-1$, so the resulting number of spacings is equal to n. To keep notation consistent with the rest of our text, we use data of length n, resulting in $n+1$ spacings. The convergence result is given on page 97 in the cited paper.

Instead of the square as in (2.6), sums of other functions of the spacings have been studied as well, such as the logarithm or the reciprocal. Additionally, statistics based on the rank of the spacings have been proposed, such as the minimum or maximum spacing. A comprehensive overview of these is given in [28].

A general challenge in goodness-of-fit testing is the specification of the parameters of the null distribution. These must be specified a priori for the goodness-of-fit test statistic distributions to hold. Estimating the parameters for the null distribution from the data generally changes the distribution of the test statistic. However, [8] shows that the asymptotic distribution of the Greenwood statistic (2.6) does *not* change when parameters are estimated, provided that the estimation is done efficiently.

A notable disadvantage of the spacings statistics is the sensitivity to ties in the data. Many procedures based on spacings fail when ties are present. Theoretically, there is no problem, as the statistics are developed for continuous data. In practice however, rounded data can cause problems.

2.2.3 Shapiro-Wilk Statistic

The Shapiro-Wilk statistic [39] can be used to test data for the composite hypothesis of normality. Let $X = X_1, \ldots, X_n$ denote the ordered data. Then, the Shapiro-Wilk (SW) statistic is given by:

$$W_n := \frac{(\sum_{i=1}^{n} a_i X_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Here, $\bar{X}$ denotes the sample mean, and a_i are constants given by:

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}},$$

where m is a vector of length n containing the expected values of the standard normal order statistics, and V the sample covariance matrix of X.

Clearly, the computation of W_n for large samples can be demanding for large samples, as the computation for the coefficients a_i involves large matrix operations and inversions. However, research has been done to approximate the coefficients a_i for large sample sizes. The approximation given in [37] is valid for samples up to $n = 2000$, and the author notes that extrapolation apprears justified. Moreover, [30] gives an approximation which can be extended to

all sample sizes, and provides critical values for sizes up to $n = 5000$.

A power study [34] claims that the Shapiro-Wilk statistic is the most powerful test for normality.

2.3 Estimator Properties

In this section, we discuss several estimator properties.

2.3.1 Behavioral Properties

Consistency

An estimator $\hat{\theta}_n$ for parameter θ and sample size n is (weakly) consistent if:

$$\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta,$$

as $n \to \infty$. The estimator is strongly consistent if it converges almost surely.

Asymptotic Normality

An estimator $\hat{\theta}_n$ for parameter θ and sample size n is asymptotically normal if:

$$\sqrt{n}(\hat{\theta}_n - \theta) \Longrightarrow \mathcal{N}(0, \Sigma),$$

as $n \to \infty$, with Σ some covariance matrix.

Efficiency

An unbiased estimator $\hat{\theta}_n$ for parameter θ and sample size n is efficient if:

$$\mathrm{Var}(\hat{\theta}_n) = \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information. Thus, the estimator $\hat{\theta}_n$ is efficient if it achieves the Cramér-Rao lower bound. The estimator $\hat{\theta}_n$ is asymptotically efficient if it achieves this bound asymptotically, i.e. for $n \to \infty$.

2.3.2 Performance Measures

Bias

The bias of an estimator $\hat{\theta}$ for parameter θ is given by:

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

Note that if $\hat{\theta}$ is a multivariate estimator, the bias is a vector.

Covariance

The covariance matrix of an estimator $\hat{\theta}$ for parameter θ is given by:

$$\mathrm{Cov}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\hat{\theta} - \mathbb{E}[\hat{\theta}])^T].$$

Note that, if $\hat{\theta}$ is univariate estimator, this is simply the variance.

For comparison among estimators the idea of concentration ellipsoids can be useful. The volume of the ellipsoids is a univariate measure of the dispersion of the estimator which can easily be compared among different estimators. The idea was first proposed by Cramér in [10], Chapter 21.10. The concentration ellipsoid is given by:

$$\mathcal{E}_c = \{t \in \mathbb{R}^n : (t - \mathbb{E}[\hat{\theta}])^T \Sigma^{-1} (t - \mathbb{E}[\hat{\theta}]) \leq c\}, \tag{2.7}$$

where Σ is the covariance matrix of $\hat{\theta}$ and c is a constant. This constant can be chosen such that we obtain a confidence ellipse, which generalizes confidence intervals for higher dimensions. Other definitions are discussed in [24], for instance in cases where Σ is singular. For our purposes, we assume Σ to be non-singular.

When the estimator is two-dimensional, it can be instructive to plot this ellipse. Naturally, this is not possible for higher dimensions. However, the volume of the ellipsoid is a measure of its dispersion across each parameter dimension.

The axis lengths of the ellipsoid along its major axes are proportional to the eigenvalues λ_i of the matrix Σ. Therefore, the volume of the ellipsoid (2.7) with $c = 1$ is given by:

$$\mathrm{Vol}(\mathcal{E}_1) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2} + 1)} \prod_{i=1}^{n} \sqrt{\lambda_i}.$$

A careful justification is given in Appendix A.2. This volume can be used to compare the variance of multidimensional estimators across all parameters simultaneously. For simulation results, the empirical covariance matrix may be used.

Mean-Squared Error

The mean-square-error (MSE) of a univariate estimator $\hat{\theta}$ for parameter θ is given by:

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

The MSE is commonly generalized for multivariate estimators in two ways; using the inner- or outer product of the vector $(\hat{\theta} - \theta)$. The generalization using the inner product is described in [15], Chapter 4.1. It is given by:

$$\mathrm{MSE}_{\mathrm{inner}}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^T(\hat{\theta} - \theta)] = \sum_{i=i}^{n} \mathbb{E}[(\hat{\theta}_j - \theta_j)^2] = \sum_{i=i}^{n} \mathrm{MSE}(\hat{\theta}_j).$$

As can be seen, this is simply the sum of the MSE of each individual parameter. The author notes this definition can be questionable when used on a mixture of location and scale parameters.

The generalization using the outer product is described in [19], Section 3.1.3b, is given by:

$$\mathrm{MSE}_{\mathrm{outer}}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = \mathrm{Cov}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})\mathrm{Bias}(\hat{\theta})^T.$$

Similarly as with the covariance matrix, we can define concentration ellipsoids using the MSE matrix. The volumes can be compared among different estimators, which takes both dispersion and bias into account. We use this second generalization in combination with the concentration ellipsoids when comparing estimators. The MSE concentration ellipsoid is given by:

$$\mathcal{E}_c^{\mathrm{MSE}} = \{t \in \mathbb{R}^n : (t - \theta)^T \left[\mathrm{MSE}_{\mathrm{outer}}(\hat{\theta})\right]^{-1}(t - \theta) \leq c\}.$$

The volume of this ellipsoid can be computed analogously to the volume of the regular concentration ellipsoid.

2.4 Parameter Estimation

In this section, we cover existing parameter point estimation procedures. Since we propose a parameter estimation procedure later ourselves, we aim to identify the estimators which might pose as good alternatives to our own to assess performance.

As we use our estimators in normal mixture models, we pay special attention to the estimators in this context.

2.4.1 Moment-based Estimation

We first examine the method of moments. The method relies on computing functions g_a corresponding to the true population moments $\mathbb{E}[X^a]$ as a function of the parameters $\theta = (\theta_1, \ldots, \theta_k)$. Using the empirical moments $\mu_a = \sum_{i=1}^{n} X_i^a$, the moment estimator $\hat{\theta}_{\text{MOM}}$ is then given as the solution to the system of equations:

$$\begin{cases} \mu_1 = g_1(\theta), \\ \mu_2 = g_2(\theta), \\ \quad \ldots \\ \mu_k = g_k(\theta). \end{cases}$$

These estimators are consistent under weak conditions. However, the estimator is not necessarily the most efficient. Other estimators, such as the maximum-likelihood estimator, can have lower variance.

Computationally, the method of moments can yield both very simple and very complex calculations, depending on the model assumed. When the assumed model is a normal mixture, the latter is definitely the case. For instance, in [26], after many algebraic manipulations, the result must be obtained from the root of a ninth-degree polynomial. While this paper was published as early as 1894, therefore done without the use of computers, models with more components result in even more complex computations.

2.4.2 Likelihood-based Estimation

Perhaps the most well-known estimation procedure is the maximum-likelihood estimator (MLE). The method has attractive theoretical properties. Under suitable regularity conditions, the MLE is consistent, asymptotically efficient and asymptotically normal. The MLE is given by:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta, x),$$
$$L(\theta, x) = f(\theta, x),$$

where L is known as the likelihood function, and f is the joint probability density of the data $X = (X_1, \ldots, X_n)$. For computational reasons, usually the logarithm of the likelihood function is used. In the case of iid data, the likelihood function becomes much simpler, as the joint density can then be factored as the product of the marginal densities. Finding the parameter estimates is usually done by solving the likelihood equations, which are obtained by setting the derivative of the (log) likelihood to 0. For many models,

these equations can be solved analytically and parameter estimates can then be obtained easily.

This is *not* the case for mixture models, where the likelihood equations are usually nonlinear and no analytic solution exists. This causes likelihood optimization to be computationally expensive, which is problematic when also considering the high dimensionality of the parameter space. Furthermore, the likelihood estimation cannot be applied naively. If the variances of each component are unconstrained, then the likelihood can increase to infinity by setting the mean of one component equal to a datapoint and letting its corresponding variance go to zero.

However, under regularity conditions, there still exists a local maximum of the likelihood that has the MLE properties of consistency, asymptotic efficiency and asymptotic normality [27][35].

This maximum is not necessarily unique; by the so called label-switching, the likelihood of a set of parameters does not change when we permute the labels on the component means, variances and weights. Finally, other lower local maxima might exist, which could make optimization harder.

Because of these problems, direct application of maximum likelihood is ineffective. However, a solution to these problems has been found in the form of the Expectation-Maximization algorithm, which we discuss in Section 2.5.

2.4.3 Spacings-based Estimation

The Maximum Spacings (MSP) estimator was proposed by [31] and [7] independently, as an alternative to the popular MLE. Mainly, the method was proposed as an alternative in cases where MLE performs poorly, with mixture models being one of these cases.

The estimator relies on the probability integral transform. Consider iid data $X = (X_1, \ldots, X_n)$ and assume the parametric model F_θ, with $\theta \in \Theta$. Then, $F_\theta(X)$ is uniform. Therefore, we can estimate θ by finding the value that results in $F_\theta(X)$ being the most "uniform". This is done by maximizing the geometric means of the spacings of U. The MSP is therefore given by:

$$\hat{\theta}_{\text{MSP}} = \arg\max_{\theta \in \Theta} \left(\prod_{i=1}^{n+1} D_i \right)^{\frac{1}{n+1}},$$

where D_i are the spacings, as defined before in (2.5). To avoid numerical problems, the logarithm of the geometric mean may be maximized instead. Note that the maximum of the geometric mean always exists, since the sum of the spacings is always equal to 1.

In cases where the MLE exists, and is therefore consistent and asymptotically efficient, the MSP has the same attractive properties, as demonstrated in the original papers. Strong consistency properties were later devised in [38]. Additionally, the MSP can have these properties in cases where the MLE does not exist.

Moreover, the MSP can have better finite-sample performance. According to [14], when densities are skewed or heavy-tailed, the MSP can be expected to perform better in finite samples.

Another setting where the MSP performs better than the MSP is in the uniform case with unknown endpoints $[a, b]$. Let the sample X be ordered. Then, the MLE estimates are given by $\hat{a}_{\text{MLE}} = X_1$ and $\hat{b}_{\text{MLE}} = X_n$. The MSP estimates are then given by:

$$\hat{a}_{\text{MSP}} = \frac{nX_1 - X_n}{n-1}, \qquad \hat{b}_{\text{MSP}} = \frac{nX_n - X_1}{n-1}. \tag{2.8}$$

The MSP estimates are the minimum-variance-unbiased estimators for these parameters.

One notable disadvantage of the MSP is that the method is not naturally generalized to multivariate observations. However, research has been done in that area; for instance, in [32], where the distances are generalized to either a geometric or a probabilistic view.

Another disadvantage of the MSP is the sensitivity to ties. Just like in the goodness-of-fit setting, the spacings statistics may break down in the case of ties in the data.

The MSP uses the geometric mean as a measure to of uniformity. However, other functions could be used as well, leading to the generalized spacings estimator, introduced in [16].

2.4.4 Minimum-distance-based Estimation

The minimum-distance estimator (MDE) was proposed by Wolfowitz in [42][†]. This MDE is based on the distance between the ECDF F_n and the true cumulative distribution function (CDF) F_θ, and is given by:

$$\hat{\theta}_{\text{MDE-Dist}} = \arg \inf_{\theta \in \Theta} \text{Dist}(F_n, F_\theta),$$

where Dist is some appropriate distance measure between the ECDF and the CDF, such as the Kolmogorov-Smirnov, Cramèr-Von Mises or Anderson-Darling statistics introduced in Section 2.2.1.

The MDE based on the ECDF and CDF has been studied especially in the case of robust estimation. In [25], it is shown that this MDE leads to excellent robustness properties.

Note however, that a minimum-distance estimator is not necessarily based on the ECDF and the CDF. We propose a MDE based on the discrepancy between the IEIP (introduced in Section 2.1) and the ECDF of the probability integral transform of the sample in Chapter 5. In this setting, both functions depend on the parameters θ; in the ECDF based MDE, only F_θ depends on the parameters.

2.5 EM Algorithm

As discussed in Section 2.4.2, maximum likelihood is problematic in the setting of normal mixtures. However, research has been done to use maximum likelihood effectively despite the problems. A very common solution is given by the Expectation-Maximization (EM) algorithm, proposed in [11][‡]. A survey of the early work on maximum likelihood and mixture densities, with emphasis on the EM algorithm and its properties, is given in [35]. The book [22] on finite mixture models also reviews more contemporary literature.

2.5.1 Description of the Algorithm

The idea behind the EM algorithm is to regard the given data as *incomplete*, where the missing information is the label of the component that "generated"

[†]Work towards the MDE was done before this paper. The cited source gives an updated version of the MDE which is superior to the methods proposed before by the same author.

[‡]The EM algorithm is actually useful in a much more general incomplete data setting. The (normal) mixture setting is just one of the cases where the algorithm can be used. In the cited paper, the authors give many examples of its use.

the datapoint. Now, observe the following:

- If we would have the parameters of the normal mixture, we can easily compute the probability of a datapoint originating from a given component;
- Conversely, if we would have the component probabilities of each datapoint, estimating the mixture density parameters is straightforward. One can simply estimate the parameters of each component separately by weighing the data with the respective probability of the component.

These two steps, the former being referred to as the "Expectation" step and the latter as the "Maximization" step, can be done iteratively given an initial guess of either the component probabilities or the parameters of the mixture.

2.5.2 Initialization Methods

The EM algorithm needs to be started with initial parameters. Many different initialization procedures have been devised. A discussion on several methods is given in [5]. We discuss procedures that generate the initial values randomly. In practice, it is often best to use multiple random initializations, and then select the estimate with the highest likelihood. A common idea, known as emEM, is to improve the initial random points by running the EM algorithm with a lax convergence criterion first before running the algorithm fully. We now discuss three random initializations. The final initialization, K-means++ initialization, was found in [5] to be the best in noise-free data.

Simple random starting values. Many initialization methods generate random initial values from various distributions. In [22], a simple random method is given to initialize the parameters of the mixture. The initial means of the mixture density are randomly generated from a normal distribution, with mean and variance equal to the sample mean and sample variance. The initial variances of the mixture density are set to the sample variance, and the initial weights of the mixture are set to $\frac{1}{k}$, where k is the number of components of the mixture. Thus, only the means are randomly generated.

Random starting values with data binning. The R package `mixtools` [3] implements a so-called binning method[†]. When we assume k mixture

[†]The package documentation does not elaborate on the method and gives only a brief description. However, we can inspect the code and reverse-engineer the method regardless.

components, the sorted data of size n is first binned in k bins of equal size. Define B_j to be index set of the jth bin, with indices:

$$\begin{aligned} B_1 &= (1, \ldots, \left\lceil \frac{n}{k} \right\rceil), \\ &\ldots \\ B_j &= ((j-1) \left\lfloor \frac{n}{k} \right\rfloor, \ldots, j \left\lceil \frac{n}{k} \right\rceil), \\ &\ldots \\ B_k &= ((k-1) \left\lfloor \frac{n}{k} \right\rfloor, \ldots, k \left\lceil \frac{n}{k} \right\rceil). \end{aligned}$$

Now, let $\bar{X}_j$ and s_j be the sample mean and standard deviation of the jth bin, respectively. Then, the initial mixture parameters are generated from the following distributions:

$$\begin{aligned} \sigma_j^{-1} &\sim \text{Exp}(s_j) \\ \mu_j &\sim \mathcal{N}(\bar{X}_j, \sigma_j) \end{aligned}$$

The weights are generated from a uniform distribution between 0 and 1, and then normalized by their sum[†]. Note that the variance is generated as the reciprocal of the exponential distribution.

K-means++ initialization. The final method we discuss here is known as K-means++ initialization. This method was originally proposed in [2] as an initialization method for the well-known K-means algorithm. In [5], this method is adapted to be used for the EM-algorithm. First, initial "centers" are generated using the K-means++ algorithm. These centers are then used to compute initial parameters for the normal mixture.

Given data $X = (X_1, \ldots, X_n)$ and a model with k components, the method first chooses centers sequentially as follows. The first center is chosen uniformly at random from X. Denote $D(X_i)$ the distance from data point X_i to the closest center already defined. Then, the other centers $c_2, \ldots, c_k$ are chosen at random from X, with probabilities equal to $\frac{D(X_i)}{\sum_{i=1}^{n} D(X_i)}$ for each X_i.

[†]The package documentation claims this is the uniform Dirichlet distribution, but this is not true. For $k = 2$, for example, we can show that the weight parameter ρ is generated from the density:

$$f_\rho(x) = \begin{cases} \frac{1}{2(1-x)^2} & 0 \leq x \leq \frac{1}{2}, \\ \frac{1}{2x^2} & \frac{1}{2} < x \leq 1. \end{cases}$$

These centers are then transformed to initial parameters for the normal mixture as follows. The data X is divided into k partitions by assigning each X_i to the closest center c_j. The initial means and variances of the mixture are then set to the sample means and variance of each partition. The initial weights are set to the fraction of datapoints assigned to each cluster.

2.5.3 Properties of the Algorithm

The EM-algorithm has the a very desirable property that the likelihood cannot decrease in successive iterations, as shown in [11]. Furthermore, the attractive properties of the MLE hold under suitable regularity conditions, as mentioned before in Section 2.4.2. Finally, the computational cost of each iteration is relatively low [35].

However, the EM-algorithm has some disadvantages as well. The convergence is comparably slow [35]. The convergence of the EM algorithm can be expected to be linear in mixtures of exponential families. In contrast, Newtons methods converges quadratically and quasi-Newton methods converge superlinearly, for example. The convergence is slower when the components are not well-separated.

Additionally, the algorithm is sensitive to the starting parameters [22]. When the initial parameters are chosen poorly, the algorithm converges to a suboptimal maximum.

Chapter 3

The o_n Statistic

We now define an alternative statistic based on the integrated empirical identity process, which was introduced in Section 2.1.

As described in [4], we have the following weak convergence:

$$2\sqrt{n+1}(I_n(t) - F_n(t)) \implies V(t). \tag{3.1}$$

The continuous mapping theorem tells us that the weak convergence still holds if a continuous mapping is applied to both sides. We might be able to improve on the d_n statistic, which uses the supremum of the absolute value as a mapping, by using another continuous mapping. While the resulting test statistic might have a distribution which is hard to describe analytically, numerical results could still be obtained and used to assess performance.

The Cramèr-Von Mises statistic improved upon the Kolmogorov-Smirnov statistic by using a squared-integral mapping on the weak convergence result (2.4). In a similar fashion, consider the following mapping:

$$m\big(x(t)\big) = \int_0^1 x(t)^2 dt.$$

We then use this mapping m on both sides of the weak convergence result (3.1) to define the o_n statistic:

$$o_n := 4(n+1)\int_0^1 (I_n(t) - F_n(t))^2 dt \implies \int_0^1 V(t)^2 dt.$$

One might improve on the o_n statistic by adding a weight function in the same spirit as the Anderson-Darling statistic. It might be tempting to add weight functions such as:

$$w_1(x) = \frac{1}{I_n(t)(1 - I_n(t))}, \quad w_2(x) = \frac{1}{F_n(t)(1 - F_n(t))}, \quad w_3(x) = \frac{1}{t(1-t)},$$

but care must be taken that the resulting integral is still well-defined.

3.1 Dependency on Spacings

The o_n statistic consists of an integral of a function containing discontinuities. However, the statistic can be simplified and expressed using only the spacings, without the need for integration. This formulation also avoids practical problems when evaluating the integral numerically.

Note that both I_n and F_n have simple expressions between two consecutive datapoints. Let $u_1, \ldots, u_n$ be the uniform order statistics. Define $u_0 = 0$ and $u_{n+1} = 1$. We now have for $t \in [u_{i-1}, u_i)$:

$$\begin{aligned} I_n(t) &= \frac{n+1}{2} \sum_{j=1}^{i-1} (u_j - u_{j-1})^2 + \frac{n+1}{2}(t - u_{i-1})^2, \\ F_n(t) &= \frac{i-1}{n}. \end{aligned}$$

Note that, in the definition of $I_n(t)$, for $i = 1$, we define the sum $j = 1$ to 0 to be the empty sum, equal to 0. For notational purposes, we introduce:

$$\begin{aligned} t_i &= u_i - u_{i-1}, \\ S_i &= \frac{n+1}{2} \sum_{j=1}^{i} t_j^2, \\ P_i &= S_i - \frac{i}{n}. \end{aligned}$$

So t_i is a spacing, S_i is the sum of squared spacings up to and including u_i, multiplied with a constant $\frac{n+1}{2}$. Then P_i is S_i minus the value of F_n between the next spacing not included in S_i.

We can now simplify o_n in the following manner:

$$
\begin{aligned}
&4(n+1)\int_0^1 (I_n(t)-F_n(t))^2dt = 4(n+1)\sum_{i=1}^{n+1}\int_{u_{i-1}}^{u_i}(I_n(t)-F_n(t))^2dt \\
&= 4(n+1)\sum_{i=1}^{n+1}\int_{u_{i-1}}^{u_i}\big(\frac{n+1}{2}\sum_{j=1}^{i-1}(u_j-u_{j-1})^2+\frac{n+1}{2}(t-u_{i-1})^2-\frac{i-1}{n}\big)^2dt \\
&= 4(n+1)\sum_{i=1}^{n+1}\int_{u_{i-1}}^{u_i}\big(P_{i-1}+\frac{n+1}{2}(t-u_{i-1})^2\big)^2dt \\
&= 4(n+1)\sum_{i=1}^{n+1}\int_{u_{i-1}}^{u_i}(P_{i-1}^2+(n+1)P_{i-1}(t-u_{i-1})^2+\frac{(n+1)^2}{4}(t-u_{i-1})^4)dt \\
&= 4(n+1)\sum_{i=1}^{n+1}P_{i-1}^2t_i+\frac{n+1}{3}t_i^3P_{i-1}+\frac{(n+1)^2}{20}t_i^5 \\
&= 4(n+1)\sum_{i=1}^{n+1}(S_{i-1}-\frac{i-1}{n})^2t_i+\frac{n+1}{3}t_i^3(S_{i-1}-\frac{i-1}{n})+\frac{(n+1)^2}{20}t_i^5 \\
&= 4(n+1)\sum_{i=1}^{n+1}\Bigg(S_{i-1}^2t_i-\frac{2(i-1)}{n}S_{i-1}t_i+\frac{(i-1)^2}{n^2}t_i+\frac{n+1}{3}S_{i-1}t_i^3 \\
&\qquad\qquad -\frac{n+1}{n}\frac{i-1}{3}t_i^3+\frac{(n+1)^2}{20}t_i^5\Bigg).
\end{aligned}
$$

Therefore, o_n can be expressed based solely on the spacings; it is therefore another spacing statistic, which were discussed in Section 2.2.2.

In the last line, the terms are given in decreasing order:

$$
\begin{aligned}
\sum_{i=1}^{n+1}S_{i-1}^2t_i-\frac{2(i-1)}{n}S_{i-1}t_i+\frac{(i-1)^2}{n^2}t_i &\sim O(1), \\
\sum_{i=1}^{n+1}\frac{n+1}{3}S_{i-1}t_i^3-\frac{n+1}{n}\frac{i-1}{3}t_i^3 &\sim O(\frac{1}{n}), \\
\sum_{i=1}^{n+1}\frac{(n+1)^2}{20}t_i^5 &\sim O(\frac{1}{n^2}).
\end{aligned}
$$

A detailed derivation of these orders can be found in Appendix A.3.

3.2 Asymptotic Distribution

It is desirable to have (asymptotic) analytical distribution results for the o_n statistic. We can obtain results in two ways:

1. We have the weak convergence result (3.1). We can analyze the distribution of the mapped version of $V(t)$ to obtain the asymptotic distribution of o_n.

2. We can express o_n with only spacings. The distribution of a spacing t_i is known, and given by:

$$t_i \sim \frac{E_i}{\sum_j^{n+1} E_j},$$

where E_i are iid exponential random variables with rate parameter equal to 1. We can use this to analyze the distribution of o_n exactly.

We study the distribution using the first method.

Asymptotic Results using Weak Convergence

We must analyze the distribution of:

$$m\big(V(t)\big) = \int_0^1 V(t)^2 dt.$$

Note that $V(t)$ is given by:

$$V(t) = 2\sqrt{5}W_t - s(\sqrt{5}-1)tW_1, \quad 0 \le t \le 1,$$

where W_t is a standard one-dimensional Brownian motion. We define:

$$B_t = W_t - tW_1, \quad 0 \le t \le 1,$$

a Brownian bridge on $[0,1]$, pinned at both ends to 0. Now, note that we can write $B'_t = B_t + tz$, which is a Brownian bridge pinned to z at $t = 1$.

Note that $V(t)$ can be interpreted as a scaled Brownian bridge with a random endpoint $\frac{W_1}{\sqrt{5}}$, and we can therefore also write:

$$V(t) = 2\sqrt{5}\left(B_t + t\frac{W_t}{\sqrt{5}}\right), \quad 0 \le t \le 1.$$

To compute the distribution of $m\big(V(t)\big)$, we can use a formula for the joint density of the square integral of a Brownian motion and its endpoint, given in [6], Equation 1.9.8, page 169:

$$\mathcal{P}_x\bigg(\int_0^t W_s^2 ds \in dy, W_t \in dz\bigg) = \frac{1}{\sqrt{2\pi}}\mathrm{ee}_y(\frac{1}{2}, t, \frac{x^2+z^2}{2}, -xz)dydz,$$

where W_s is standard Brownian motion, $\mathcal{P}_x$ is the probability measure when the process W_s started at x, and ee_y is a special inverse Laplace transform.

We can use this to compute the probability measure of the integrated squared Brownian bridge $B_s + sz$. Note that:

$$\mathcal{P}_x\bigg(\int_0^t W_s^2 ds \in dy \;\bigg|\; W_t \in dz\bigg) = \\ \mathcal{P}_x\bigg(\int_0^t W_s^2 ds \in dy, W_t \in dz\bigg)\mathcal{P}_x\bigg(W_t \in dz\bigg)^{-1}.$$

So, for the Brownian bridge $B_s + sz$, it holds:

$$\mathcal{P}_x\bigg(\int_0^t (B_s+sz)^2 ds \in dy\bigg) = \mathcal{P}_x\bigg(\int_0^t W_s^2 ds \in dy \;\bigg|\; W_t = z\bigg).$$

Therefore, the probability density function of $m(B_s + sz)$ is now given by:

$$f_{m(B_s+sz)}(y) = f_{m(W_s),W_1}(y,z)f_{W_1}^{-1}(z),$$

where $f_{m(W_s),W_1}(y,z)$ is the joint probability density of $m(W_s)$ and W_1, and $f_{W_t}(z)$ is the marginal probability density of W_t. We can now compute the lower quantile of the distribution of $m(B_s + sz)$ as:

$$\begin{aligned}\mathbb{P}\bigg(\int_0^1 (B_s+sz)^2 < \alpha\bigg) &= \int_{-\infty}^{\alpha} f_{m(B_s+sz)}(y)dy \\ &= \int_{-\infty}^{\alpha} f_{m(W_s),W_1}(y,z)f_{W_1}^{-1}(z)dy. \qquad (3.2)\end{aligned}$$

To analyze (3.2), we must first consider the function ee_y, which is given by:

$$\begin{aligned}\mathrm{ee}_y(\nu,t,z,x) &= \mathcal{L}_\gamma^{-1}\bigg(\bigg(\frac{\sqrt{2\gamma}}{\mathrm{sh}(t\sqrt{2\gamma})}\bigg)^\nu \exp\bigg(-\frac{z\sqrt{2\gamma}\mathrm{ch}(t\sqrt{2\gamma})}{\mathrm{sh}(t\sqrt{2\gamma})} - \frac{x\sqrt{2\gamma}}{\mathrm{sh}(t\sqrt{2\gamma})}\bigg)\bigg) \\ &= \sum_{k=0}^{\infty}\frac{(-z)^k}{k!}\sum_{l=0}^{\infty}\frac{(-x)^l}{l!}s_y(\nu+k+l, \nu+k+l, t, z+kt),\end{aligned}$$

where s_y is another special inverse Laplace transform. In our case, $\nu = \frac{1}{2}$, $x = 0$, $t = 1$, and then this expression simplifies to:

$$\mathrm{ee}_y(\frac{1}{2}, 1, z, 0) = \sum_{k=0}^{\infty} \frac{(-z)^k}{k!} s_y(\frac{1}{2} + k, \frac{1}{2} + k, 1, z + k).$$

Now, s_y is given by:

$$\begin{aligned} s_y(\mu, \nu, t, z) &= \mathcal{L}_\gamma^{-1}\left(\frac{(2\gamma)^{\frac{\mu}{2}}}{\mathrm{sh}^\nu(t\sqrt{2\gamma})} e^{-z\sqrt{2\gamma}}\right) \\ &= 2^\nu \sum_{l=0}^{\infty} \frac{\Gamma(\nu + l)\exp(-\frac{(\nu t + z + 2lt)^2}{4y})}{\sqrt{2\pi} y^{1+\frac{\mu}{2}} \Gamma(\nu) l!} D_{\mu+1}\left(\frac{\nu t + z + 2lt}{\sqrt{y}}\right), \end{aligned}$$

where D_ρ is the parabolic cylinder function. For the parameters considered, we obtain:

$$\begin{aligned} &s_y(\frac{1}{2} + k, \frac{1}{2} + k, 1, z + k) = \\ &\qquad 2^{\frac{1}{2}+k} \sum_{l=0}^{\infty} \frac{\Gamma(\frac{1}{2} + k + l)\exp(-\frac{(\frac{1}{2} + k + z + 2l)^2}{4y})}{\sqrt{2\pi} y^{\frac{5}{4}+\frac{k}{2}} \Gamma(\frac{1}{2} + k) l!} D_{\frac{3}{2}+k}\left(\frac{\frac{1}{2} + k + z + 2l}{\sqrt{y}}\right). \end{aligned}$$

To compute the quantiles of the distribution of $V(t)$, we can use the law of total probability. Define $w_1 = \frac{W_1}{\sqrt{5}}$. We then have:

$$\mathbb{P}\left(\int_0^1 V(t)^2 < \alpha\right) = \mathbb{E}\left[\mathbb{P}\left(\int_0^1 (B_t + t w_1)^2 < \frac{\alpha}{20}\right) \middle| W(1)\right].$$

We can now use this to obtain the quantiles of $m\big(V(t)\big)$, as:

$$\begin{aligned} &\mathbb{P}\left(\int_0^1 V(t)^2 < \alpha\right) = \mathbb{E}\left[\int_{-\infty}^{\frac{\alpha}{20}} f_{m(W_s), W_t}(y, w_1) f_{W_t}^{-1}(w_1) dy\right] \\ &= \mathbb{E}\left[\exp(\frac{w_1^2}{2}) \int_{-\infty}^{\frac{\alpha}{20}} \mathrm{ee}_y(\frac{1}{2}, 1, \frac{w_1^2}{2}, 0) dy\right] \\ &= \mathbb{E}\left[\exp(\frac{w_1^2}{2}) \int_{-\infty}^{\frac{\alpha}{20}} \sum_{k=0}^{\infty} \frac{(-\frac{w_1^2}{2})^k}{k!} s_y(\frac{1}{2} + k, \frac{1}{2} + k, 1, \frac{w_1^2}{2} + k) dy\right]. \end{aligned}$$

We can substitute the expression for s_y and subsequently D_ρ. However, this leads to a complex expression without hope for simplification. It seems that the analytic computation of the expectation becomes intractable, and with it, the quantile computation.

Chapter 4

Goodness-of-fit Testing

The statistics d_n and o_n can be used in a goodness-of-fit hypothesis testing procedure. Goodness-of-fit testing was introduced in Section 2.2. We study goodness-of-fit testing using the d_n and o_n statistic in the case of simple null hypotheses. We consider the regular case where the parameters of the null distribution are specified a priori, and the more practical case where these are estimated from the data.

As illustrated in Section 2.2, if our simple null hypothesis is that the data came from F_θ, we can compute the statistics for the transformed data $F_\theta(X)$. Then, we can reject the null if the value of the statistic is significant, provided we know the distribution of the d_n and o_n statistic for uniform data. In this setting, attractively, the statistics are distribution-free under the null hypothesis.

Specifying the null distribution F_0 is not always straightforward, however. The parameters θ of the distribution must be known in order to find the probability integral transform $F_\theta(X)$. Estimating these parameters from the data itself then seems attractive. However, the distribution of the test statistics can change when computed on data transformed with a distribution with estimated parameters. Moreover, the statistic can then also depend on the distribution F, resulting in the statistic no longer being distribution-free under the null. Therefore, to use the statistics in this setting, we need to find the distribution of the statistic conditional on the model F.

In the case where parameters are specified a priori, the goodness-of-fit hypothesis test can be formulated as follows. Given data X and hypothesized distribution F_θ, with θ specified a priori, the resulting competing hypotheses are:

$$H_0 : X \sim F_\theta \qquad \text{vs.} \qquad H_1 : X \not\sim F_\theta.$$

If the parameters are estimated from the data, our hypotheses change. Given data X and hypothesized family of distributions F_θ, with $\theta \in \Theta$, and $\hat{\theta}$ an estimate of the parameters of the distribution, the resulting competing hypotheses are:

$$H_0 : X \sim F_{\hat{\theta}} \qquad \text{vs.} \qquad H_1 : X \not\sim F_{\hat{\theta}}.$$

We study the statistics d_n and o_n in both settings. First, we compute quantiles of the statistics, both with fully specified null distributions and estimated parameters; the latter depends on the model F. Then, we compare their power with two other goodness-of-fit statistics; the Anderson-Darling and the Shapiro-Wilk statistic.

4.1 Quantiles

The quantiles of the statistics are approximated using a Monte Carlo approach. We compute quantiles for the statistics d_n and o_n. We also compute quantiles for the Anderson-Darling statistic.

The statistics are computed for N samples of size n and empirical quantiles are obtained. The distribution of the statistics might depend on the sample size n, so the quantiles are computed separately for different sample sizes.

For the fully-specified null distribution, we compute the values of the statistics on uniform samples. The distribution of the statistics might change and be dependent on the distribution when parameters are estimated. For the case of estimated parameters, samples from the relevant null distribution are drawn and the probability integral transform is computed using estimated parameters. For the normal and exponential distribution, the parameters are estimated using maximum likelihood. For the uniform distribution, the minimum-variance-unbiased estimator (2.8) is used.

The results are computed using $N = 10^6$ samples. The results for the statistics d_n, o_n and A_n are given in Table 4.1, Table 4.2 and Table 4.3.

	20	30	50	100	200	1000	10000
A priori	5.585	5.858	6.101	6.341	6.498	6.658	6.730
Uniform	5.130	5.529	5.897	6.234	6.438	6.652	6.741
Normal	4.126	4.459	4.795	5.116	5.345	5.588	5.691
Exponential	4.781	5.070	5.340	5.564	5.712	5.874	5.964

Table 4.1: Approximate 95% quantiles of the d_n statistic distribution for different sample sizes and null distributions, either specified a priori or with estimated plugged-in parameters, computed using simulation of $N = 10^6$ samples

	20	30	50	100	200	1000	10000
A priori	9.477	10.460	11.440	12.397	12.954	13.358	13.397
Uniform	7.960	9.275	10.667	11.958	12.704	13.328	13.440
Normal	4.260	4.926	5.705	6.589	7.203	7.755	7.928
Exponential	5.622	6.430	7.271	8.128	8.614	9.042	9.116

Table 4.2: Approximate 95% quantiles of the o_n statistic distribution for different sample sizes and null distributions, either specified a priori or with estimated plugged-in parameters, computed using simulation of $N = 10^6$ samples

	20	30	50	100	200	1000	10000
A priori	2.498	2.504	2.495	2.492	2.494	2.494	2.499
Uniform	1.867	2.033	2.180	2.314	2.399	2.470	2.484
Normal	0.722	0.731	0.740	0.746	0.750	0.750	0.750
Exponential	1.295	1.306	1.312	1.319	1.318	1.319	1.321

Table 4.3: Approximate 95% quantiles of the Anderson-Darling statistic distribution for different sample sizes and null distributions, either specified a priori or with estimated plugged-in parameters, computed using simulation of $N = 10^6$ samples

4.2 Power

The power of the statistics is approximated using a Monte Carlo approach. Each alternative distribution is sampled $N = 10^5$ times for samples of sizes

Distribution name	Distribution parameters
Symmetric Normal	$1/2\mathcal{N}(-0.88, 0.45^2) + 1/2\mathcal{N}(0.88, 0.45^2)$
Asymmetric Normal	$1/5\mathcal{N}(-1.68, 0.2^2) + 4/5\mathcal{N}(0.42, 0.6^2)$
Symmetric Beta	$1/2\text{Beta}(2, 8) + 1/2\text{Beta}(8, 2)$
Asymmetric Beta	$1/2\text{Beta}(2, 6) + 1/2\text{Beta}(12, 2)$
Gamma	$\text{Gamma}(0.3, 0.9)$
Gamma Mixture	$3/10\text{Gamma}(1.1, 2) + 7/10\text{Gamma}(5, 1.2)$

Table 4.4: Specifications of the alternative distributions. Here, $N(\mu, \sigma^2)$ denotes the normal distribution.

$n \in \{20, 30, 50, 100\}$. The alternative distributions are described in Table 4.4 and plotted in Figure 4.1. The estimators for the relevant parameters are (naturally) the same as those used for the computation of the quantiles and are reported in Section 4.1. The SW statistic is computed using `shapiro.test` included in the `stats` package in R [29]. The resulting power of the statistics is plotted in Figure 4.2.

A priori parameters The statistics d_n and o_n perform quite similarly. In most cases, d_n performs slightly better, except for the gamma case. In most cases, the statistics d_n and o_n considerably outperform the Anderson-Darling statistic, except for the gamma case, where the Anderson-Darling performs extremely well.

Estimated parameters We conclude the following:

- Across the board, the tests gain more power by using estimated parameters. However, the d_n and o_n gain less power compared to the Anderson-Darling statistic, especially in the normal setting.
- The Anderson-Darling statistic performs best in the gamma cases and in the normal-mixture cases; in the latter case, Shapiro-Wilk performs slightly worse.
- The d_n and o_n statistic perform best in the beta-mixture cases. The d_n statistic performs slightly better than the o_n statistic.

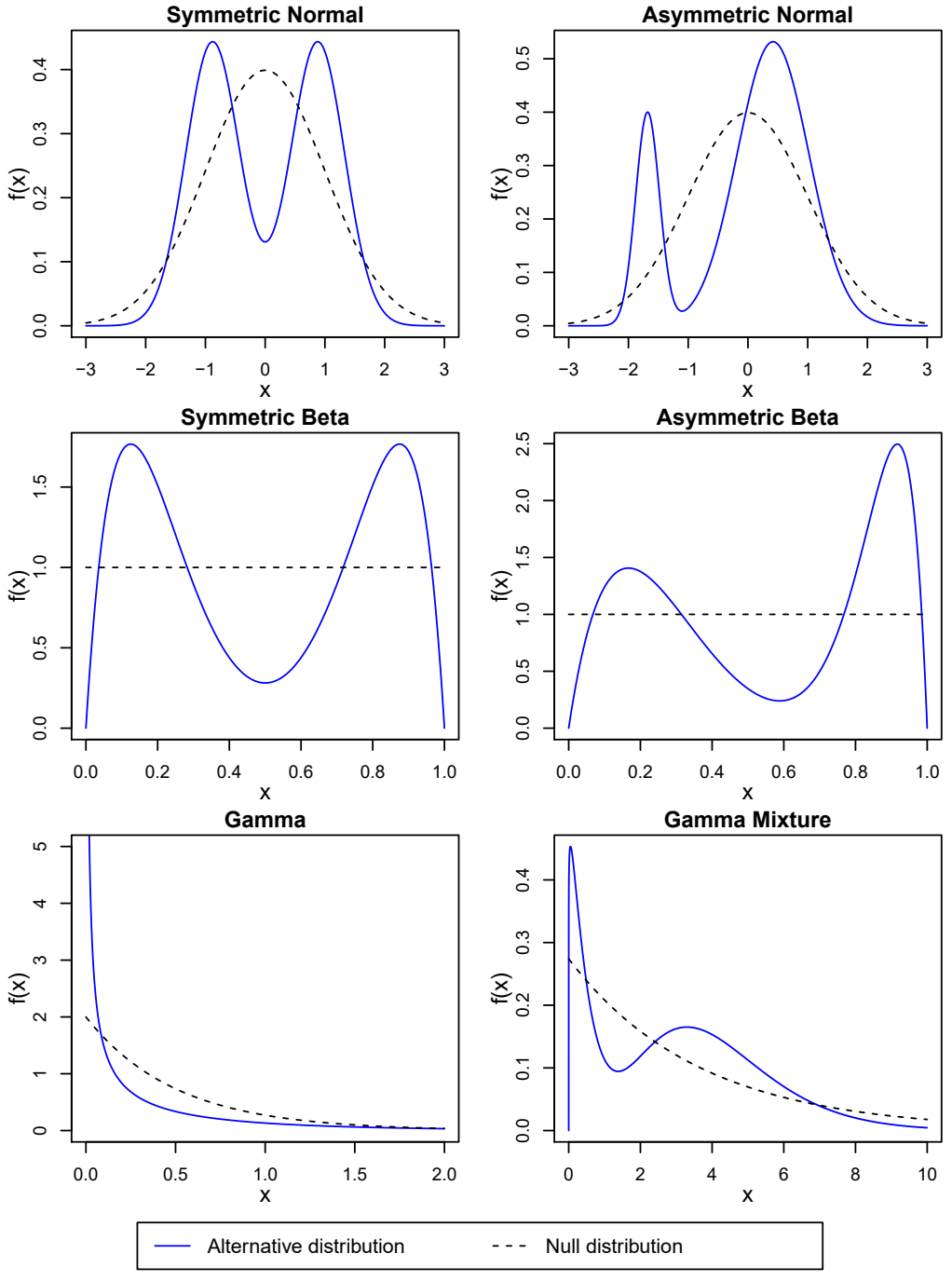

Figure 4.1: The alternative distributions considered for the power comparison, along with the assumed null distribution in each case.

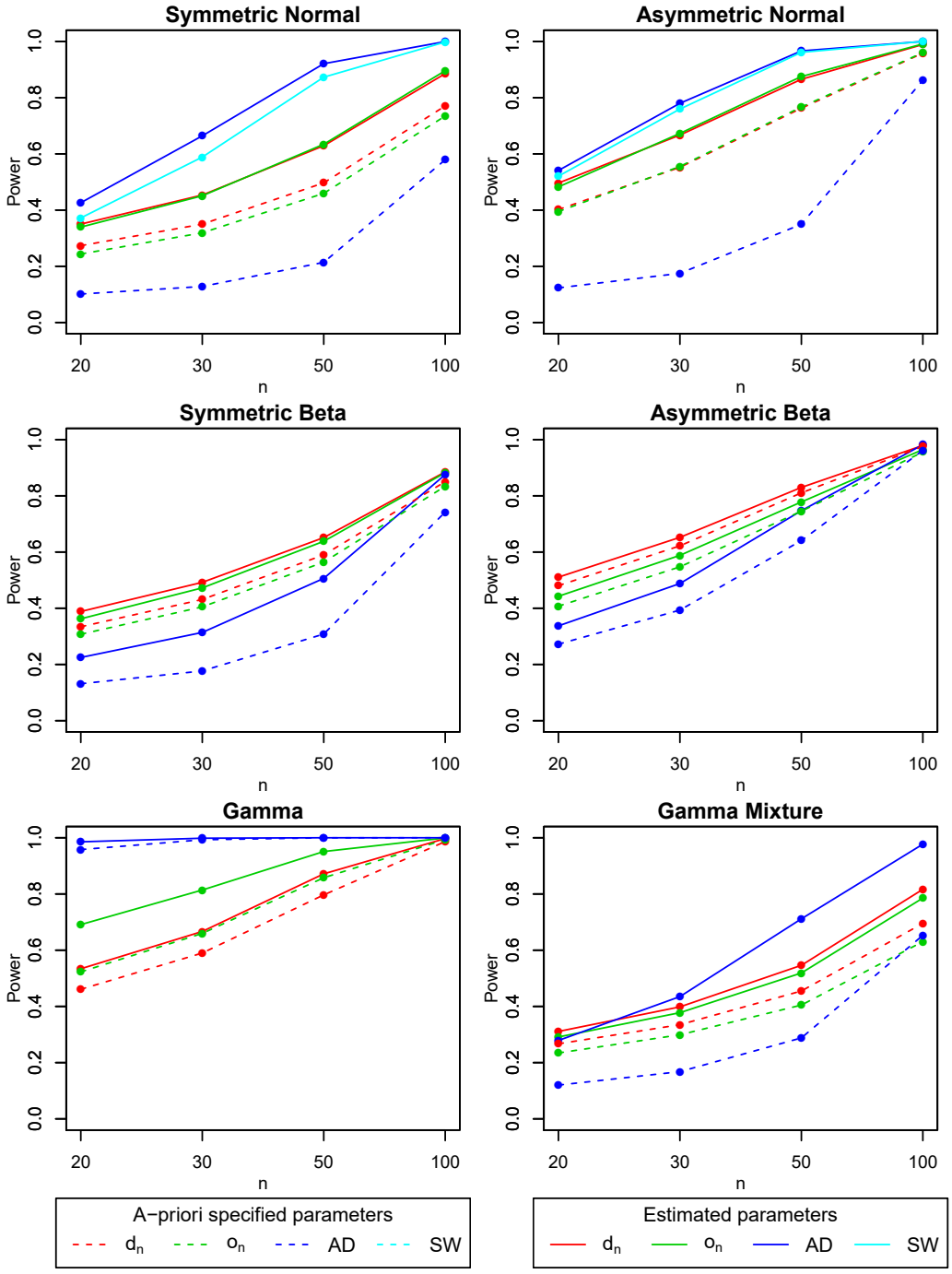

Figure 4.2: The power of the different statistics under different alternative distributions. Dashed lines specifiy the power when parameters are specified a priori, solid lines when parameters are estimated from the data. The Shapiro-Wilk test is only applied in the case of normal null distributions.

Chapter 5

Minimum-distance Parameter Estimation

The statistics based on the empirical identity process can also be used in the context of parameter estimation. Since the statistics serve as distance metrics between I_n and F_n, we can use them to construct minimum-distance estimators (MDE's).

Consider the following context. Given data $X = (X_1, \ldots, X_n)$, we assume a parametric model $X \sim F_\theta$ with unknown $\theta \in \Theta$. We can compute the probability integral transform $F_\theta(X)$ conditionally on the parameters θ. We can then compute the IEIP and ECDF of the transformed sample $F_\theta(X)$. As $F_\theta(X)$ depends on θ, so do the IEIP and ECDF. We denote these by $I_n^{(\theta)}$ and $F_n^{(\theta)}$ respectively.

The minimum-distance estimator is then given by:

$$\hat{\theta} = \text{arginf}_{\theta \in \Theta} \text{Dist}(I_n^{(\theta)}, F_n^{(\theta)}),$$

where Dist can be any statistic that measures the distance between I_n and F_n.

We consider this estimator in two parametric models; the exponential distribution and the dual normal mixture distribution. The exponential distribution is considered mainly for its simplicity, having only a single parameter. It is therefore only studied briefly. Our main focus is on the dual normal mixture model, which we study extensively.

For both models, the estimator performance is studied. Additionally, we study the sensitivity to initial parameters for the optimization; note that

the minimum distance estimator is the solution to an optimization problem. Finding the minima requires specification of initial parameters, and the final estimate may be dependent on this initialization. For the normal mixture model, we also study the robustness of the estimator. Finally, we also apply the estimators on a real dataset: Fisher's Iris data, included in R [29].

5.1 Exponential Distribution

In this section, we study the MDE's in the exponential parametric model.

5.1.1 Initialization Sensitivity

First, we examine the initial parameter sensitivity of the MDE in the one-dimensional exponential case. For initialization in the exponential case, we supply an optimization interval.

Consider optimization surface, given by the following function:

$$g_{\text{Dist}}(\theta) = \text{Dist}(I_n^{(\theta)}, F_n^{(\theta)}), \tag{5.1}$$

which is the function minimized for the MDE. Plotting g gives some insight in the behavior of the MDE. The resulting plots for several datasets are given in Figure 5.1. As can be seen in the plots, the functions g_{d_n} and g_{o_n} have minimums at generally the same location and seem to agree. Note that g_{d_n} has kinks where the location of the supremum in the statistic changes.

It is interesting to see if an optimization interval that is too broad could influence the result. If we choose an interval of the form $(0, \lambda_{\max})$, picking $\lambda_{\max}$ too large results in estimates at local minima instead of global minima. An example of this is given in Figure 5.2.

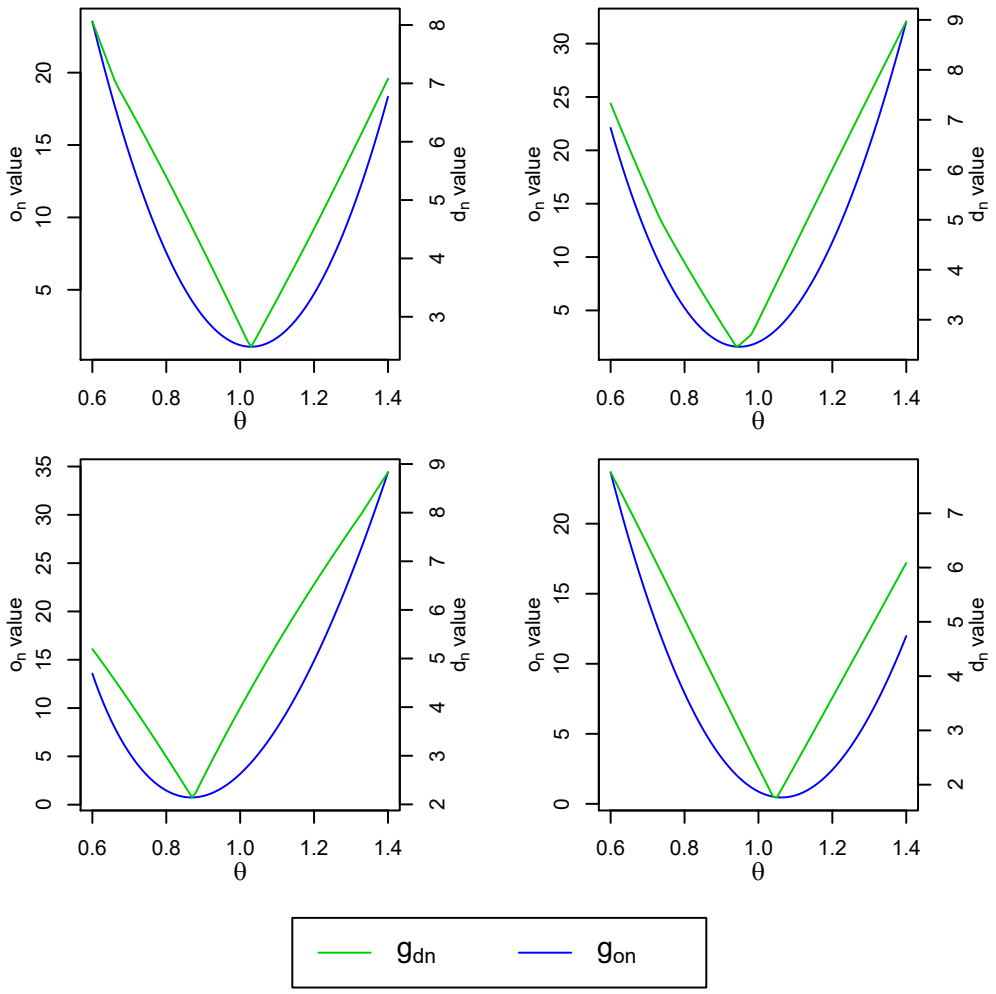

Figure 5.1: Plots of $g_{d_n}(\theta)$ and $g_{o_n}(\theta)$ for four different exponential data sets with true rate parameter θ_{true} equal to 1, of sample size $n = 100$, for $\theta \in [0.6, 1.4]$. Note the within one plot, the y-axis is different for the two functions.

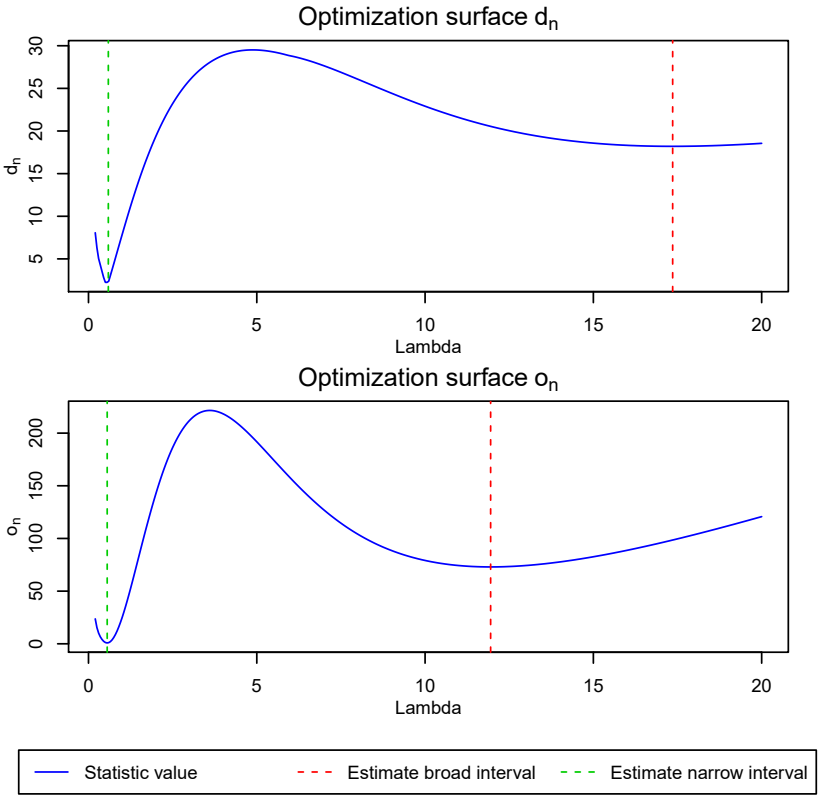

Figure 5.2: A specific case in the exponential setting with true rate parameter θ_{true} equal to 1. In this case, the MDE selects an estimate that is too large, when the optimization interval is too broad. The data set is of size $n = 20$.

5.1.2 Performance Study

The performance of the MDE's is studied by simulation. We simulate samples from the exponential distribution with rate parameter $\theta_{\text{True}} = 1$ of different sizes and compare the estimator performance in each case. For the MDE and MSP estimators, we need to supply an optimization interval. This is set to $(0, 2.5)$. The results are given in Figure 5.3.

We conclude that in this setting, the MSP performs the best overall, outperforming the MLE. As we discussed in Section 2.4.3, we should expect the MSP to outperform the MLE as the density is skewed. The MDE-d_n and MDE-o_n perform worse compared to the other estimators.

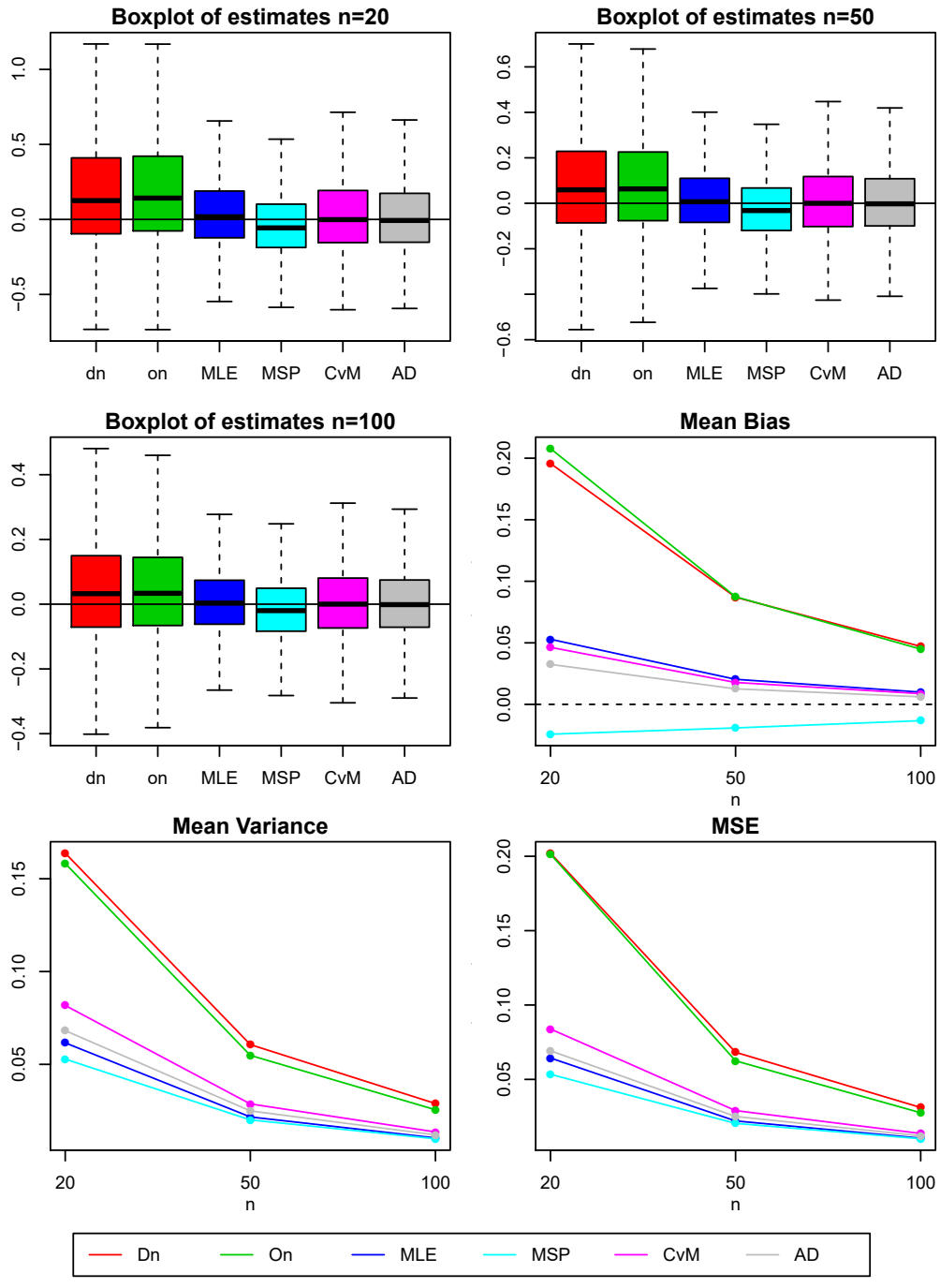

Figure 5.3: Performance measures of estimators in the exponential case with true rate parameter θ_{true} equal to 1. The true parameter has been subtracted from the estimates in the boxplots. Estimators have been initialized with the true parameters. Results are based on $N = 10^5$ samples of various sizes.

5.2 Dual Normal Mixture Distribution

In this Section, we study the behavior and performance of the MDE estimators in the dual normal mixture setting. The dual normal mixture density is given by:

$$f(x;\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) = \rho\frac{1}{\sigma_1}\phi(\frac{x-\mu_1}{\sigma_1}) + (1-\rho)\frac{1}{\sigma_2}\phi(\frac{x-\mu_2}{\sigma_2}), \tag{5.2}$$

where ϕ is the standard normal density. Four different sets of parameters are considered, which are given and named in Table 5.1. The mixtures are plotted in Figure 5.4, along with a normal density having the same mean and variance as the mixture for reference.

We assess the performance of our MDE in these distributions. For each model, we consider the MDE using both the d_n and the o_n statistic.

For the performance analysis, we compare the performance of the MDE-d_n and MDE-o_n against MDE's based on the Cramèr-Von Mises and Anderson-Darling statistic, as introduced in Section 2.4.4. We compare them with the MSP, as introduced in Section 2.4.3. Finally, we also compare the results with the well-known MLE, introduced in Section 2.4.2. We use the MLE in combination with the EM algorithm, introduced in Section 2.5. All estimators are implemented ourselves, except for the MLE-EM estimator. The EM algorithm from package `mixtools` [3] is used.

	μ_1	μ_2	σ_1	σ_2	ρ
Symmetric bimodal	-0.88	0.88	0.45	0.45	0.50
Asymmetric bimodal	-1.68	0.42	0.20	0.60	0.20
Symmetric unimodal	-0.40	0.40	0.45	0.45	0.50
Asymmetric unimodal	-0.10	0.40	0.60	0.20	0.80

Table 5.1: Parameter values of the different normal mixtures considered.

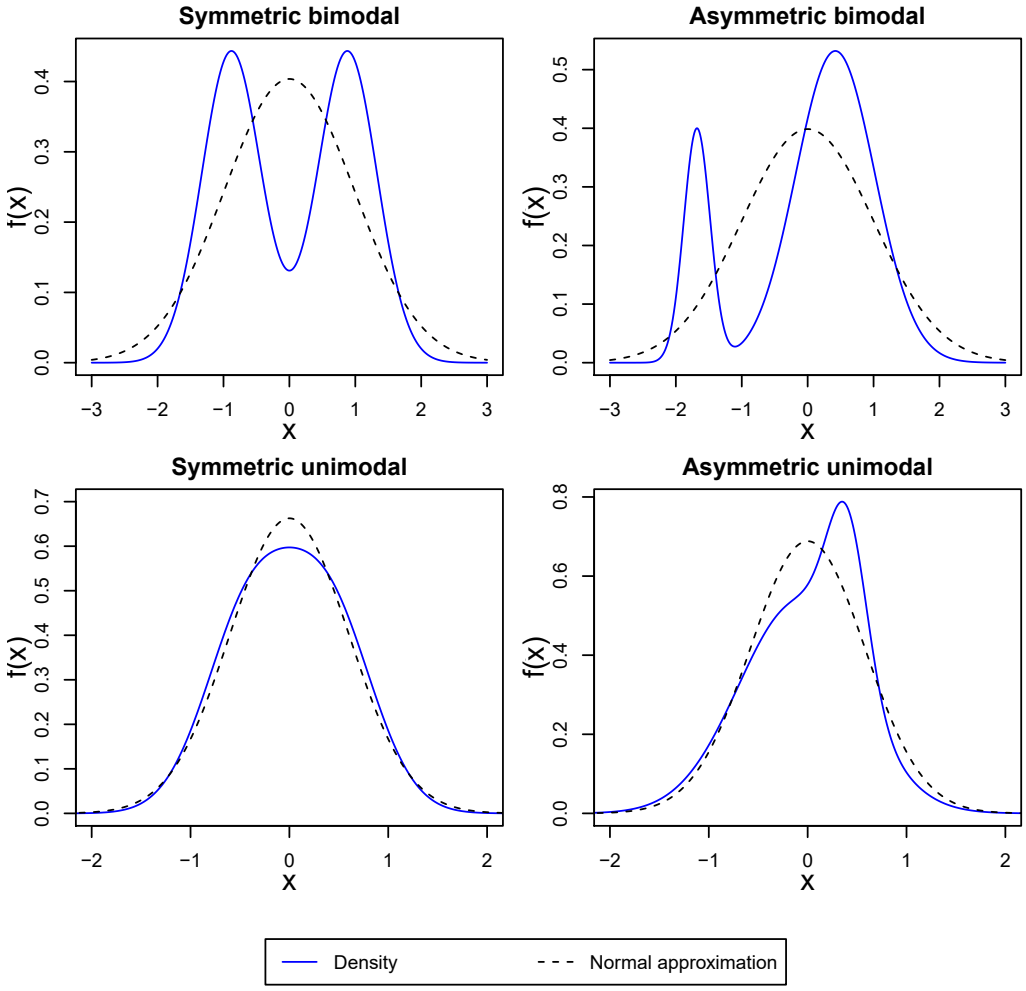

Figure 5.4: Density plots of the different normal mixtures considered, along with its normal approximation, i.e. a normal density having the same mean and variance as the mixture, for reference.

5.2.1 Interpretation of Performance Measures

Evaluating estimates for a dual normal mixture is non-trivial. While at first it may seem like we can simply compute marginal or joint performance measures of the parameters, further inspection shows these can be misleading.

Consider the case where the weight parameter is measured at one of the extremes, e.g. $\rho = 1$. In this case, $f(x; \hat{\theta})$ does *not* change for any value of μ_2, and interpretation of the MSE of μ_1 becomes difficult. Furthermore, the volume of the MSE concentration ellipsoid is dependent on this MSE, and its interpretation becomes more difficult too.

Hellinger distance. Because of the problem, one could argue that it might be better to consider distance measures between the estimated density $f(x; \hat{\theta})$, which is the density (5.2) with plugged-in estimates, and the true density $f(x; \theta)$. One could consider for instance the Hellinger distance[†] as a measure of the distance. However, considering such distances can lead to preference of parameters sets where the weight parameter is measured at the extremes. This is best highlighted by an example.
Consider the symmetric unimodal mixture, defined in Table 5.1, and take the Hellinger distance as a distance measure. Then, the distance for the parameter set $\theta_A = (\mu_1, 0, \sigma_1, 0.6, 0)$ is approximately 0.00136, for any value of μ_1 and σ_1. The distance for $\theta_B = (-0.4, 0.4, 0.4, 0.4, 0.5)$, which are almost equal the true parameters, except slightly underestimating the standard deviations, has distance 0.00169. Therefore, θ_A would be preferred according to this measure. However, this parameter set corresponds to a single-component normal distribution, while we might prefer to find the optimal dual-component parameter estimates.

Modality. Another disadvantage of considering the distance measures is that it does not take other qualities of the estimated density $f(x; \hat{\theta})$ into account, such as the modality[‡]. For instance, suppose we have to estimates

[†]The (squared) Hellinger distance between functions f and g is given by:

$$H^2(f, g) = 1 - \int \sqrt{f(x)g(x)}dx. \tag{5.3}$$

[‡]Finding the number of modes of the dual normal mixture is not trivial. Sufficient conditions for unimodality are relatively easy, but necessary and sufficient conditions are somewhat harder. However, conditions are given in [36], which we use. The conditions involve checking if the weight parameter lies within an interval, for which the boundaries are found by finding roots of a polynomial.

θ_{uni} and θ_{bi}, and suppose the Hellinger distance of θ_{bi} is lower than θ_{uni}. If the true densitiy is unimodal, we might judge θ_{uni} to be a better estimate, if the modality of the final distribution is of importance. Whether or not this modality is of importance is dependent on the application.

Finally, in many applications the parameters have clear interpretation and are interesting in their own right. They must therefore be estimated carefully. It is therefore not wise to judge an estimator merely by its performance regarding the distance between $f(x;\hat{\theta})$ and $f(x;\theta)$

Conclusion. In conclusion, we can therefore not limit ourselves to merely (marginal) performance measures of either parameter values or estimated densities, and we must consider both to give a reasonable interpretation of estimator quality. However, this makes it particularly difficult to compare estimators in this setting. In particular, we must pay attention to the weight parameters estimated. When these are estimated at the extremes, the MSE of the other parameters must be interpreted with care.

5.2.2 Initialization Sensitivity

We now study the initialization sensitivity of the MDE's in the normal mixture model with density (5.2). We consider the density with parameters of the symmetric bimodal normal mixture, defined in Table 5.1. For the MDE's, must now solve a 5-dimensional optimization problem and we must start the optimization routine with an initial guess. We study the sensitivity of the estimators to this initial guess, and study different methods for finding reasonable initial values.

Since the optimization surface g, defined in (5.1), is now 5-dimensional, we cannot plot the entire surface. Instead, we plot the surface of g when parameters μ_2, σ_2 and ρ are set to their true values. The resulting level plots are given in Figure 5.5.

Note that the functions g_{d_n} and g_{o_n} generally agree on the area of the minimum. However, we note that the value of g_{o_n} increases much faster when moving away from the minimum, while g_{d_n} has areas further from its minimum with relatively low values as well. This could indicate that optimization procedures over g_{d_n} might be more sensitive to the initial parameters than over g_{o_n}.

Therefore, we also study the case where we first find a minimum using the

MDE-o_n estimator, and then find the final set of parameters using the MDE-d_n estimator, with initial values equal to the old MDE-o_n estimates. This "composition" of estimators is denoted by MDE-$d_n(o_n)$.

We now study the sensitivity of the MDE's to the initial parameters using simulation. We also use the MLE-EM algorithm for reference. We again consider the symmetric bimodal normal mixture with density (5.2), with parameters defined in Table 5.2. We use a fixed sample size of $n = 100$. For all simulation results, we order the estimates as $\mu_1 < \mu_2$.

Deterministic Pertubations

We first compare the initialization sensitivity to *deterministic* perturbations of the initial parameters. We compare the performance when the estimators are initialized with the true parameters, except for a fixed perturbation in μ_1, σ_1 or ρ. We perturb μ_1 by adding 0.18, σ_1 by adding 0.2 and ρ by adding 0.2. This allows us to see if the methods are sensitive to a single parameter in particular.

The simulation results show that the estimators are not sensitive to an initial parameter in particular. However, we do see that MDE-d_n is more sensitive to the perturbation than the other estimators. Results are deferred to Appendix B.1.1. Boxplots of the attained likelihood are given in Figure B.1. The volumes of the MSE concentration ellipsoids, as defined in Section 2.3.2, are given in Figure B.2. The MSE of each parameter is given in Figure B.3 and the bias of each parameter in Figure B.4.

Random Initializations

Next, we compare the initialization sensitivity when using *random* initialization methods. We use the three methods discussed in Section 2.5.2.

The simulation results show that the K-means++ initialization is superior to the other methods. The other two random methods perform worse. Plots of the attained likelihood are given in Figure 5.6. Other results are deferred to Appendix B.1.1. Volumes of the MSE concentration ellipsoids, as defined in Section 2.3.2, are given in Figure B.5. The MSE of each parameter is given in Figure B.6 and bias of each parameter in Figure B.7.

Conclusions. With respect to initialization sensitivity, we conclude for this particular model:

- The MDE's seem equally sensitive to deterministic perturbations in any parameter.
- Out of the three random initializations considered, K-means++ performs the best for all estimators.
- The MDE-d_n method is most sensitive to the initialization. The MDE-o_n and MDE-$d_n(o_n)$ are much more stable. The MLE-EM is even more stable.
- The composed estimator MDE-$d_n(o_n)$ does not improve the initialization sensitivity, nor the performance when compared to MDE-o_n.

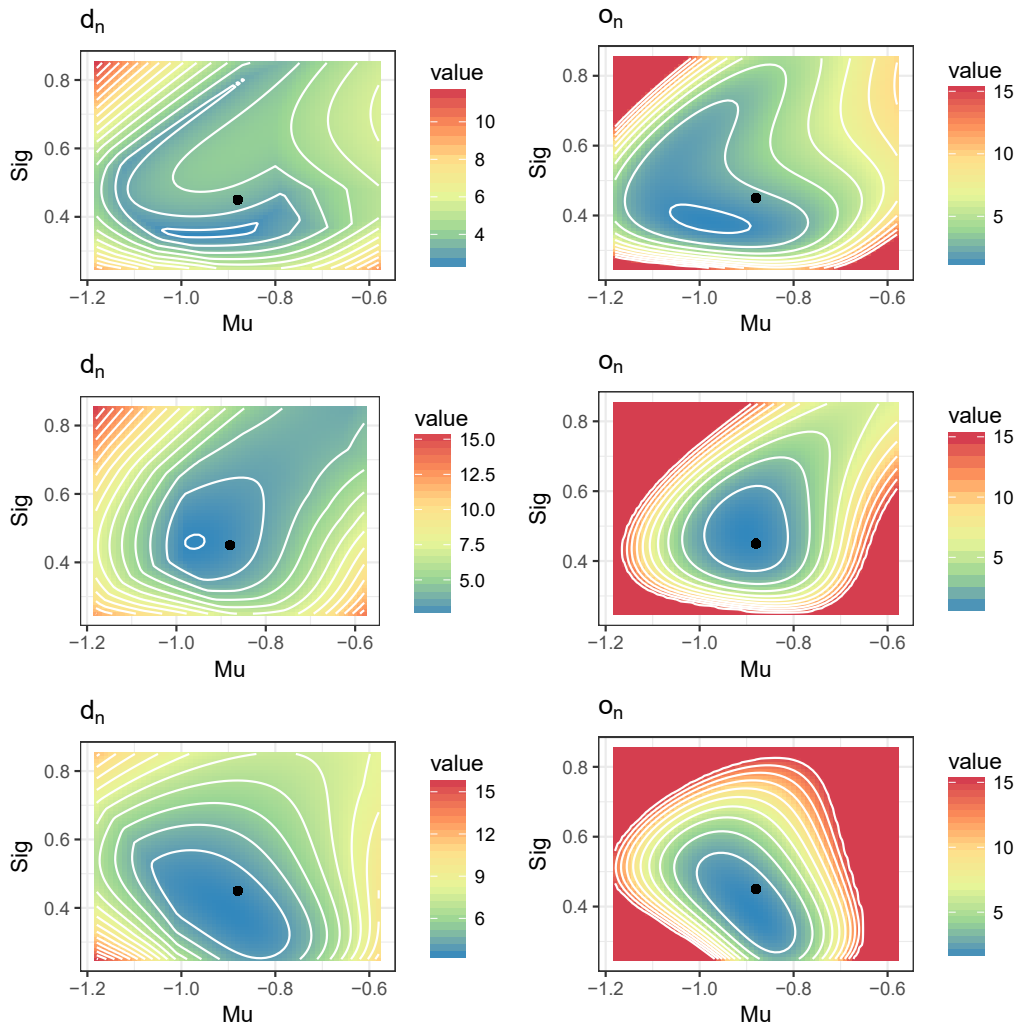

Figure 5.5: Plots of g_{d_n} and g_{o_n} for three symmetric bimodal normal mixture datasets of size $n = 100$. The parameters μ_2, σ_2 and ρ are set to their true values. A point is included in each plot indicating the true parameter values of μ_1 and σ_1. The value of the o_n statistic is limited to 15 to ensure granularity around the minimum.

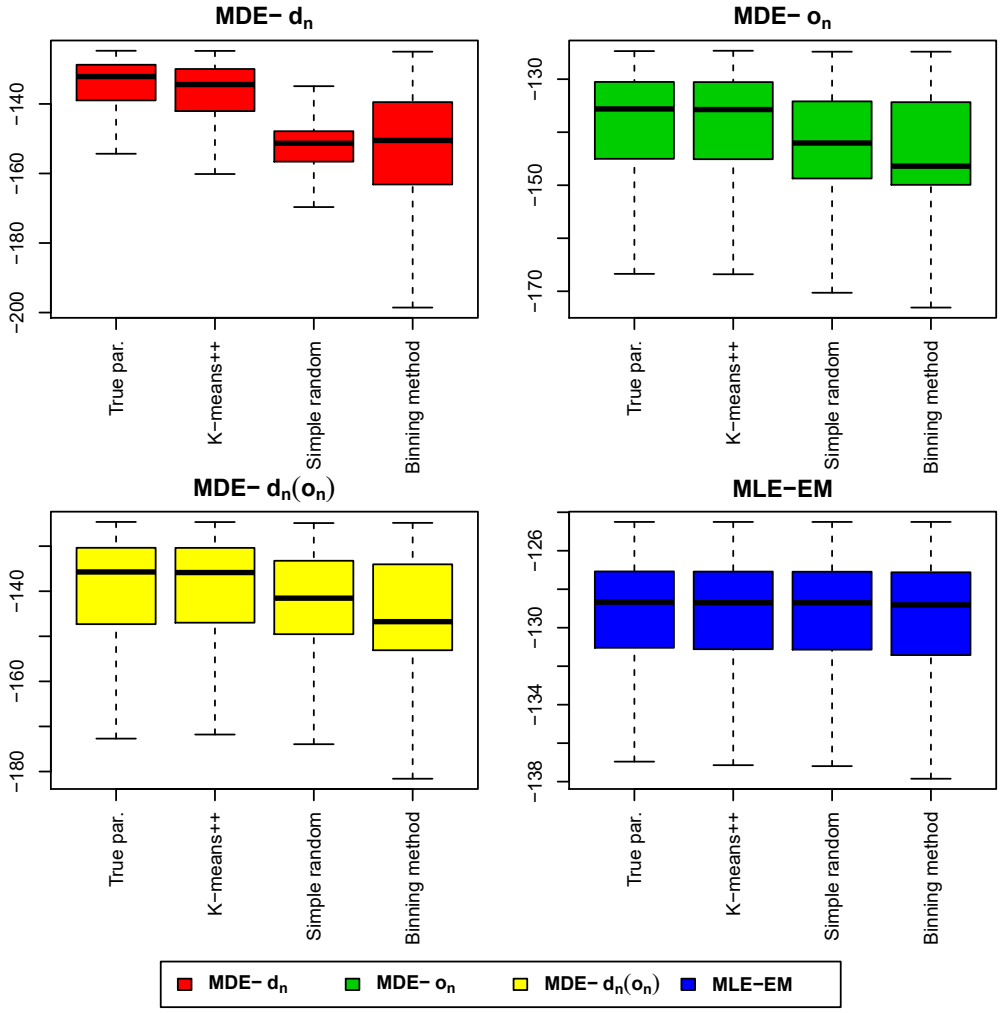

Figure 5.6: Boxplots of the attained likelihoods of the estimators using different random initialization schemes, for the symmetric bimodal normal mixture distribution. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

5.2.3 Performance Study

Initialization with True Parameters

We now study the performance of the estimators in the normal mixture model with density (5.2), when all methods are initialized with the true parameters. We study all four different dual normal mixtures models given in Table 5.1. For all simulation results, we order the estimates as $\mu_1 < \mu_2$.

The performance is assessed by simulation. The histograms of the weight parameters is given in Figure B.8. The volumes of the MSE concentration ellipsoids, as defined in Section 2.3.2, are given in Figure 5.7. The MSE for each parameter is given in Figure 5.8. The mean Hellinger distances, as defined in (5.3), are given in Figure 5.9. For the bimodal models, the densities with plugged-in estimates are nearly always bimodal for all estimators considered. For the unimodal models, the fractions of densities with plugged-in estimates that result in the correct number of modes are given in Figure 5.10.

Boxplots of the estimates of each parameter for each of the four distributions are deferred to Appendix B.1.2 and given in Figure B.9, Figure B.10, Figure B.11 and Figure B.12, for sample size $n = 50$.

There is a clear difference in the performance of the estimators between the bimodal and the unimodal mixture densities. We observe:

- The weight parameters are always distributed fairly clustered around the true values, except for the MLE-EM and MSP in the symmetric unimodal distribution, which is distributed somewhat uniformly, but not excessively at the extremes. Furthermore, there are no extreme estimates of means or variances of a single component with very small weight. Therefore, the MSE performance measures can be interpreted quite clearly in all cases.

- For the symmetric mixtures, the MDE perform noticeably worse in the second component compared to the first, both in terms of the means and the variances. They also overestimate the weight of the first component. Since the components in these symmetric distributions only differ in their mean location, having equal variance and weight in the mixture, this is not expected[†].
- For the unimodal mixtures, the MDE-d_n is considerably better at producing estimates for which the density has only a single mode. Especially in small samples, the MLE-EM erroneously results in a bimodal mixture often.

Conclusions. We conclude the following in this setting, when initializing with the true parameters:

- The MDE-o_n is inferior to the MDE-d_n.
- The MDE-d_n could be preferred for unimodal mixtures. The MSE is of an order of magnitude lower than the other estimators. It also results in densities with a single mode the most.
- The MLE-EM could be preferred for bimodal mixtures. For low sample sizes and symmetric mixtures, the MDE-d_n may give comparable or better results. For low sample sizes and asymmetric mixtures, the MSP may give comparable better results.

[†]When we relabel the two components, the MDE perform worse in the first component, and now underestimates its weight. It is therefore not caused by labels; the MDE perform worse and underestimate the weight of the component with the *rightmost mean*. At first sight, the fact that we use the *lower* empirical identity process for our inference seems like a possible cause for this behavior. However, this is not the case, as illustrated in Appendix A.1.

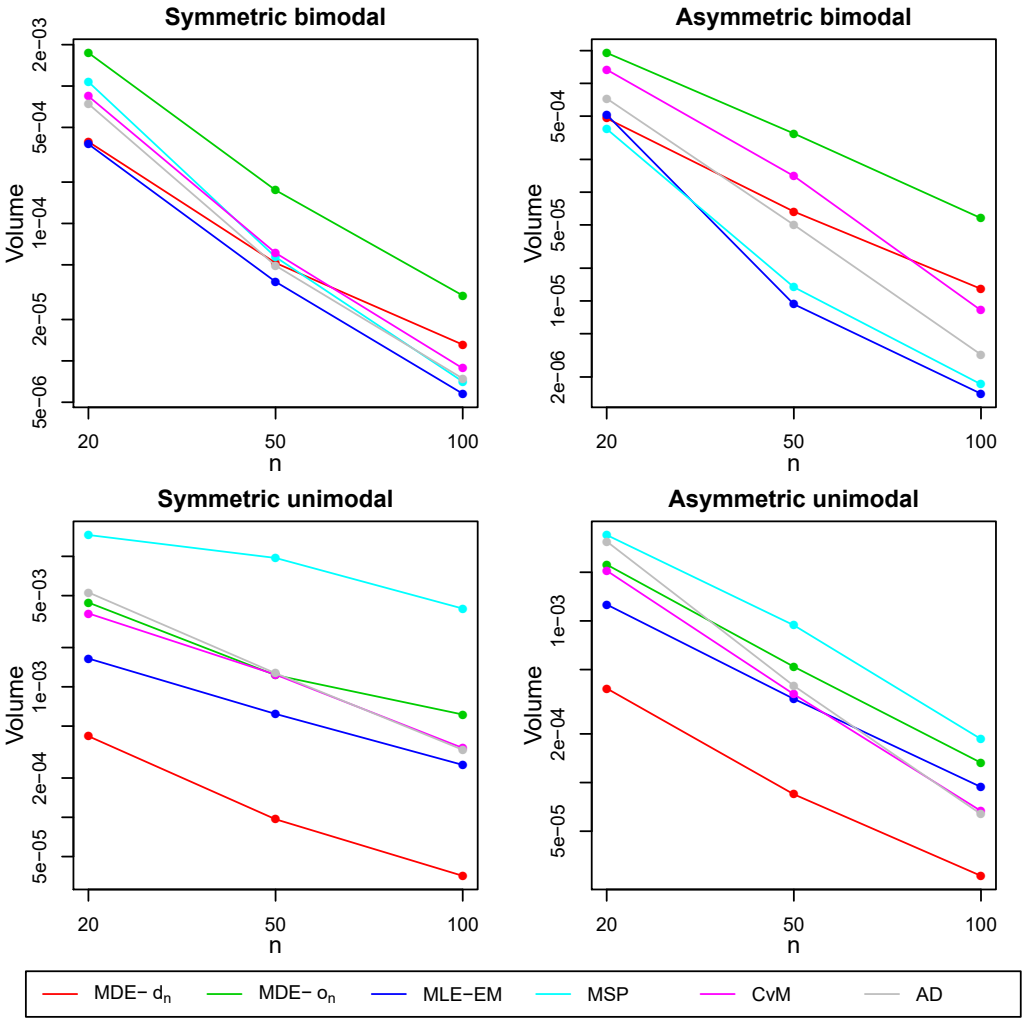

Figure 5.7: Volumes of the MSE concentration ellipsoids for the four normal mixture distributions, for different values of n. Note that the y-axis is in logarithmic scale. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples.

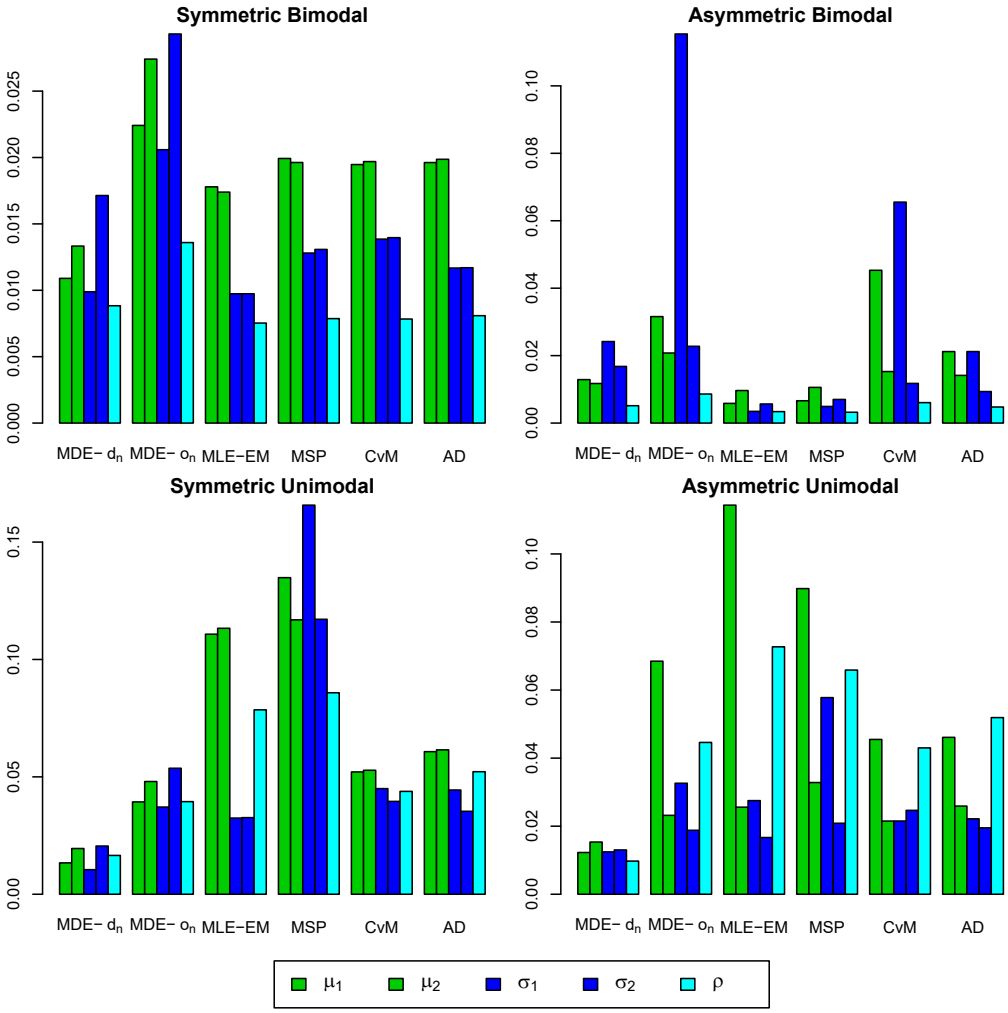

Figure 5.8: Barplot of the MSE of each parameter for the four normal mixture distributions, with sample size $n = 50$. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples.

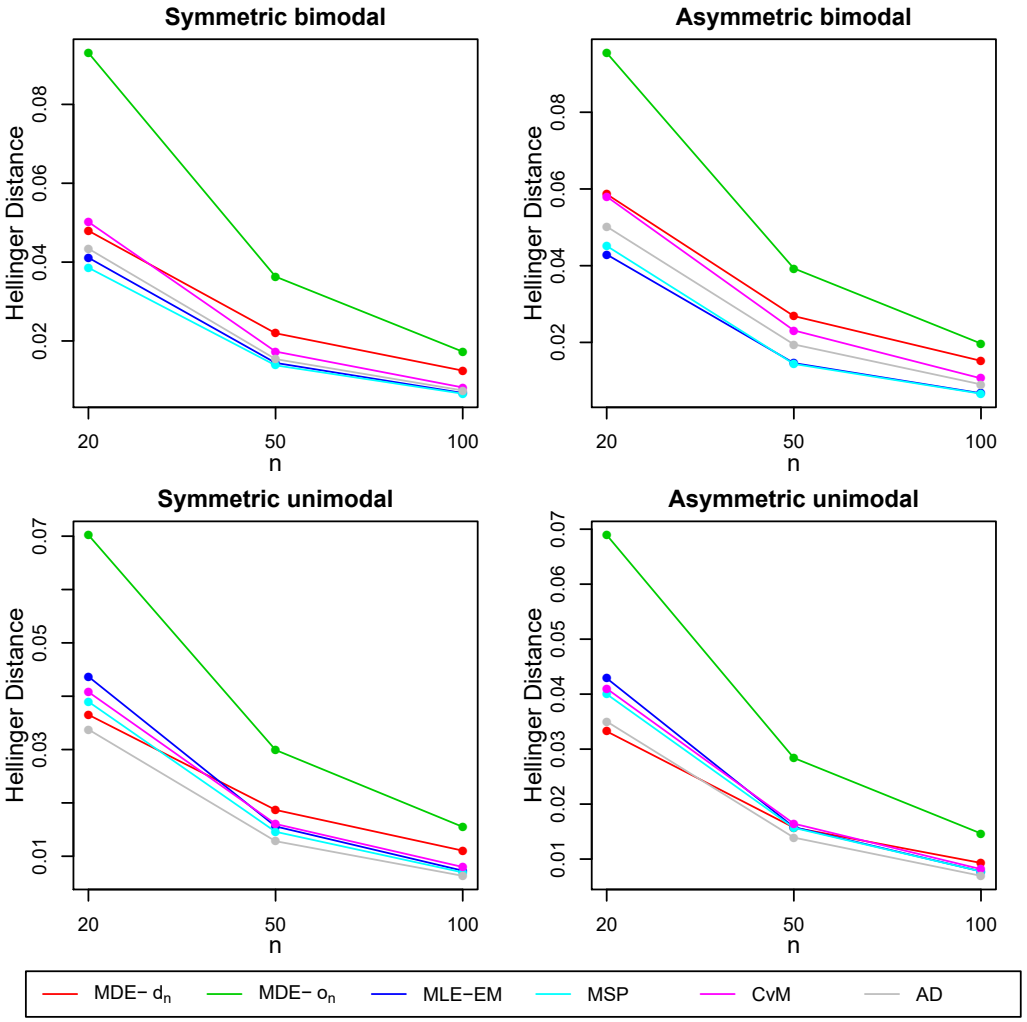

Figure 5.9: Means of the Hellinger distances between the true CDF and the estimated CDF, for the four normal mixture distributions, for different values of n. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples.

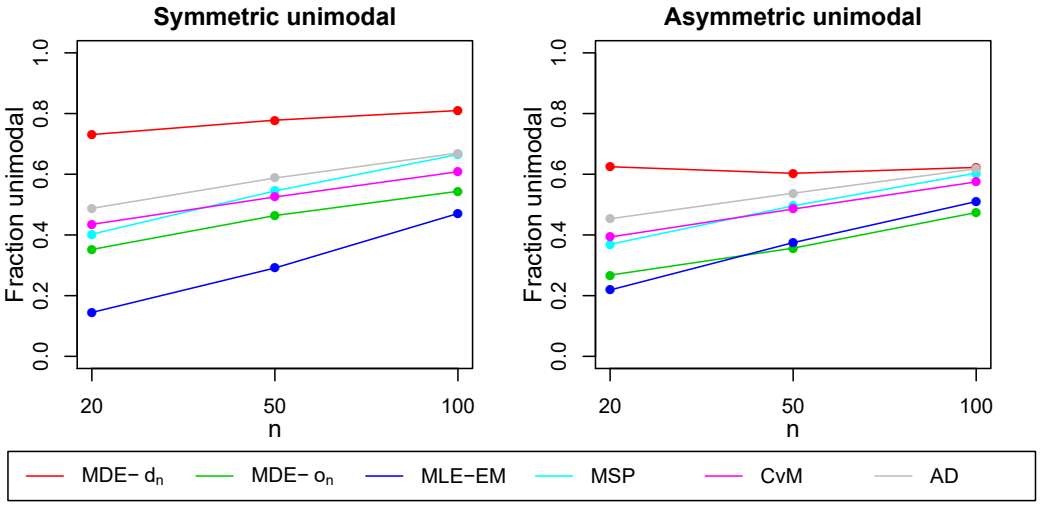

Figure 5.10: Fraction of the estimates that result in the density with plugged-in estimates to have the correct number of modes, for the two unimodal normal mixture distributions, for different values of n. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples.

Random Initialization

We now study the performance of the estimators in the normal mixture model with density (5.2), under random initialization. In Section 5.2.3, we observed good performance of the MDE-d_n estimator in the case of unimodal mixtures. However, this performance was obtained when initializing with the true parameters. In practice, we must use an initialization method, and we observed in Section 5.2.2 that the initialization sensitivity of the MDE-d_n method is relatively high. We therefore now assess the performance when initializing with a random initialization. We consider the MDE methods based on d_n and o_n, as well as their composition, defined in Section 5.2.2. We compare the performance with the MLE-EM estimator. The MSP and other MDE estimators are not considered.

We concluded in Section 5.2.2 that the K-means++ method performs the best, so we use it for this performance analysis. In Section 5.2.3, we concluded the MDE-d_n to have good performance for unimodal mixtures. Therefore, we study the performance of the MDE-d_n in the unimodal mixtures, defined in Table 5.1, under K-means++ initialization.

We study the performance by simulation. The histograms of the weight estimates are given in Figure B.13. The volumes of the MSE concentration ellipsoids are given in Figure 5.11. The MSE of each parameter is given in Figure 5.12. The fractions of correctly estimated modes are given in Figure 5.14. The mean Hellinger distances, as defined in (5.3), are given in Figure 5.13.

The weight estimates do not tend towards the extremes excessively, although the MLE-EM estimates are distributed more uniformly on the interval than the other estimators. Like in Section 5.2.3, the MSE performance measures are still interpretable. Note that the estimated weights of the asymmetric mixture are usually quite far from the truth for all estimators.

We observe that in terms of MSE, the MLE-EM and MDE-d_n are now similar. The marginal MSE's are slightly lower for MDE-d_n, but when we also consider the dependence between parameter estimates through the MSE matrix and the corresponding concentration volume, the MLE-EM and MDE-d_n are similar. The mean Hellinger distances of the densities with plugged-in parameters are lower for the MLE-EM. The fraction of estimates that result in unimodal densities are similar for the MLE-EM and MDE-d_n.

Conclusions. We conclude that, in the random initialization setting, under unimodal mixtures, the MLE-EM and MDE-d_n method perform similarly.

Multiple initializations. We see that the excellent performance of the MDE-d_n estimator when initializing with the true parameters is reduced when initializing with the K-means++ initializaton. One could wonder if this lost performance can be recovered by initializing multiple times, and selecting the estimate with the lowest distance. In Appendix B.1.2, we include a discussion regarding this idea.

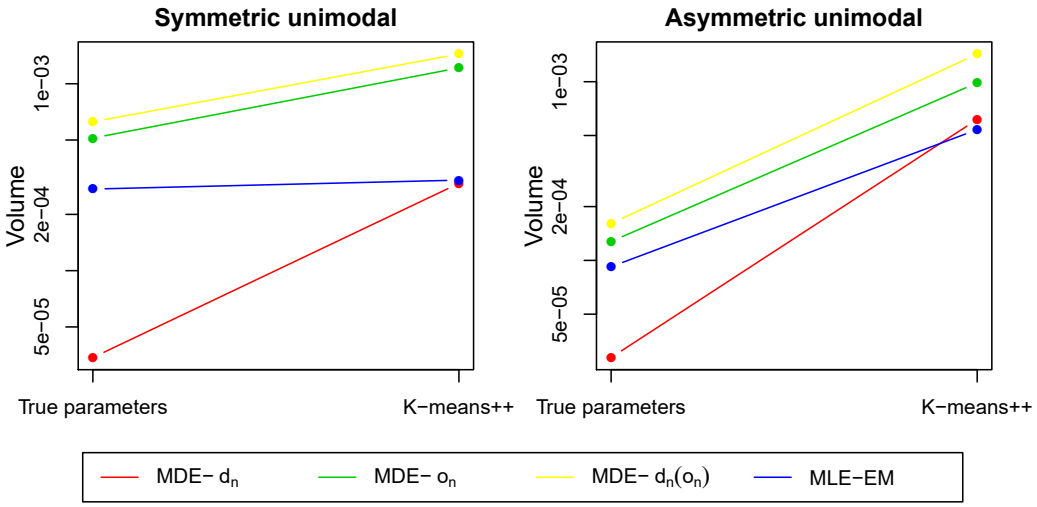

Figure 5.11: Volumes of the MSE concentration ellipsoids, using true and K-means++ initialization, for the two unimodal normal mixture distributions. Note that the y-axis is in logarithmic scale. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

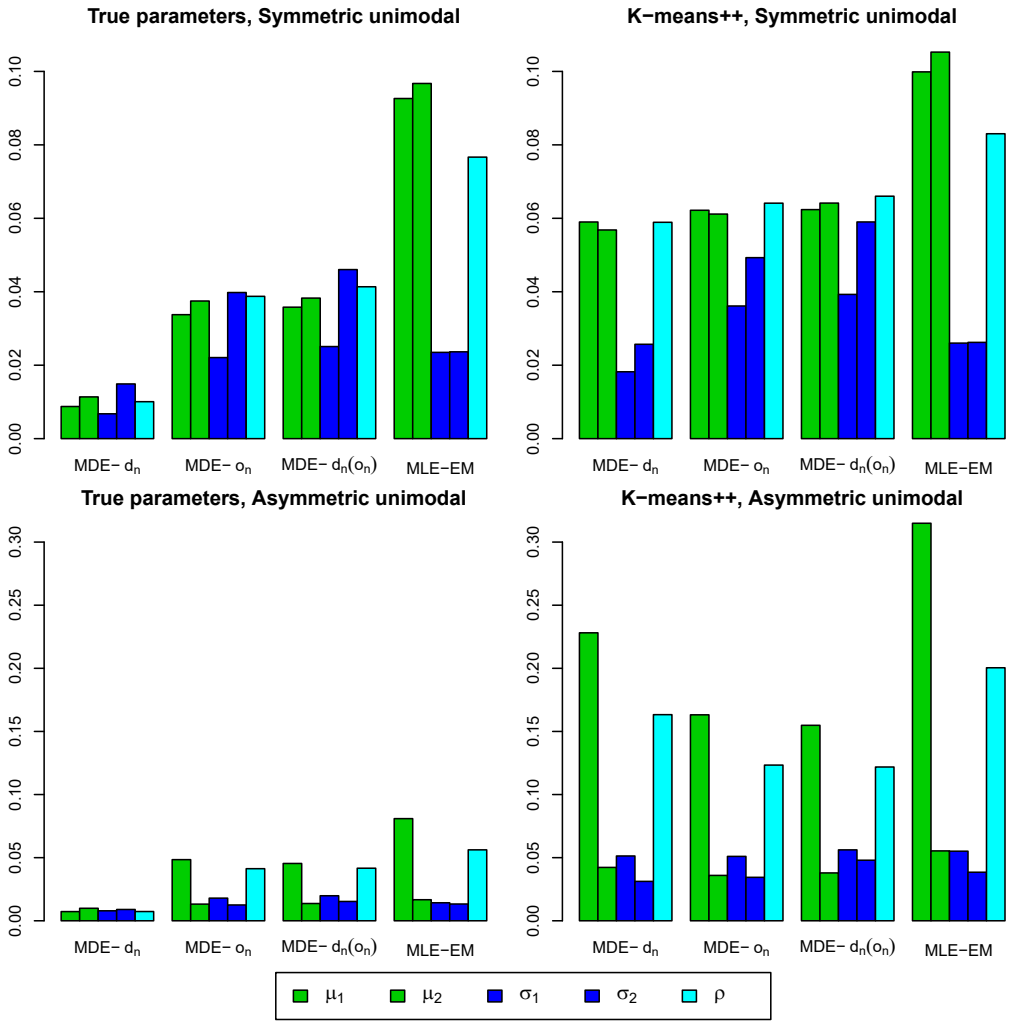

Figure 5.12: Barplots of the MSE of the estimators, using true and K-means++ initialization, for the two unimodal normal mixture distributions. Note that the y-axis is of different scale for the first two and the second two barplots. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

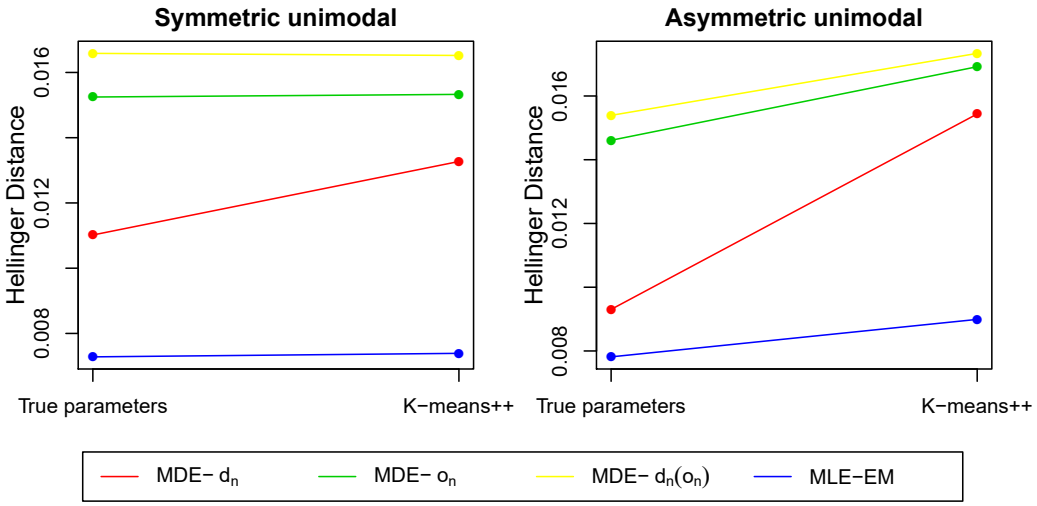

Figure 5.13: Means of the Hellinger distances between the true CDF and the estimated CDF, using true and K-means++ initialization, for the two unimodal normal mixture distributions. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

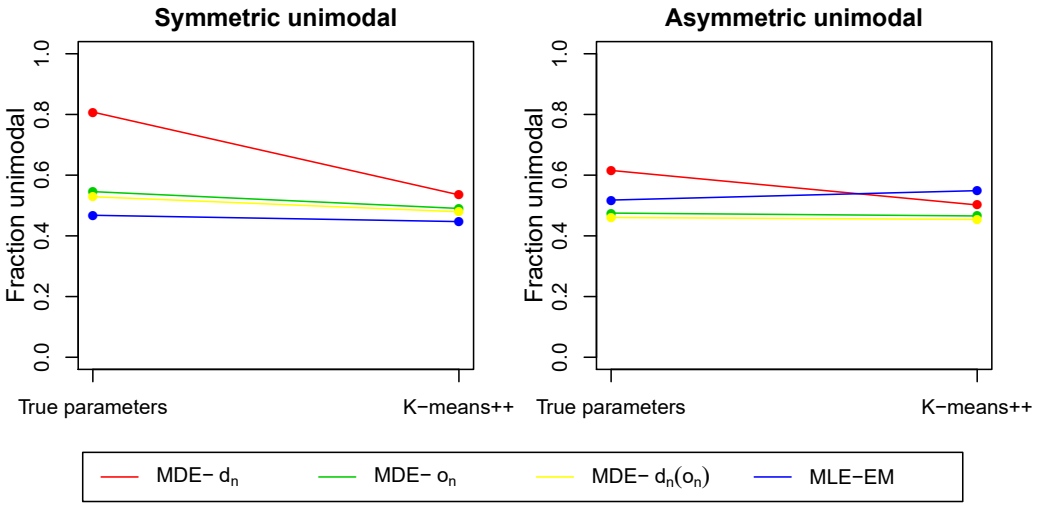

Figure 5.14: Fraction of the estimates that result in the density with plugged-in parameters having the correct number of modes, using true and K-means++ initialization, for the two unimodal normal mixture distributions. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

5.2.4 Robustness Study

Initialization with True Parameters

Minimum distance estimators have been studied in the setting of robust estimation, as discussed in Section 2.4.4. We now study the robustness of the minimum distance estimators based on the d_n and o_n statistic in the normal mixture model, when methods are initialized with the true parameters.

We consider the dual normal mixture densities as in Table 5.1, but instead of normal components as in (5.2), we use components of t-distributions:

$$f^{(\nu)}(x; \mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \rho \frac{1}{\sigma_1} f_t(\frac{x - \mu_1}{\sigma_1}; \nu) + (1 - \rho) \frac{1}{\sigma_2} f_t(\frac{x - \mu_2}{\sigma_2}; \nu), \quad (5.4)$$

where $f_t(x, \nu)$ is the density of the t-distribution with ν degrees of freedom. Thus, $f^{(\nu)}$ is a mixture of two t-distributions, where each component has ν degrees of freedom. The parameters μ_1, μ_2, σ_1, σ_2 and ρ are estimated with the assumption that $f^{(\nu)}$ is a *normal* mixture. Naturally, this model assumption is violated; the smaller the value of ν, the more this assumption is violated. Smaller values of ν increase the probability of outliers. The value of ν is not estimated; the estimators assume $\nu = \infty$.

As shown in Section 5.2.3, the MDE-d_n performs very well compared to the MLE-EM estimator in the setting of unimodal normal mixtures, when methods are initialized with the true parameters. We therefore check for robustness in the same setting. We use the t-distribution components as in (5.4) as a contamination, and use the same parameterizations as the unimodal densities as given in Table 5.1. The mixture densities are plotted in Figure 5.15 for various values of ν.

We compute results for sample size $n = 100$ for various values of ν to see the influence of the outliers on the estimates. The volumes of the MSE concentration ellipsoids, as defined in Section 2.3.2, are given in Figure 5.16. The MSE for each parameter is given in Figure 5.17. The mean Hellinger distances, as defined in (5.3) are given in Figure 5.18. Histograms of the weight parameter ρ are given in Figure 5.19 and Figure 5.20.

We note the following:

- The estimates of the weight parameter from the MLE-EM algorithm tend to be close to either 0 or 1 in the contaminated case. This means the distribution with plugged-in parameters is essentially a single component normal distribution. The MDE-d_n has no such behavior.
- The MSP performs dramatically worse compared to the other estimators, having extremely high MSE especially for the first scale parameter.
- The MDE-AD is notably sensitive to the outliers in the symmetric unimodal case. Overall, it is not as robust as the MDE-CvM.

Conclusions. We conclude the following for this robustness setting:

- The MDE-d_n performs very well even under the alternative model. It is very robust to outliers in this setting. The MDE-o_n is likewise robust, but has lower performance.
- The MLE-EM has considerably larger MSE and tends to return parameterizations that consist of one dominant component. The estimator is not robust.

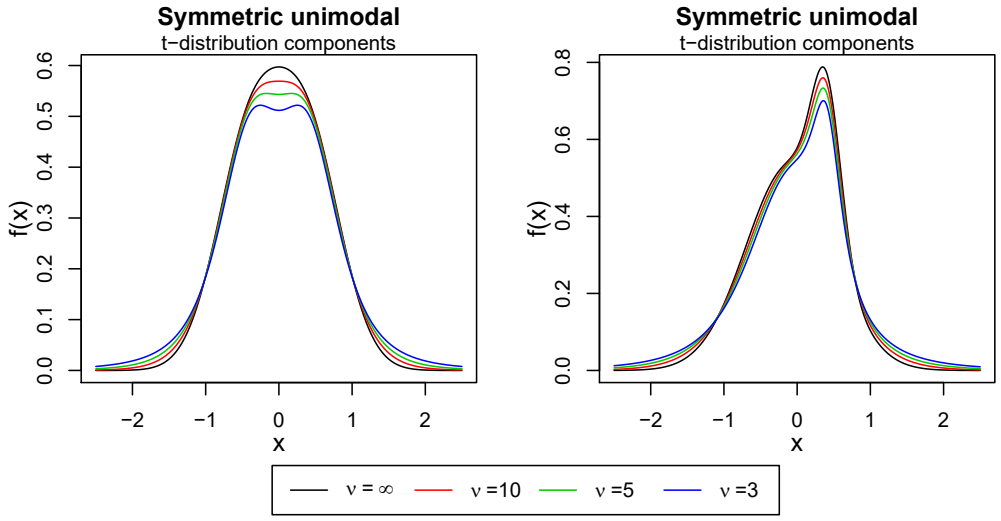

Figure 5.15: Plots of the two distributions considered for the robustness study for different values of ν. The distributions consist of two components of t-distributions with equal degrees of freedom ν. The location and scale parameters are equal to those of the normal mixture distributions with the same names.

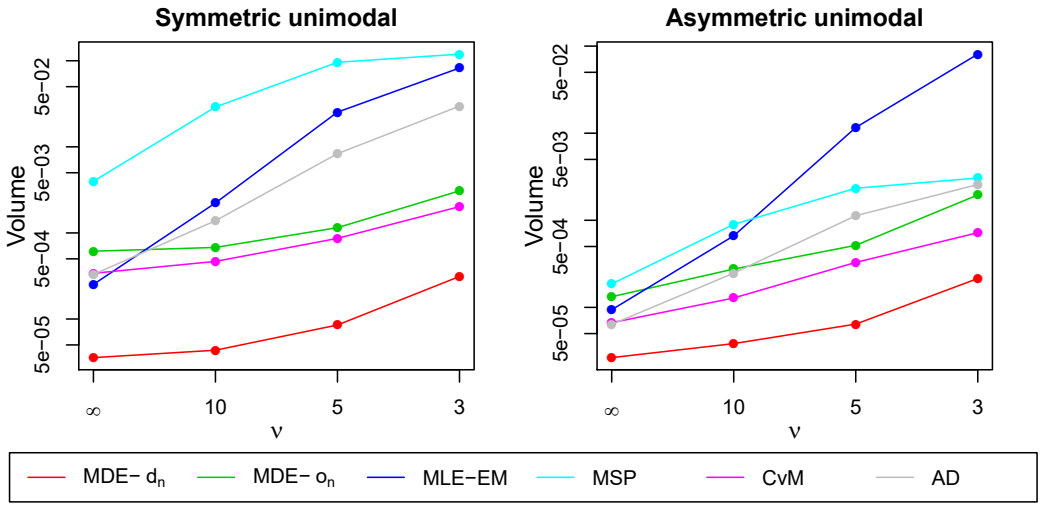

Figure 5.16: Volumes of the MSE concentration ellipsoids for the two unimodal mixture distributions, with t-distribution components with degrees of freedom ν, for different values of ν. Note that the y-axis is in logarithmic scale. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples of size $n = 100$.

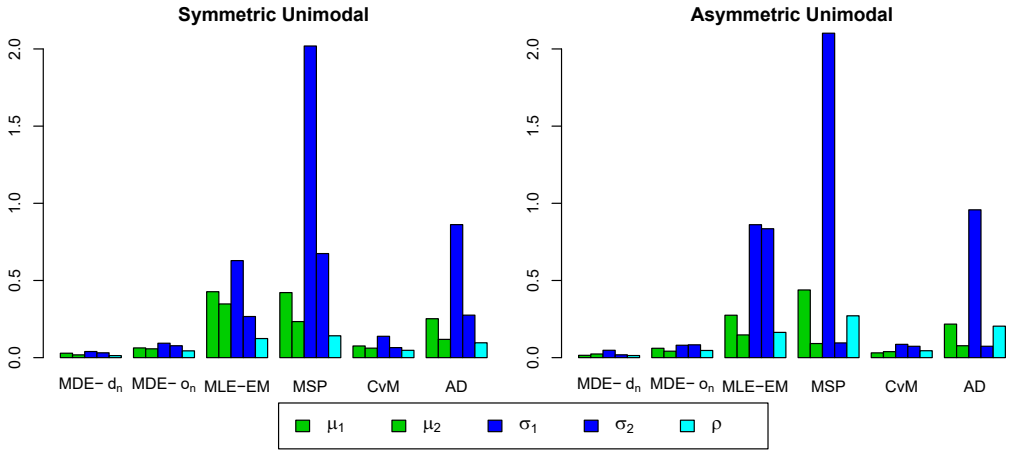

Figure 5.17: Barplot of the MSE of each parameter for the two unimodal mixture distributions, with t-distribution components with degrees of freedom $\nu = 5$. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples of size $n = 100$.

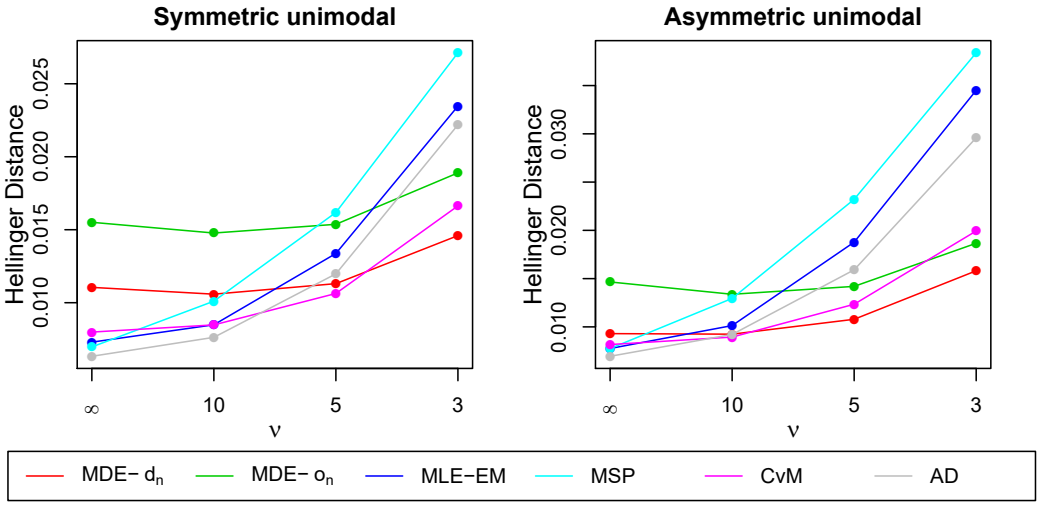

Figure 5.18: Means of the Hellinger distances between the true CDF and the estimated CDF, for the two unimodal mixture distributions, with t-distribution components with degrees of freedom ν, for different values of ν. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples of size $n = 100$.

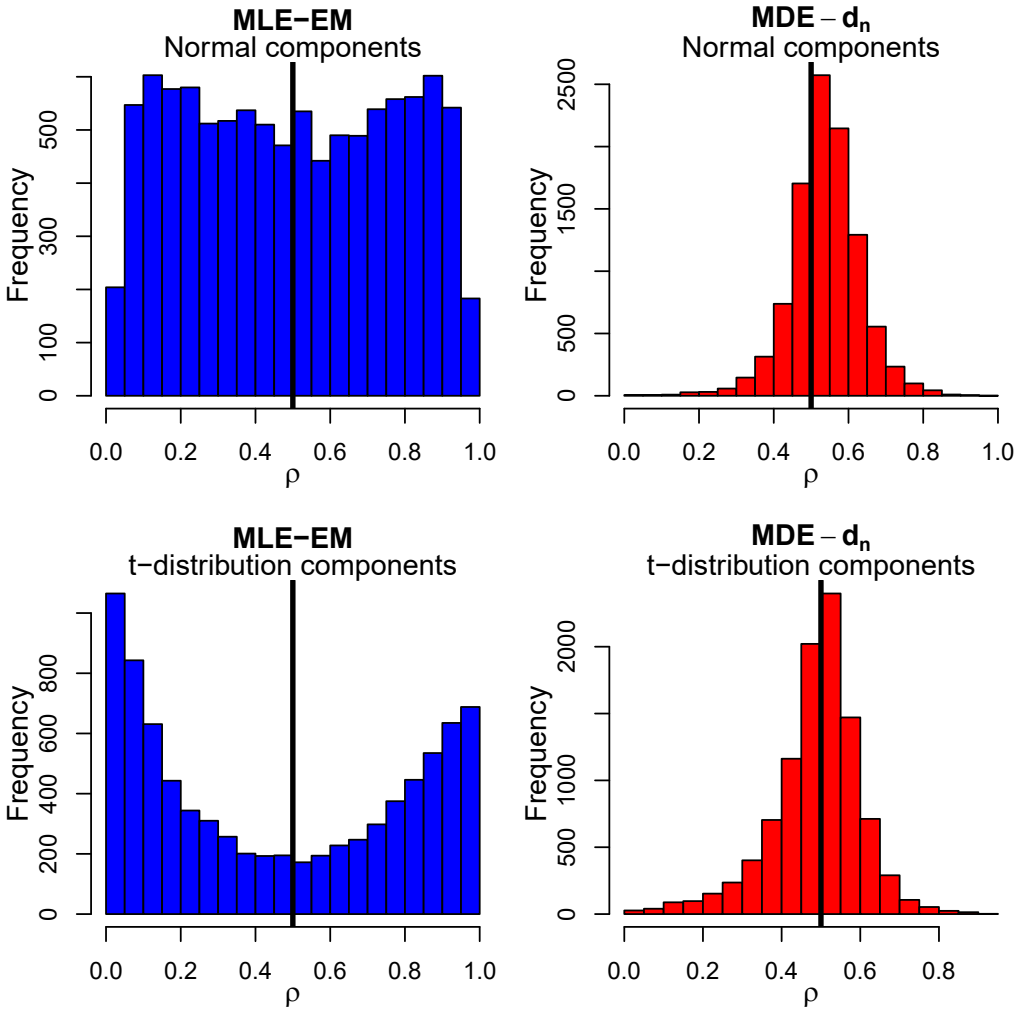

Figure 5.19: Histograms of the weight parameter ρ for the MLE-EM and MDE-d_n estimator for the symmetric unimodal mixture distribution, with either normal components or t-distribution components with $\nu = 3$. The true value of ρ is indicated with a line. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples of size $n = 100$.

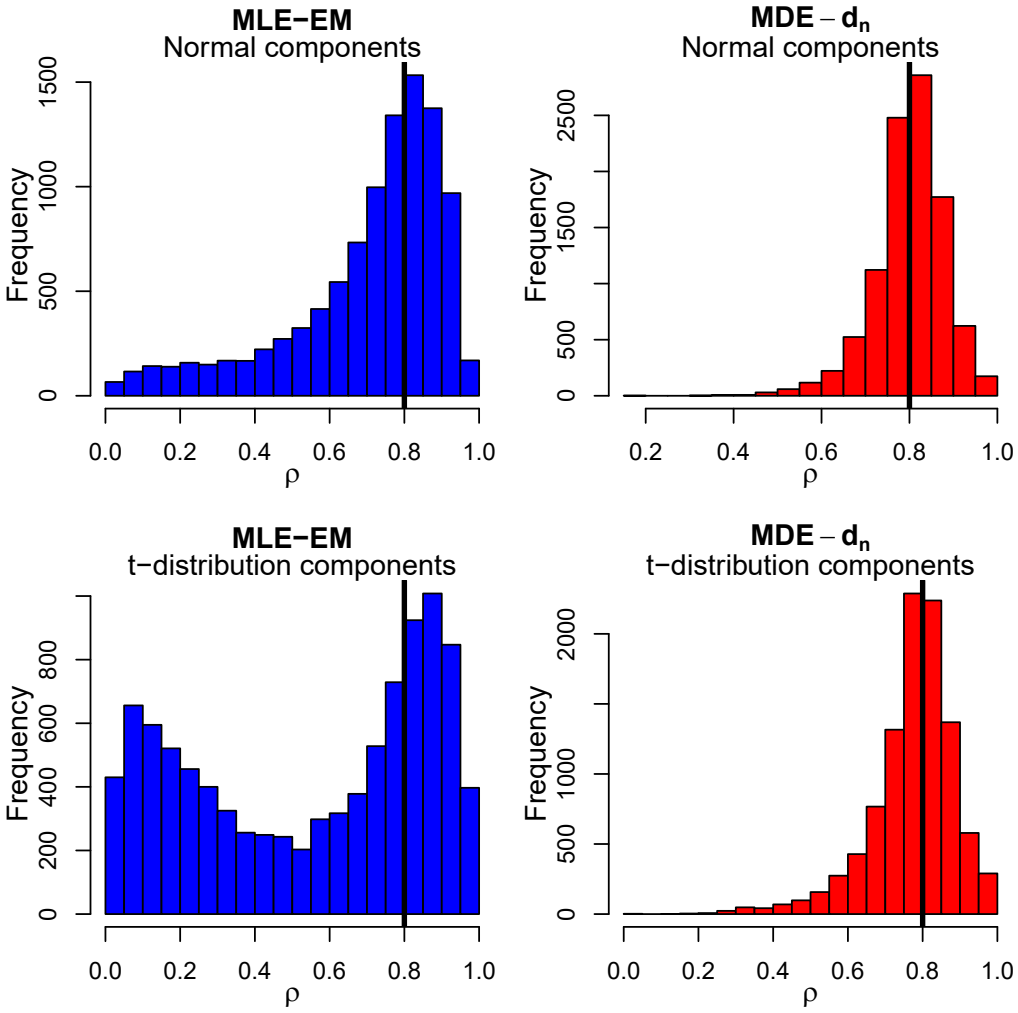

Figure 5.20: Histograms of the weight parameter ρ for the MLE-EM and MDE-d_n estimator in the asymmetric unimodal mixture distribution, with either normal components or t-distribution components with $\nu = 3$. The true value of ρ is indicated with a line. Estimators are intialized with the true parameters. Results are based on $N = 10^4$ samples of size $n = 100$.

Random Initialization

In Section 5.2.4, we observed that the MDE-d_n is quite robust in a normal mixture setting, when the actual components are t-distributed, when the method was initialized with the true parameters. However, in practice, we naturally must use an initialization method. We now study the robustness when the MDE-d_n is initialized with the K-means++ initialization. We compare against the MLE-EM estimator, initialized with the same method. We consider the symmetric unimodal distribution, as defined in Table 5.1.

To study the robustness, we study the distribution of the weight estimates. We compare the distribution when the components are either normally distributed or t-distributed, when the methods are initialized with K-means++ initialization. The histogram of the weights is given in Figure 5.21.

For the MDE-d_n, the weight parameter estimates are more spread out when using K-means++ initialization as opposed to initializing with true parameters. This can be seen when comparing Figure 5.19 to Figure 5.21. However, the distribution is again similar for either normal or t-distributed components.

For the MLE-EM, if components are normal, the distribution of the weight parameter estimates when methods are initialized with K-means++ initialization is very similar to the distribution when initialized with true parameters. This can be seen when comparing Figure 5.19 to Figure 5.21. However, for t-distribution components, the weights are estimated even more often near their extremes when K-means++ initialization is used as opposed to initialization with true parameters.

Conclusions. In conclusion, in this symmetric unimodal setting under K-means++ initialization, the MDE-d_n seems to retain its robustness properties. Conversely, the sensitivity of the MLE-EM estimator to the outliers is exacerbated when random initialization is used, and the weight parameter tends to be estimated at the extremes even more often.

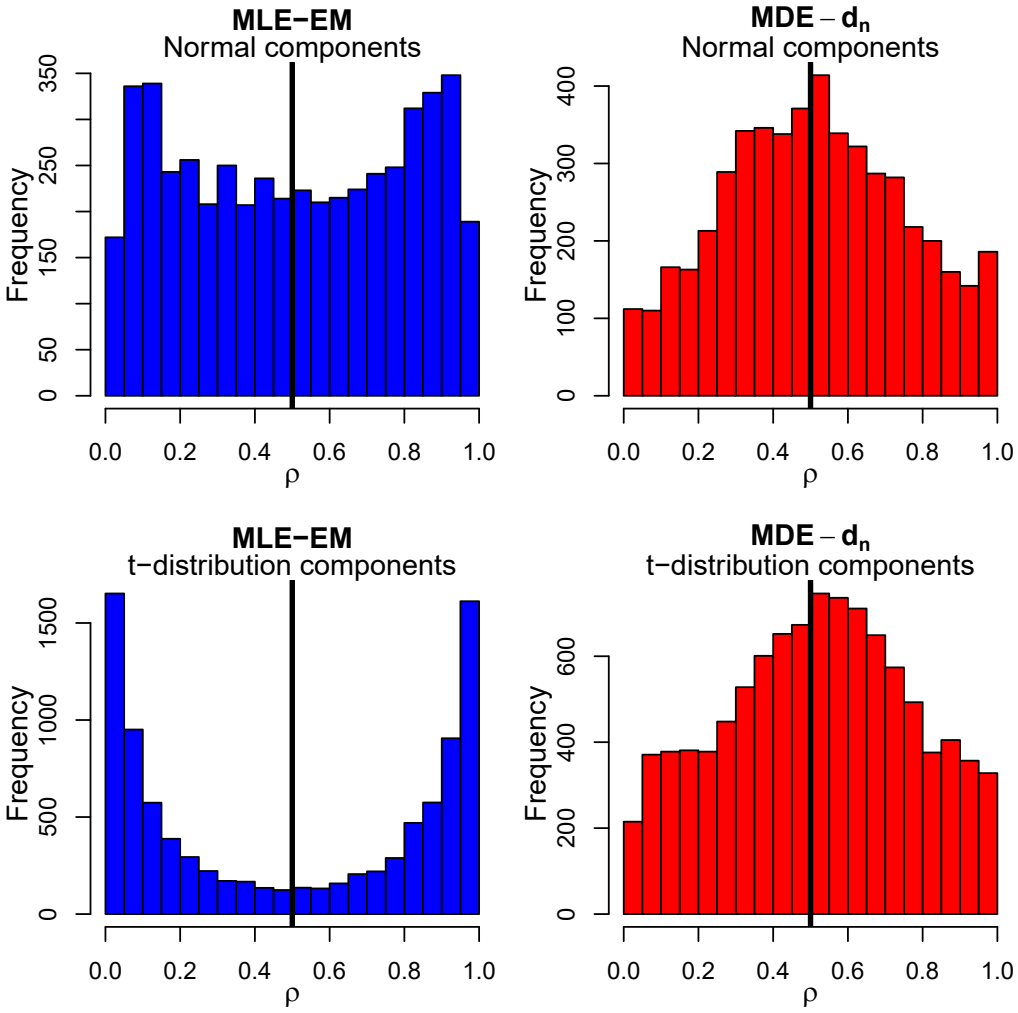

Figure 5.21: Histograms of the weight parameter ρ for the MLE-EM and MDE-d_n estimator in the symmetric unimodal distribution, with either normal components or t-distribution components with $\nu = 3$, when the methods are initialized with K-means++ initialization. The true value of ρ is indicated with a line. Results are based on $N = 10^4$ samples of size $n = 100$.

5.2.5 Application to Fisher's Iris Data

We now apply the MDE-d_n estimator in a practical application. Consider Fisher's Iris data, distributed with R [29]. The data contains measurements of sepal length, sepal width, petal length and petal width for 50 flowers from three different species. We restrict ourselves to the data of the two most similar species; *Versicolor* and *Virginica*. We consider the petal length of the two species.

We model this data as a dual normal mixture without using knowledge of the labels. We cannot apply the MDE-d_n estimator directly, as the data is rounded and contains ties. Therefore, we first add some noise to each datapoint, normally distributed with mean 0 and standard deviation 0.02. This ensures the data contains no more ties. The low standard deviation ensures that the probability of changing the order of two datapoints is small.

We compute the normal mixture parameters using the MDE-d_n and MLE-EM estimator, initializing 20 times with K-means++ initialization and choosing the parameter values with the lowest distance and highest likelihood respectively.

To compare the results, we will compute the normal mixture parameters using the data labels. Then, we can then simply estimate the component parameters by using the familiar MLE of a normal distribution on the separate groups.

The mixture densities with plugged-in estimates are plotted in Figure 5.22. The components of each of these estimated densities, along with the labeled data, is plotted in Figure 5.23. As we can see, the MDE-d_n and MLE-EM perform similarly, as expected in this setting based on our simulation results in Section 5.2.3. The MLE-EM places a higher weight on the first component. Both are reasonably close to the normal mixture distribution which was estimated using the labels.

Data labeling. Now, we can use these mixture densities to compute the *expected labels* of each datapoint. We can compare the probabilities of a datapoint belonging to each component, and select the label corresponding to the highest probability. Using this labeling method, the mixture from the MDE-d_n mislabels 13 datapoints, while the MLE-EM mislabels 18 datapoints. Using the same procedure for our benchmark distribution, the mixture which was computed *with label knowledge*, results in 7 misla-

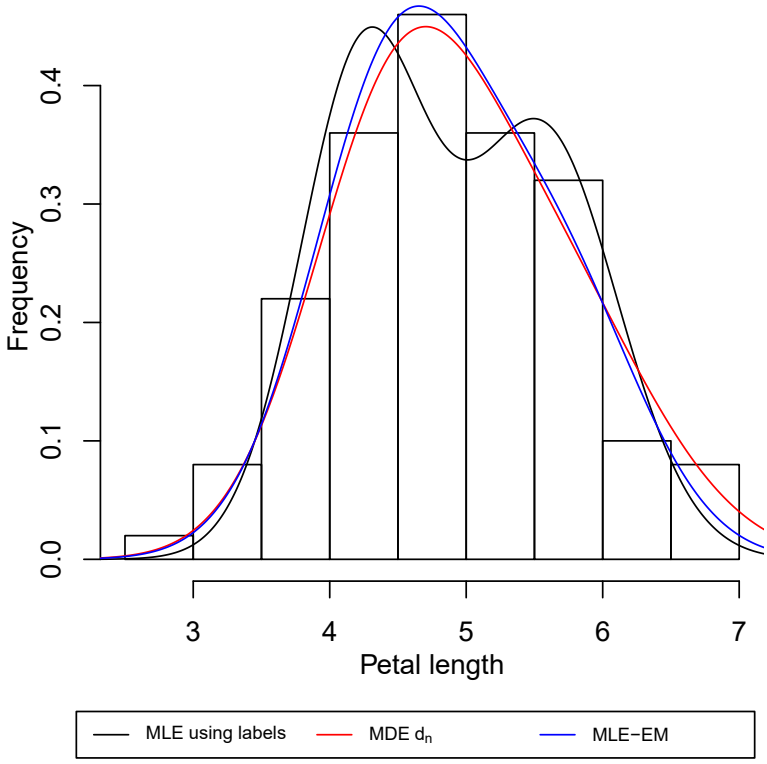

Figure 5.22: Dual normal mixture distributions estimated using the MDE-d_n, MLE-EM and MLE using the labels of the data, for the petal length of Fisher's Iris data, along with the histogram of the data. Only the the species *Versicolor* and *Virginica* are included.

beled datapoints. Therefore, the MDE-d_n performs quite well in this setting, outperforming the MLE-EM by a small margin.

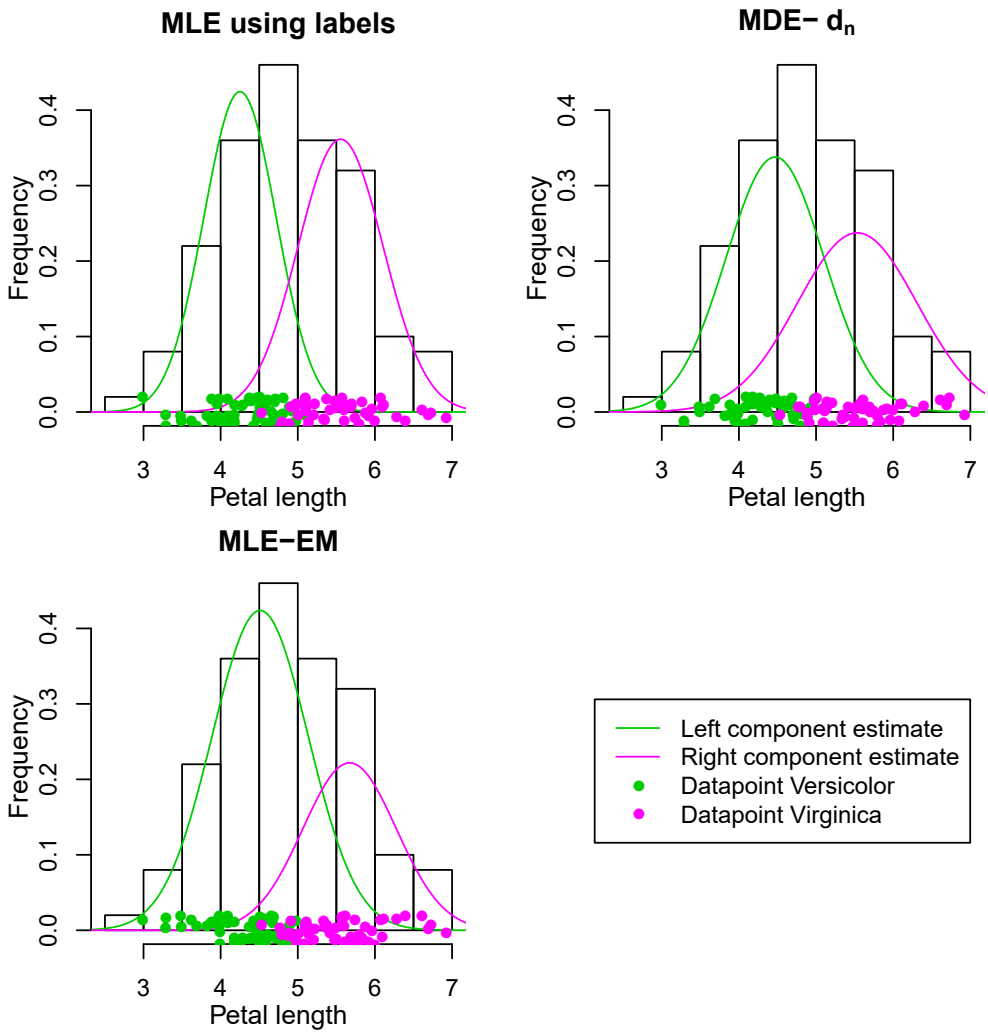

Figure 5.23: Dual normal mixture distribution components estimated using the MDE-d_n, MLE-EM and MLE using the labels of the data, for the petal length of Fisher's Iris data, along with the histogram of the data. Only the species *Versicolor* and *Virginica* are included.

Chapter 6

Conclusion

We have used the d_n and o_n statistics in two statistical inference settings; goodness-of-fit hypothesis testing and parameter estimation.

Goodness of fit. For goodness of fit, the d_n and o_n statistic perform similarly. Compared to the Anderson-Darling statistic, they are more powerful statistics for rejecting beta mixtures when the null hypothesis is uniform, both when parameters are specified a priori and when they are estimated. They also perform well in other settings where parameters are specified a priori, but are outperformed when the parameters are estimated from the data.

Therefore, the d_n and o_n statistic are powerful statistics and preferred over the Anderson-Darling statistic in the goodness-of-fit setting with a uniform null, where the alternative is suspected to have clustering.

Parameter estimation. For the parameter estimation, the MDE-d_n and MDE-o_n statistic have different performance. The MDE-o_n estimator is less sensitive to initial parameter specification. However, it is generally outperformed by the MDE-d_n estimator. When all are initialized with true parameters, the MDE-d_n estimator performs considerably better than the MLE-EM estimator on 2-component normal mixtures when the components have significant overlap. If initialized once with the K-means++ method, the MDE-d_n method performs similarly to the MLE-EM. Additionally, the MDE-d_n estimator has excellent robustness properties in case the components have heavier tails. The MLE-EM algorithm is considerably less robust in this setting.

Therefore, the MDE-d_n statistic is a powerful estimator for dual normal

mixtures where the components have large overlap. It could be a viable alternative to MLE-EM estimation in this setting. Furthermore, the MDE-d_n could be preferred especially in case of data with suspected outliers, considering it is significantly more robust than the MLE-EM.

Bibliography

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *Annals of Mathematical Statisticss*, 23(2):193–212, 1952.

[2] David Arthur and Sergei Vassilvitskii. K-Means++: the Advantages of Careful Seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 8:1027–1025, 2007.

[3] Tatiana Benaglia, Didier Chauveau, David R. Hunter, and Derek S. Young. mixtools: An R package for analyzing mixture models. *Journal of Statistical Software*, 32(1):1–29, 2009.

[4] Enrico Bibbona, Giovanni Pistone, and Mauro Gasparini. The Empirical Identity Process: asymptotics and applications. *The Canadian Journal of Statistics*, 2017.

[5] Johannes Blömer and Kathrin Bujna. *Simple Methods for Initializing the EM Algorithm for Gaussian Mixture Models.* dec 2013.

[6] Andrei N. Borodin and Paavo Salminen. *Handbook of Brownian motion - Facts and Formulae.* Springer, second edition, 2002.

[7] R. C. H. Cheng and N. A. K. Amin. Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45(3):394–403, 1983.

[8] R. C. H. Cheng and M. A. Stephens. A Goodness-of-Fit Test Using Moran Statistic with Estimated Parameters. *Biometrika*, 76(2):385–392, 1989.

[9] Harald Cramér. On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):141–180, 1928.

[10] Harald Cramér. *Mathematical Methods of Statistics*, volume 42. Princeton University Press, 1946.

[11] A. P. Dempster, N. M. Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38, 1977.

[12] Monroe D. Donsker. An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society*, 6, 1951.

[13] Monroe D. Donsker. Justification and Extension of Doob's Heuristic Approach to the Kolmogorov- Smirnov Theorems. *Ann. Math. Statist.*, 23(2):277–281, 1952.

[14] Magnus Ekström. Alternatives to maximum likelihood estimation based on spacings and the Kullback-Leibler divergence. *Journal of Statistical Planning and Inference*, 138(6):1778–1791, 2008.

[15] Bernard Flury. *A First Course in Multivariate Statistics*. Springer Texts in Statistics. Springer, 1997.

[16] Kaushik Ghosh and S. Rao Jammalamadaka. A general estimation method using spacings. *Journal of Statistical Planning and Inference*, 93(1-2):71–82, 2001.

[17] Major Greenwood. The Statistical Study of Infectious Diseases. *Journal of the Royal Statistical Society*, 109(2):85, 1946.

[18] S. Rao Jammalamadaka and M. N. Goria. A test of goodness-of-fit based on Gini's index of spacings. *Statistics and Probability Letters*, 68(2):177–187, 2004.

[19] George G. Judge, W. E. Griffiths, R. Carter Hill, Helmut Lütkepohl, and Tsoung-chao Lee. *The Theory and practice of econometrics*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, second edition, 1985.

[20] Bradford F. Kimball. On the Asymptotic Distribution of the Sum of Powers of Unit Frequency Differences. *The Annals of Mathematical Statistics*, 21(2):263–271, 1950.

[21] A. N. Kolmogorov. Sulla Determinazione Empirica di una Legge di Distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

[22] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[23] P. A. P. Moran. The Random Division of an Interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1):92, 1947.

[24] Kenneth Nordström. The Concentration Ellipsoid of a Random Vector Revisited. *Econometric Theory*, 7(3):397–403, 1991.

[25] William C. Parr and William R. Schucany. Minimum Distance and Robust Estimation. *Journal of the American Statistical Association*, 75(371):616–624, 1980.

[26] Karl Pearson. Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 185:71–110, 1894.

[27] B. Charles Jr. Peters and Homer F. Walker. An Iterative Procedure for Obtaining Maximum-Likelihood Estimates of the Parameters for a Mixture of Normal Distributions. *SIAM Journal on Applied Mathematics*, 35(2):362–378, 1978.

[28] R. Pyke. Spacings. *Journal of the Royal Statistical Society. Series B (Methodological)*, 27(3):395–449, 1965.

[29] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[30] M. Mahibbur Rahman and Z. Govindarajulu. A modification of the test of Shapiro and Wilk for normality. *Journal of Applied Statistics*, 24(2):219–236, 1997.

[31] Bo Ranneby. The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method. *Scandinavian Journal of Statistics*, 11(2):pp. 93–112, 1984.

[32] Bo Ranneby, S. Rao Jammalamadaka, and Alex Teterukovskiy. The maximum spacing estimation for multivariate observations. *Journal of Statistical Planning and Inference*, 129(1-2 SPEC. ISS.):427–446, 2005.

[33] J. S. Rao and Morgan Kuo. Asymptotic results on the Greenwood Statistic and some of its generalizations. *Journal of the Royal Statistical Society, Series B*, 46(2):228–237, 1984.

[34] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[35] Richard A. Redner and Homer F. Walker. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 26(2):195–239, 1984.

[36] C. A. Robertson and J. G. Fryer. Some descriptive properties of normal mixtures. *Scandinavian Actuarial Journal*, 1969(3-4):137–146, 1969.

[37] Patrick Royston. Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2(3):117–119, 1992.

[38] Yongzhao Shao and Marjorie G. Hahn. Strong Consistency of the Maximum Product of Spacings Estimates with Applications in Nonparametrics and in estimation of unimodal densities. *Annals of Statistics*, 51(1):31–49, 1999.

[39] S. S. Shapiro and M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3/4):591, 1965.

[40] M. A. Stephens. Introduction to Kolmogorov (1933) On the Empirical Determination of a Distribution. In *Breakthroughs in Statistics, Springer Series in Statistics*, pages 93–105. Springer, New York, jan 1992.

[41] Richard Von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit.* Springer, 1928.

[42] J. Wolfowitz. The Minimum Distance Method. *Annals of Mathematical Statisticss*, 28(1):75–88, 1957.

Appendix A

Mathematical Appendix

A.1 Derivations of Expressions for Integrated Empirical Identity Process

In this appendix, we derive simple expressions for the integrated empirical identity process, for the upper, lower and central version. We show that, for the d_n statistic, the choice of IEIP is not relevant, as all three produce the same statistic.

In Section 2.1, several definitions were already presented. To avoid confusion with the main text, where I_n is used for the lower IEIP, we use $I_n^L(t)$ for the lower IEIP and I_n^C for the central IEIP and other corresponding functions. This notation is only used in this appendix.

Let $u_1, \ldots, u_n$ be ordered uniform random variables, and define $u_0 = 0$ and $u_{n+1} = 1$. Denote $F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(u_i \leq t)$, the ECDF. Denote $Q_n(u) = \inf\{t \in [0,1] : F_n(t) \geq u\}$ the empirical quantile function. The empirical identity processes are given by:

$$
\begin{aligned}
R_n^L(t) &:= \begin{cases} 0 & \text{if } 0 \leq t < u_1 \\ Q_n(F_n(t)) & \text{if } u_1 \leq t \leq 1 \end{cases}, \\
R_n^U(t) &:= \begin{cases} Q_n(F_n(t) + \frac{1}{n}) & \text{if } 0 \leq t < u_n \\ 1 & \text{if } u_n \leq t \leq 1 \end{cases}, \\
R_n^C(t) &:= \frac{R_n^L(t) + R_n^U(t)}{2}.
\end{aligned}
$$

We can simplify these expressions by noting that the values of these functions is constant between consecutive datapoints. Note that, for $t \in [u_{i-1}, u_i)$:

$$Q_n(F_n(t)) = Q_n(\frac{i-1}{n}) = u_{i-1},$$
$$Q_n(F_n(t) + \frac{1}{n}) = Q_n(\frac{i}{n}) = u_i,$$

leading to simple expressions for R_n^L and R_n^U which we use later. The empirical identity processes for $t \in [0,1]$ are given by:

$$Y_n^L(t) := (n+1)(R_n^L(t) - t),$$
$$Y_n^U(t) := (n+1)(R_n^U(t) - t),$$
$$Y_n^C(t) := (n+1)(R_n^C(t) - t) = \tfrac{1}{2}(Y_n^L(t) + Y_n^U(t)).$$

The IEIP's for $t \in [0,1]$ are then defined as:

$$I_n^L(t) := -\int_0^t Y_n^L(u)du = (n+1)\int_0^t \big(u - R_n^L(u)\big)du,$$
$$I_n^U(t) := \int_0^t Y_n^U(u)du = (n+1)\int_0^t \big(R_n^U(u) - u\big)du,$$
$$I_n^C(t) := \int_0^t \frac{1}{2}(Y_n^L(u) + Y_n^U(u))du = \frac{1}{2}(I_n^L(t) + I_n^U(t)).$$

We can now simplify these expressions. Let $t \in [u_{i-1}, u_i]$. Define the sum from 1 to 0 as the empty sum, equal to 0. Then, for I_n^L, we have:

$$\begin{aligned} I_n^L(t) &= (n+1)\int_0^t \big(u - R_n^L(u)\big)du \\ &= (n+1)\sum_{j=1}^{i-1}\int_{u_{j-1}}^{u_j}(u - u_{j-1})du + (n+1)\int_{u_{i-1}}^{t}(u - u_{i-1})du \\ &= \frac{n+1}{2}\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 + \frac{n+1}{2}(t - u_{i-1})^2. \end{aligned}$$

For I_n^U, we have:

$$\begin{aligned} I_n^U(t) &= (n+1)\int_0^t \big(R_n^U(u) - u\big)du \\ &= (n+1)\sum_{j=1}^{i-1}\int_{u_{j-1}}^{u_j}(u_j - u)du + (n+1)\int_{u_{i-1}}^{t}(u_i - u)du \\ &= \frac{n+1}{2}\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 - \frac{n+1}{2}(u_i - t)^2 + \frac{n+1}{2}(u_i - u_{i-1})^2. \end{aligned}$$

For $I_n^C(t)$, we have:

$$\begin{aligned}
I_n^C(t) &= \frac{1}{2}(I_n^L(t) + I_n^U(t)) \\
&= \frac{n+1}{4}\Bigg(2\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 + (t - u_{i-1})^2 - (u_i - t)^2 + (u_i - u_{i-1})^2\Bigg) \\
&= \frac{n+1}{4}\Bigg(2\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 + t^2 - 2u_{i-1}t + u_{i-1}^2 - u_i^2 + 2u_i t - t^2 \\
&\qquad\qquad + u_i^2 - 2u_i u_{i-1} + u_{i-1}^2\Bigg) \\
&= \frac{n+1}{4}\Bigg(2\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 + 2(u_i - u_{i-1})t + 2u_{i-1}^2 - 2u_i u_{i-1}\Bigg) \\
&= \frac{n+1}{2}\Bigg(\sum_{j=1}^{i-1}(u_j - u_{j-1})^2 + (u_i - u_{i-1})t - u_{i-1}(u_i - u_{i-1})\Bigg).
\end{aligned}$$

A plot of the three IEIP's, together with the ECDF is given in Figure A.1.

Note that all three IEIP's are increasing between datapoints (u_{i-1}, u_i), while the ECDF is constant. Therefore, to compute the supremum:

$$\sup_{t\in[0,1]} |I_n(t) - F_n(t)|,$$

it is sufficient to consider the maximum difference at- and just before each discontinuity. Since the values of the IEIP's around the discontinuities are the same, this supremum is also the same, irrespective of the choice of lower, upper or central IEIP. Therefore, the d_n statistic, which uses this supremum, is equal for any of the three IEIP's.

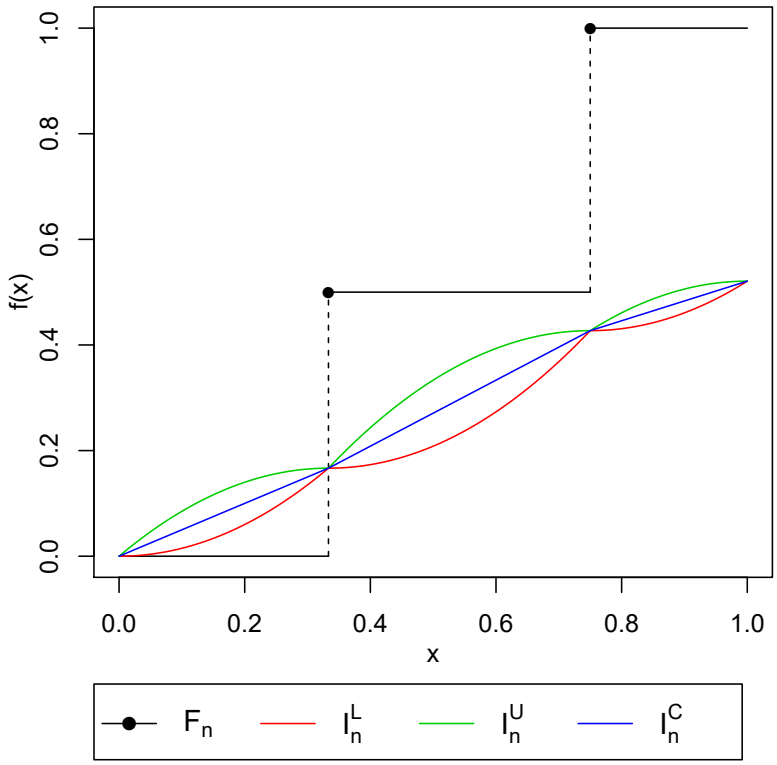

Figure A.1: Plots of the three versions of the IEIP, along with the ECDF, for data $u = (\frac{1}{3}, \frac{3}{4})$.

A.2 Volume of the Concentration Ellipsoid

In this appendix, we show that the volume of the ellipsoid given by:

$$\mathcal{E}_1 = \{t \in \mathbb{R}^n : (t - \mathbb{E}[\hat{\theta}])^T \Sigma^{-1} (t - \mathbb{E}[\hat{\theta}]) \leq 1\},$$

is equal to:

$$\text{Vol}(\mathcal{E}_1) = \frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \prod_{i=1}^{n} \sqrt{\lambda_i}, \tag{A.1}$$

where λ_i are the eigenvalues of Σ. We assume Σ^{-1} to be non-singular and symmetric[†]. Since Σ is symmetric, so is its inverse, and both are therefore diagonalizable. Let $\Sigma^{-1} = S^T D S$ be its diagonalization. A well known linear algebra result gives λ_i^{-1} are the eigenvalues of Σ^{-1}. Now, we have:

$$\begin{aligned}
&(t - \mathbb{E}[\hat{\theta}])^T \Sigma^{-1} (t - \mathbb{E}[\hat{\theta}]) \leq 1, \\
\Rightarrow &(S(t - \mathbb{E}[\hat{\theta}]))^T D (S(t - \mathbb{E}[\hat{\theta}])) \leq 1, \\
\Rightarrow &u^T D u \leq 1, \\
\Rightarrow &\sum_{i=1}^{n} \frac{u_i^2}{\sqrt{\lambda_i}^2} \leq 1.
\end{aligned}$$

Therefore, this ellipse has axis lengths equal to $\sqrt{\lambda_i}$ along its principal axis - which are given by the eigenvectors. Since the volume of an ellipsoid is given by $\frac{\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2}+1)} \prod_{i=1}^{n} a_i$, with a_i the lengths of its principal axis, we obtain the result (A.1).

A.3 Order Analysis o_n

In Section 3.1, we demonstrate that o_n can be expressed based solely on spacings, given by:

$$\begin{aligned}
o_n = 4(n+1) \sum_{i=1}^{n+1} S_{i-1}^2 t_i - \frac{2(i-1)}{n} S_{i-1} t_i + \frac{(i-1)^2}{n^2} t_i \\
+ \frac{n+1}{3} S_{i-1} t_i^3 - \frac{n+1}{n} \frac{i-1}{3} t_i^3 + \frac{(n+1)^2}{20} t_i^5.
\end{aligned}$$

[†]In the context of concentration elipsoids, Σ is the sample covariance matrix or the MSE matrix, both of which are symmetric.

We note that these terms are of the following order:

$$\sum_{i=1}^{n+1} S_{i-1}^2 t_i - \frac{2(i-1)}{n} S_{i-1} t_i + \frac{(i-1)^2}{n^2} t_i \sim O(1),$$

$$\sum_{i=1}^{n+1} \frac{n+1}{3} S_{i-1} t_i^3 - \frac{n+1}{n} \frac{i-1}{3} t_i^3 \sim O(\frac{1}{n}),$$

$$\sum_{i=1}^{n+1} \frac{(n+1)^2}{20} t_i^5 \sim O(\frac{1}{n^2}).$$

These were obtained in the following manner. First off, we observe that $t_i \sim O(\frac{1}{n})$. Next, note that:

$$S_i = \frac{n+1}{2} \sum_{j=1}^{i} t_j^2 \sim O(n \frac{i}{n^2}) = O(\frac{i}{n})$$

We can then analyze each of the terms, obtaining the orders:

$$\sum_{i=1}^{n+1} S_{i-1}^2 t_i \sim O(\sum_{i=1}^{n+1} \frac{(i-1)^2}{n^2} \frac{1}{n}) = O(\frac{1}{n^3} \sum_{i=0}^{n} i^2) = O(\frac{1}{n^3} \frac{1}{6} n(n+1)(2n+1)) = O(1),$$

$$\sum_{i=1}^{n+1} \frac{2(i-1)}{n} S_{i-1} t_i \sim O(\sum_{i=1}^{n+1} \frac{2(i-1)}{n} \frac{i-1}{n} \frac{1}{n}) = O(\frac{2}{n^3} (\sum_{i=0}^{n} i^2) = O(\frac{2}{n^3} (\frac{1}{6} n(n+1)(2n+1)) = O(1),$$

$$\sum_{i=1}^{n+1} \frac{(i-1)^2}{n^2} t_i \sim O(\sum_{i=1}^{n+1} \frac{(i-1)^2}{n^2} \frac{1}{n}) = O(\frac{1}{n^3} (\sum_{i=0}^{n} i^2) = O(\frac{1}{n^3} (\frac{1}{6} n(n+1)(2n+1)) = O(1),$$

$$\sum_{i=1}^{n+1} \frac{n+1}{3} S_{i-1} t_i^3 \sim O(\sum_{i=1}^{n+1} \frac{n+1}{3} \frac{i-1}{n} \frac{1}{n^3}) = O(\frac{n+1}{3n^4} \sum_{i=0}^{n} i) = O(\frac{n+1}{3n^4} \frac{1}{2} n(n+1)) = O(\frac{1}{n}),$$

$$\sum_{i=1}^{n+1} \frac{n+1}{n}\frac{i-1}{3}t_i^3 \sim O(\sum_{i=1}^{n+1} \frac{n+1}{n}\frac{i-1}{3}\frac{1}{n^3}) = O(\frac{n+1}{3n^4}\sum_{i=0}^{n} i)$$
$$= O(\frac{n+1}{3n^4}\frac{1}{2}n(n+1)) = O(\frac{1}{n}),$$

$$\sum_{i=1}^{n+1} \frac{(n+1)^2}{20}t_i^5 \sim O(\sum_{i=1}^{n+1} \frac{(n+1)^2}{20}\frac{1}{n^5}) = O(\frac{(n+1)^2}{20n^5}\sum_{i=0}^{n} 1)$$
$$= O(\frac{n(n+1)^2}{20n^5}) = O(\frac{1}{n^2}).$$

Appendix B

Simulation Results

In this appendix, we give simulation results that were not given in the main text.

B.1 Dual Normal Mixture Distribution

B.1.1 Initialization Sensitivity

Deterministic Pertubations

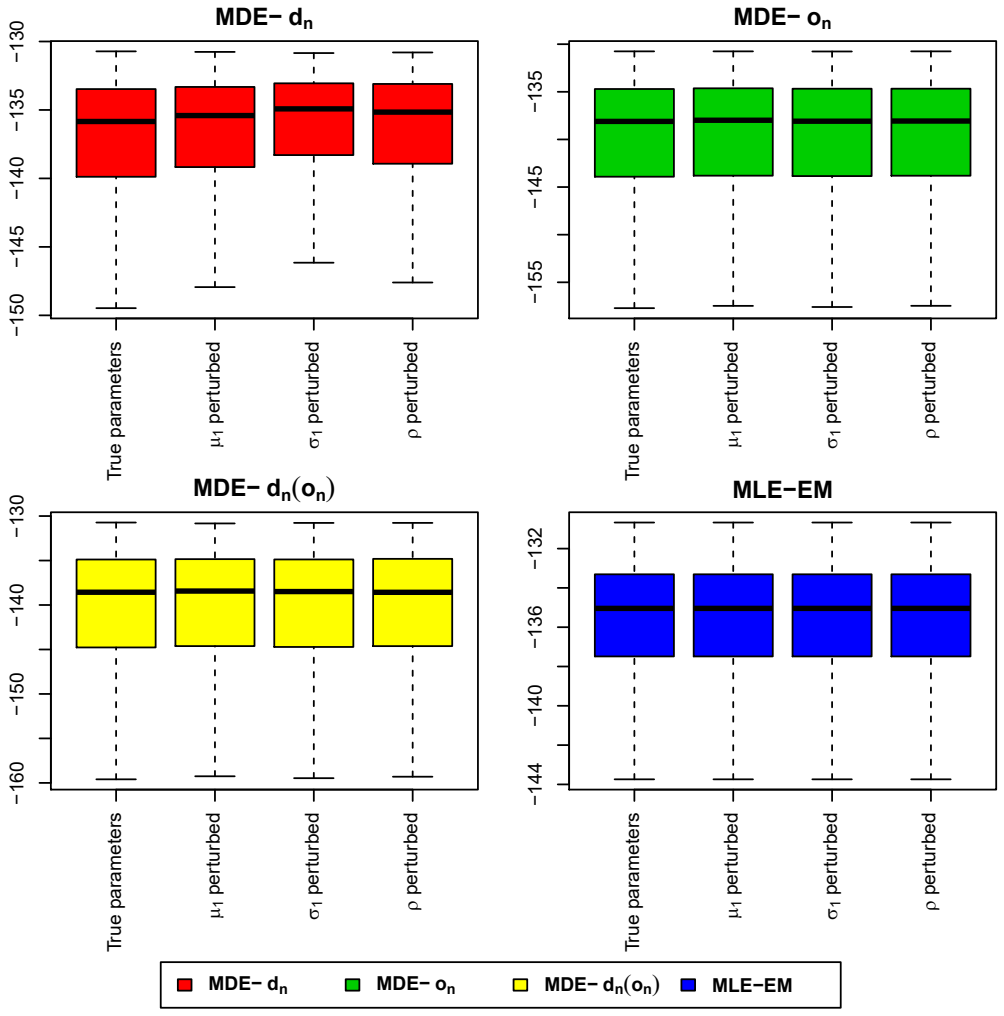

Figure B.1: Boxplots of the attained likelihoods of the estimators using different deterministic intialization pertubations, in the symmetric bimodal normal mixture model. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

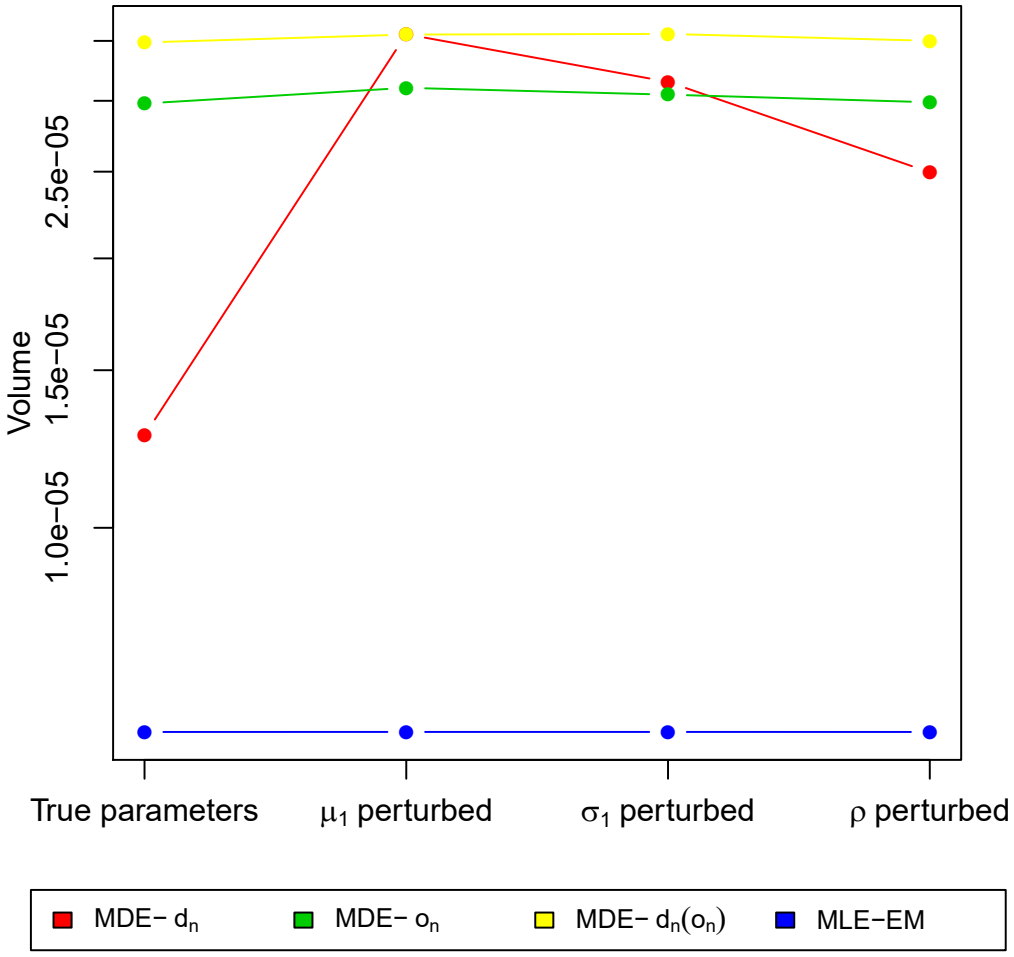

Figure B.2: Volumes of the MSE concentration elipsoids of the estimators using different deterministic intialization pertubations, in the symmetric bimodal normal mixture model. Note that the y-axis is in logarithmic scale. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

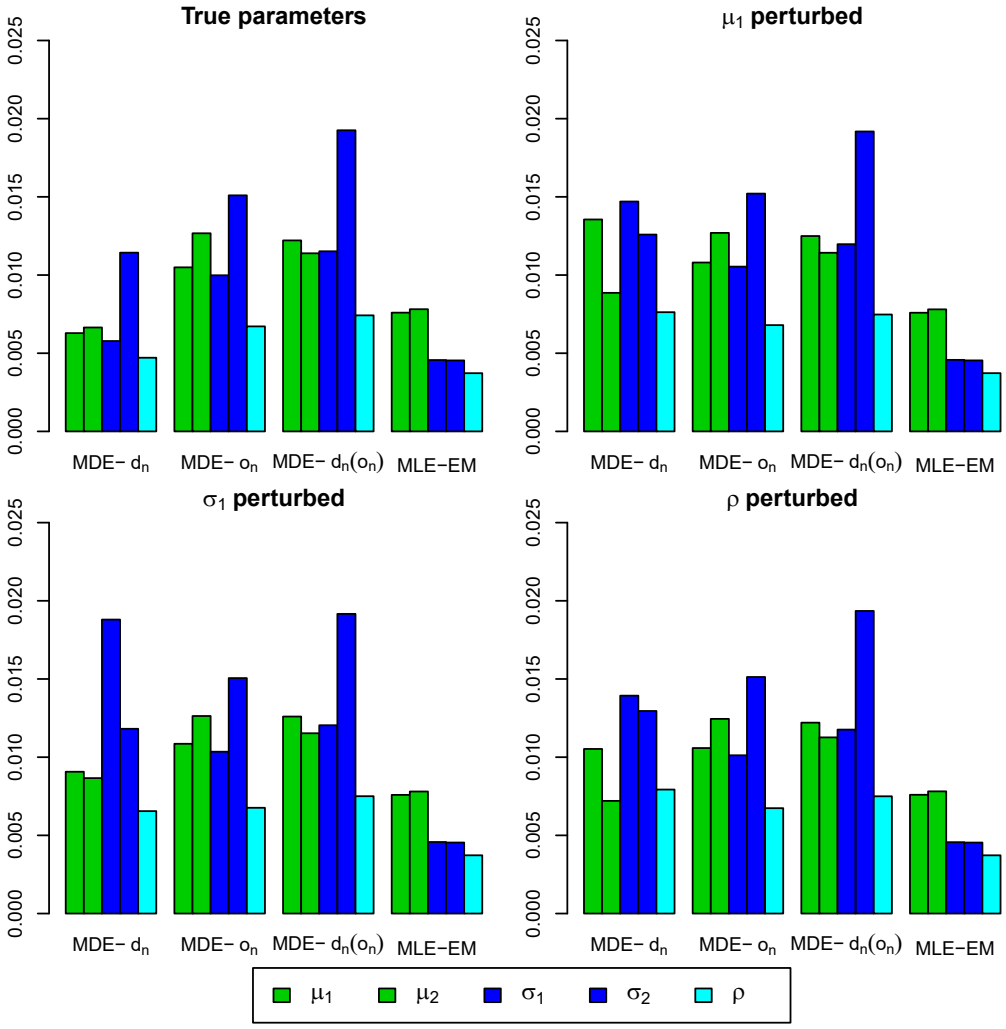

Figure B.3: Barplots of the MSE of the estimators using different deterministic intialization pertubations, in the symmetric bimodal normal mixture model. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

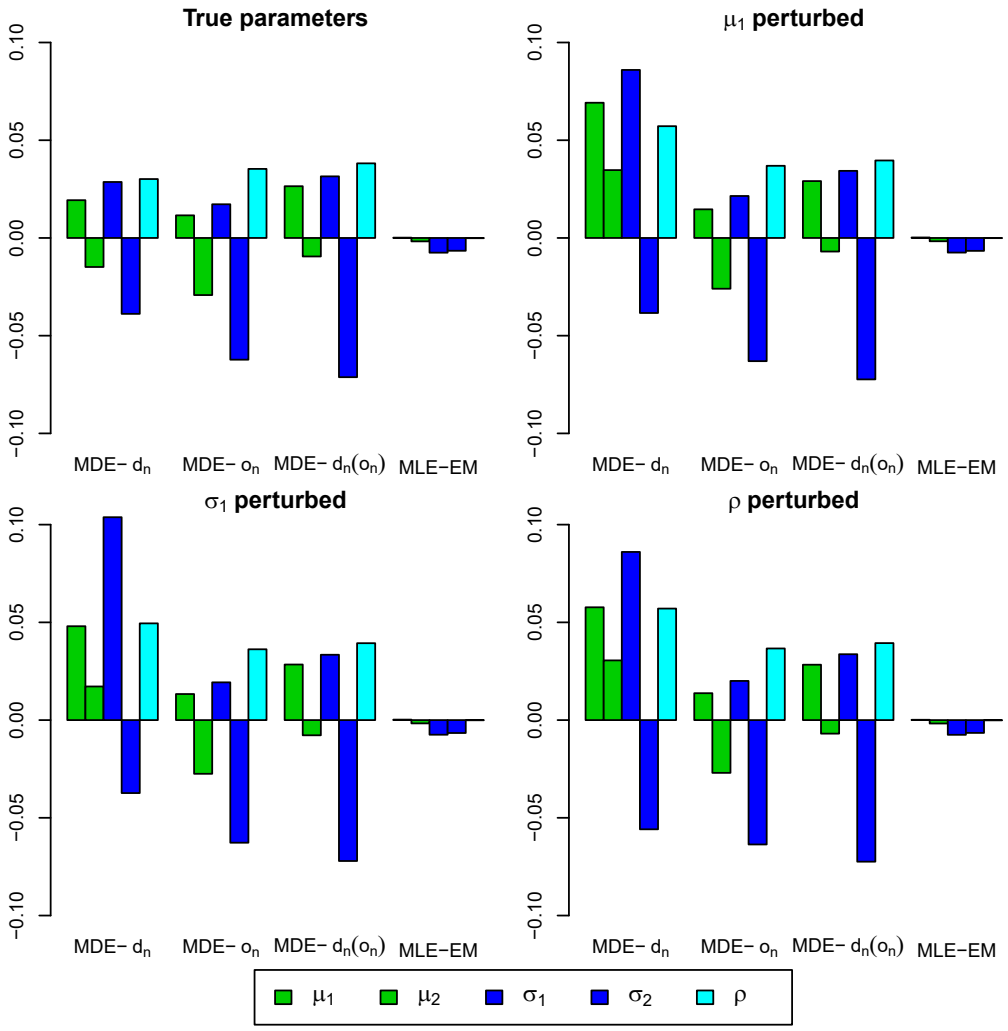

Figure B.4: Barplots of the bias of the estimators using different deterministic intialization pertubations, in the symmetric bimodal normal mixture model. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

Random Initializations

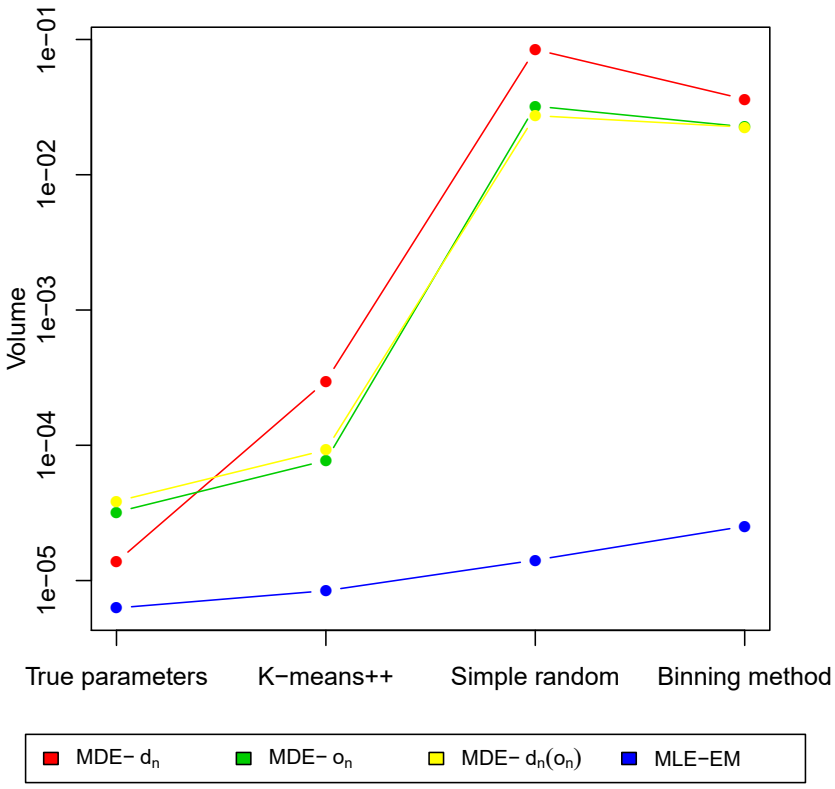

Figure B.5: Volumes of the MSE concentration elipsoids of the estimators using different random initialization schemes, for the symmetric bimodal normal mixture model. Note that the y-axis is in logarithmic scale. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

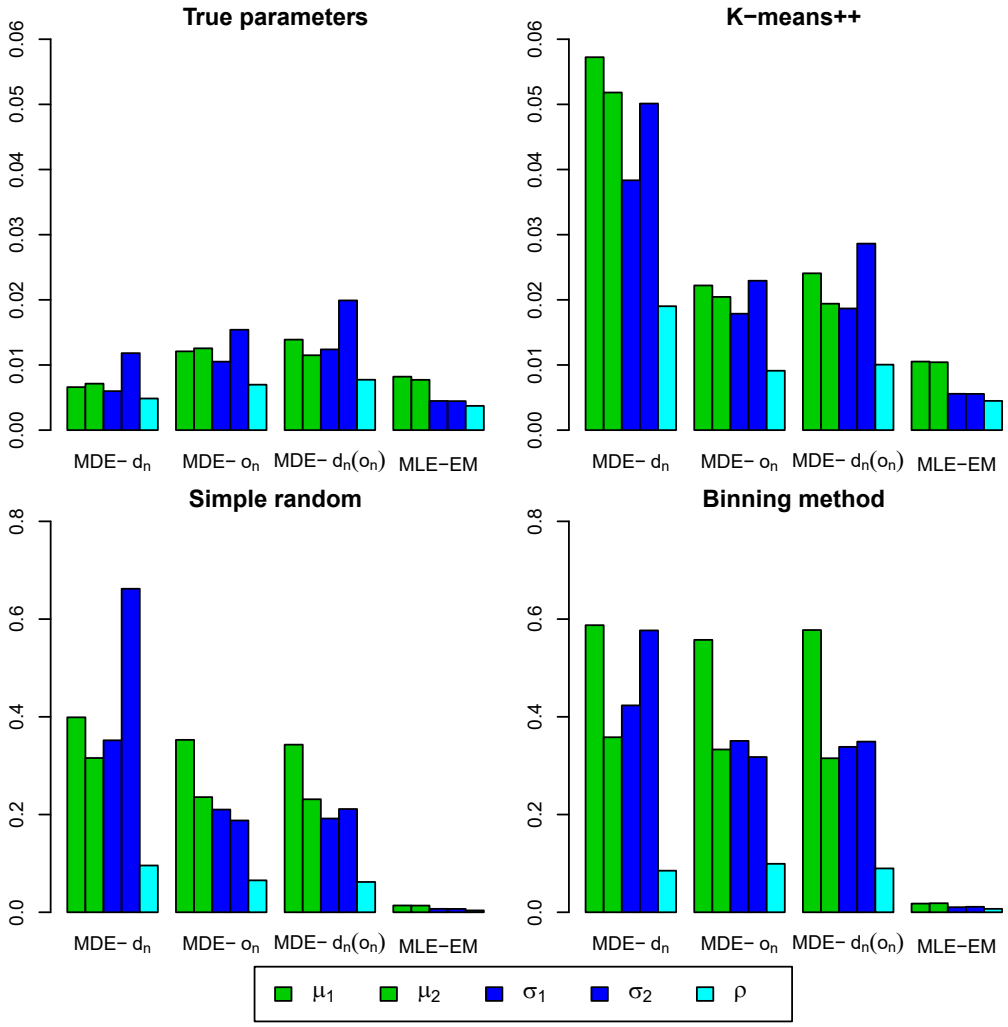

Figure B.6: Barplots of the MSE of the estimators using different random initialization schemes, for the symmetric bimodal normal mixture distribution. Note that the y-axis is of different scale for the first two and the second two barplots. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

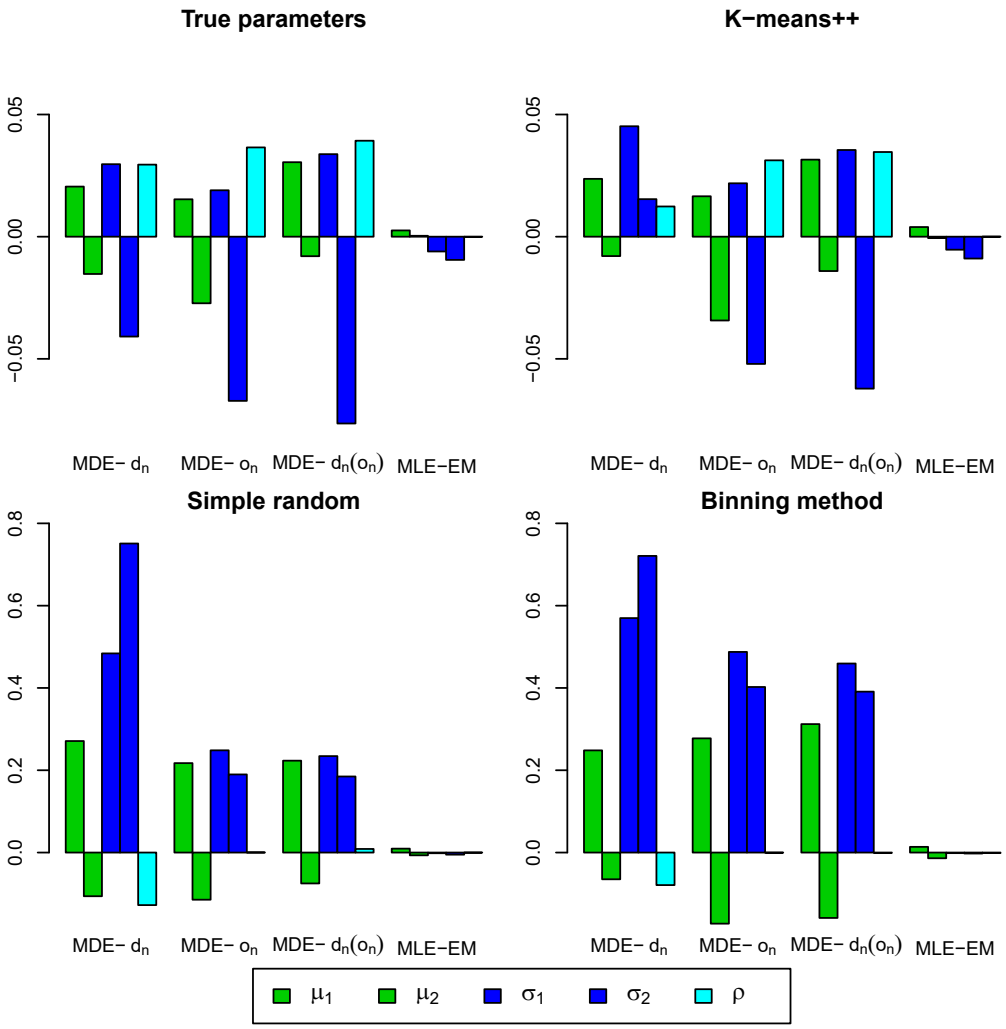

Figure B.7: Barplots of the bias of the estimators using different random initialization schemes, for the symmetric bimodal normal mixture distribution. Note that the y-axis is of different scale for the first two and the second two barplots. Results are based on $N = 5 \cdot 10^3$ samples of size $n = 100$.

B.1.2 Performance Study

Initialization with True Parameters

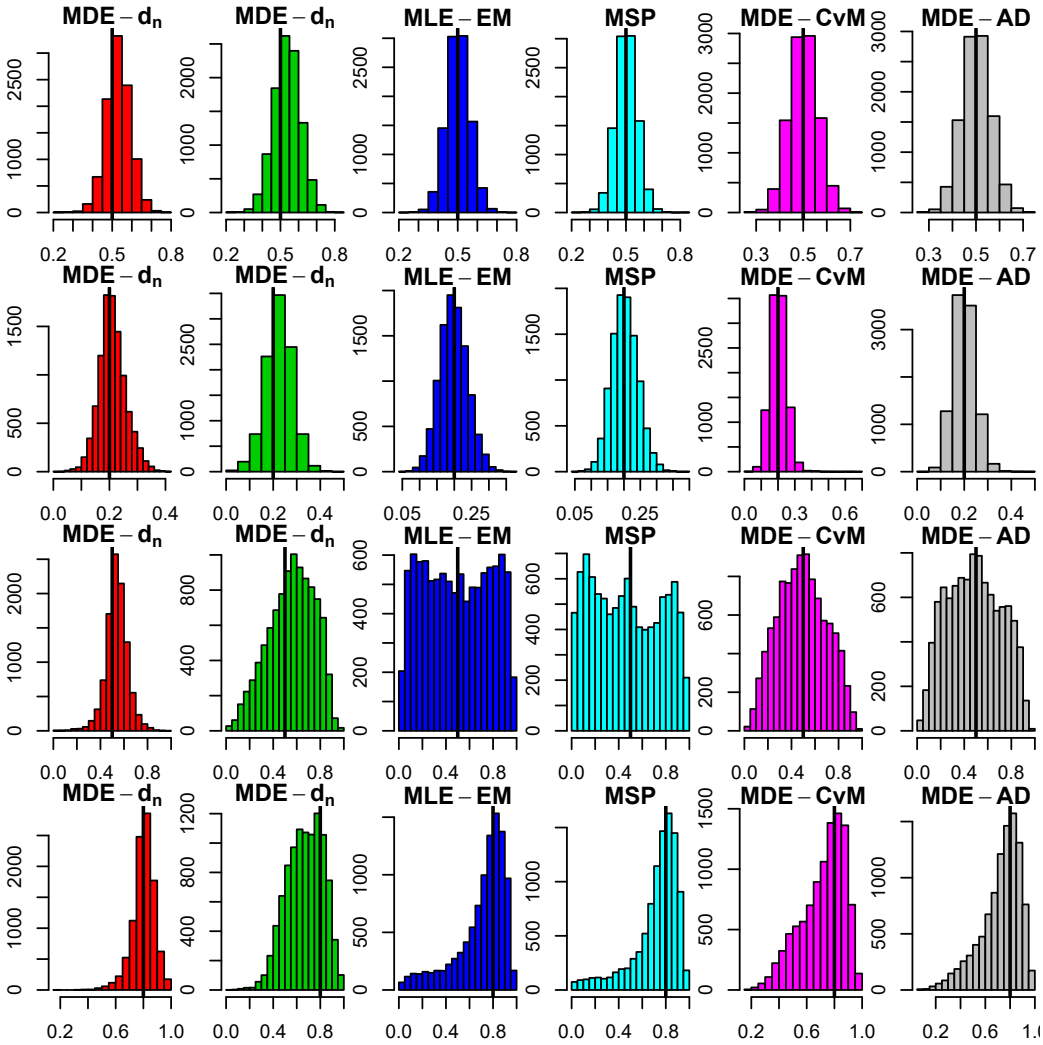

Figure B.8: Histograms of the weight parameter estimates for the four normal mixture distributions, when estimators are initialized with true parameters. Each row corresponds to a distribution. From top to bottom: symmetric bimodal, asymmetric bimodal, symmetric unimodal, asymmetric unimodal. The results are based on $N = 10^4$ samples of size $n = 100$.

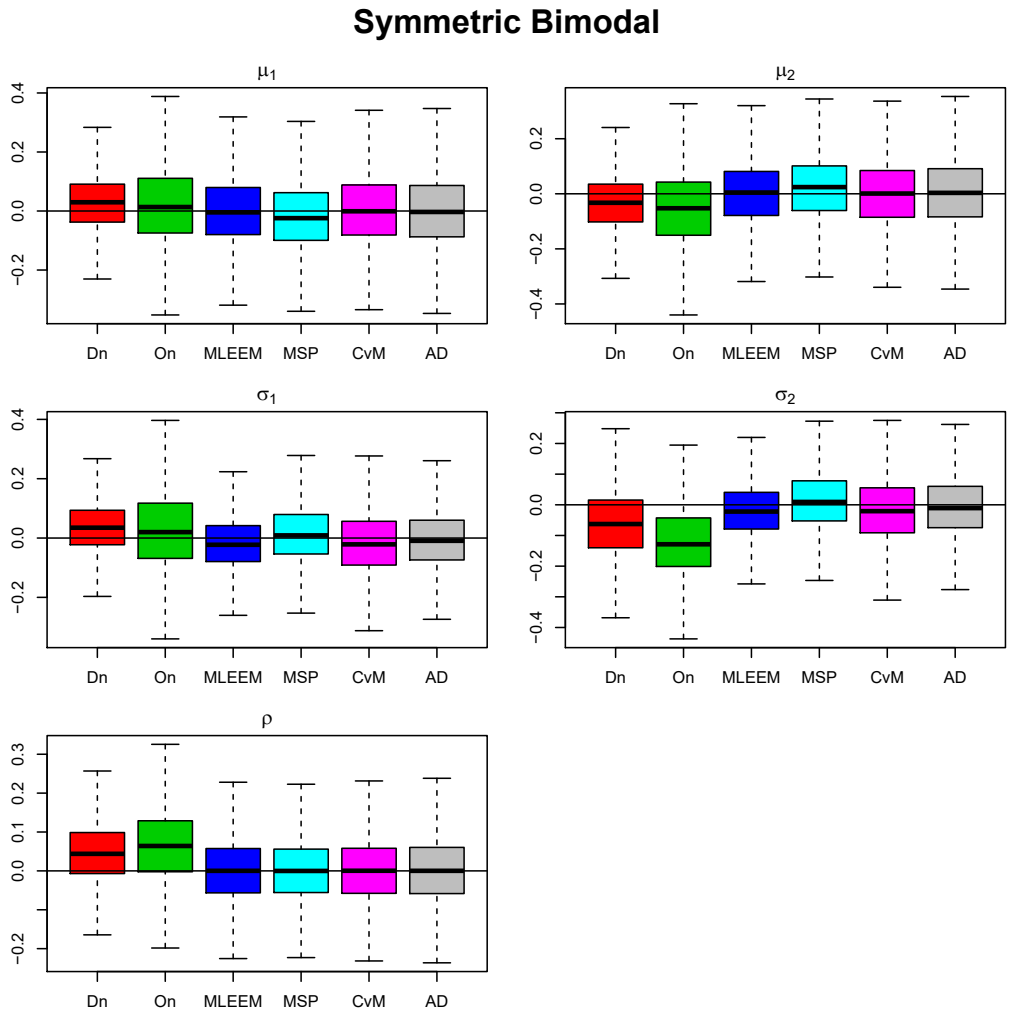

Figure B.9: Boxplots of parameter estimates of the symmetric bimodal normal mixture distribution, when estimators are initialized with true parameters. True parameter values are subtracted and outliers are not shown. The results are based on $N = 10^4$ samples of size $n = 100$.

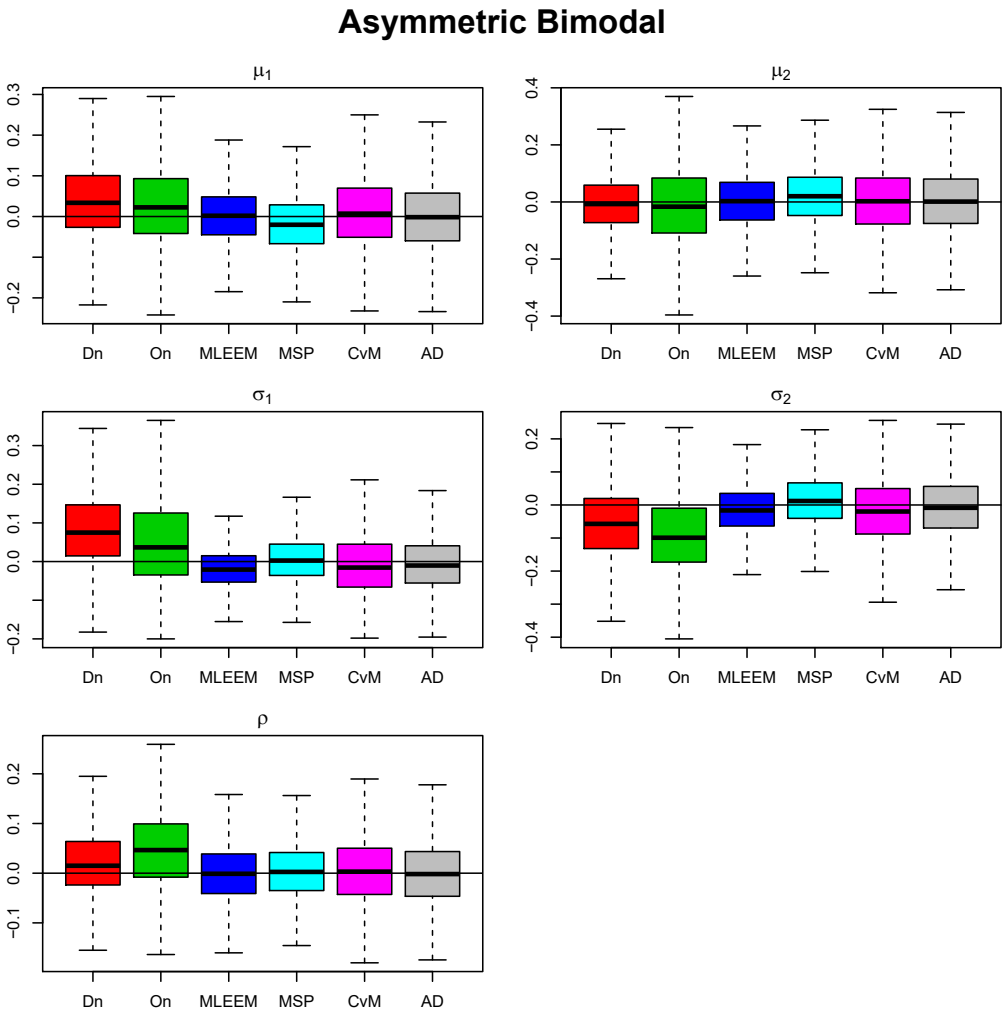

Figure B.10: Boxplots of parameter estimates of the asymmetric bimodal normal mixture distribution, when estimators are initialized with true parameters. True parameter values are subtracted and outliers are not shown. The results are based on $N = 10^4$ samples of size $n = 100$.

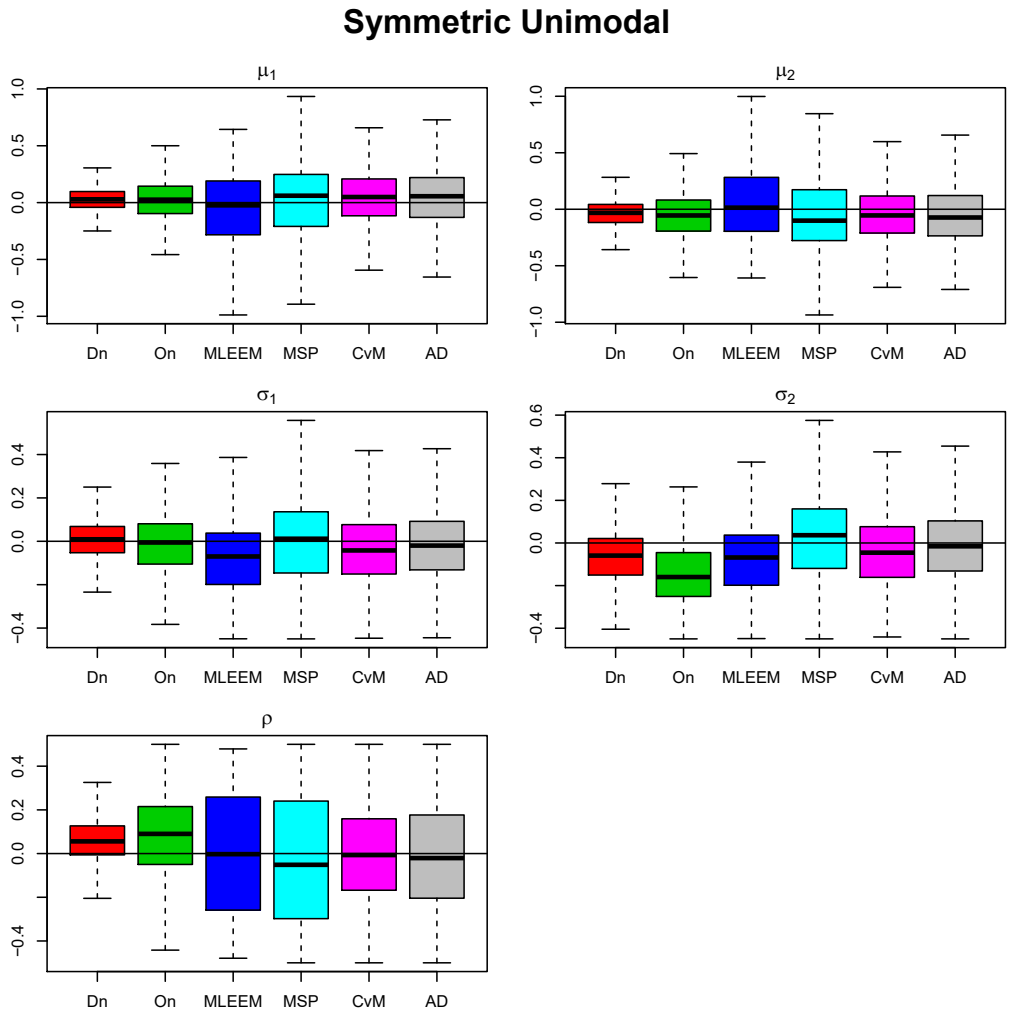

Figure B.11: Boxplots of parameter estimates of the symmetric unimodal normal mixture distribution, when estimators are initialized with true parameters. True parameter values are subtracted and outliers are not shown. The results are based on $N = 10^4$ samples of size $n = 100$.

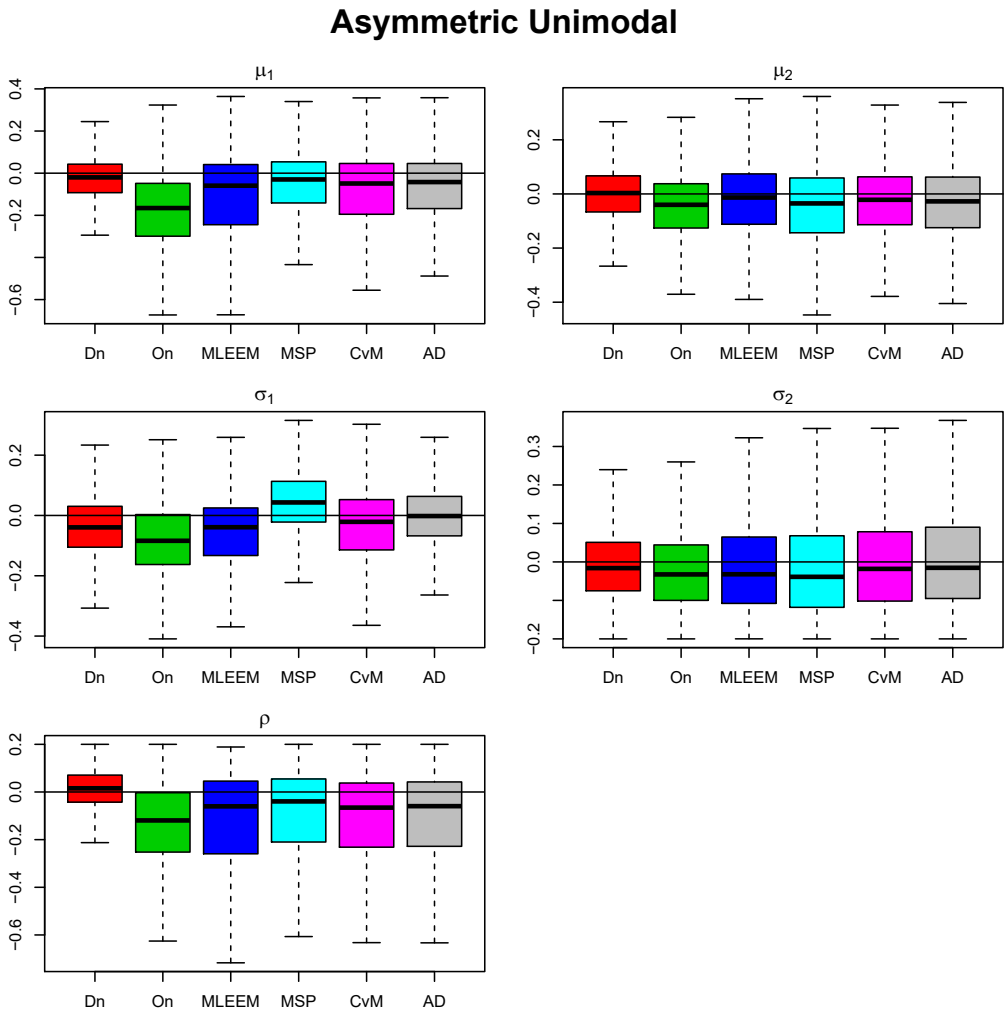

Figure B.12: Boxplots of parameter estimates of the asymmetric unimodal normal mixture distribution, when estimators are initialized with true parameters. True parameter values are subtracted and outliers are not shown. The results are based on $N = 10^4$ samples of size $n = 100$.

Single Random Initialization

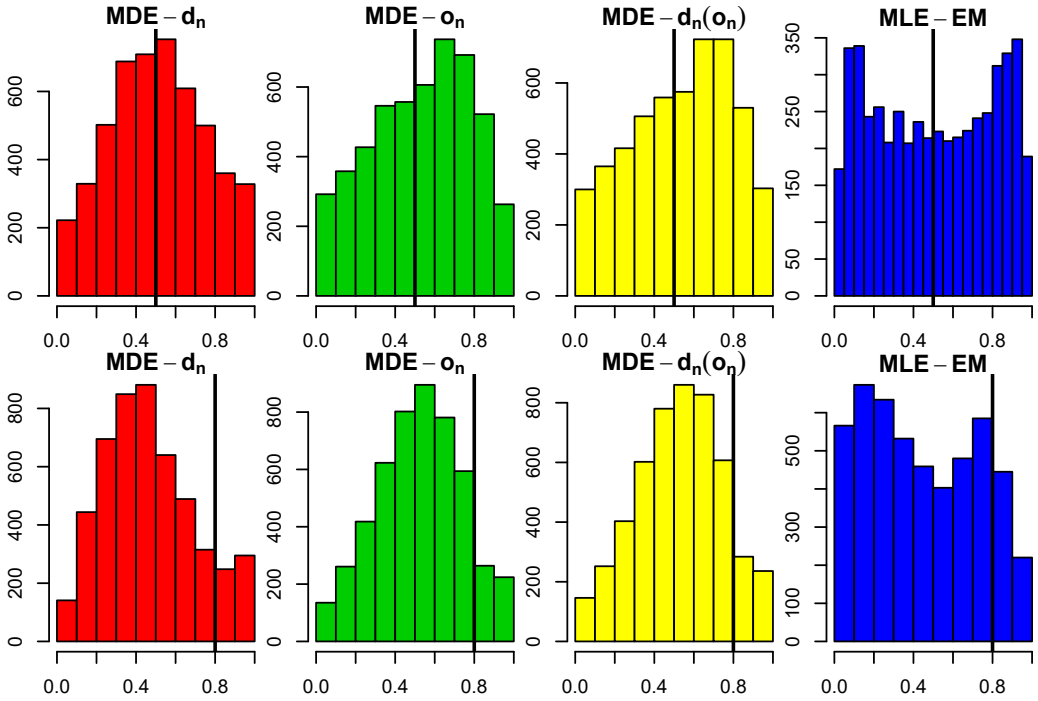

Figure B.13: Histograms of the weight parameters for the two unimodal normal mixture distributions, when estimators are initialized with K-means++ initialization. Each row corresponds to a distribution. From top to bottom: symmetric unimodal, asymmetric unimodal. The results are based on $N = 10^4$ samples of size $n = 100$.

Multiple Random Initializations

We have reviewed the performance of the MDE-d_n in the unimodal mixture models when the method was initialized once with K-means++ initialization in Section 5.2.3. A common way to improve initialization is by considering multiple initial parameters. The final estimate is then selected as the estimate with the lowest distance measure.

In Section 5.2.3, we observe that transitioning from true initialization to random initialization resulted in lost performance for the MDE-d_n. It is therefore interesting to see if we can recover some of this performance by initializing multiple times. We consider the symmetric unimodal mixture, with parameters given in Table 5.1.

First off, we are interested in how many times we should initialize. For this, we consider the mean *distance* attained for different number of initializations. The results are given in Figure B.14. As we can see, the first few extra iterations are the most rewarding, especially for the MDE-d_n. After 5 iterations or more, the distance measure of both methods decrease only slightly for each additional initialization.

Next, we study the *performance* for different number of initializations. The Hellinger distances, as defined in (5.3), and the volumes of the MSE concentration ellipsoid are given in Figure B.15.

We observe some striking behavior; more initializations can *worsen* the results when considering these performance measures. However, note that we must be careful when interpreting the results, as discussed in Section 5.2.1. Consider the histogram of the weight parameter estimates, given in Figure B.16 and Figure B.17. We see that, for the MDE-d_n, for larger number of iterations, the distribution of the weight parameter becomes slightly more uniform over the interval [0,1] as opposed to clustered around $\frac{1}{2}$. For the MLE-EM, the weight parameters are estimated at the extremes even more often. As a result, the performance measures across the different number of initializations are hard to compare and interpret. The marginal MSE of some parameters does decrease when we have more initializations, but as others increase, this is not reflected in the volume of the MSE concentration ellipsoids.

We find that, for both estimators, local minima exist that provide very good estimates for the symmetric unimodal normal mixture. However, when we allow multiple initializations, the optimizer finds other minima, which have even lower distance measures, but do not necessarily correspond to better estimates when we consider the Hellinger distance or volume of MSE concentration ellipsoid. These minima correspond with estimates where the weight parameter tends more towards the extremes. This is true for both the MDE-d_n and MLE-EM estimator.

Conclusions. In this symmetric unimodal setting, we conclude the following. The performance of the MDE-d_n lost when initializing with K-means++ as opposed to initializing with true initialization, cannot be recovered by initializing more often. Instead, multiple initialization leads to estimates of the weights further from the true values, as these points are associated with lower distance value d_n. This leads to estimates with higher MSE and Hellinger

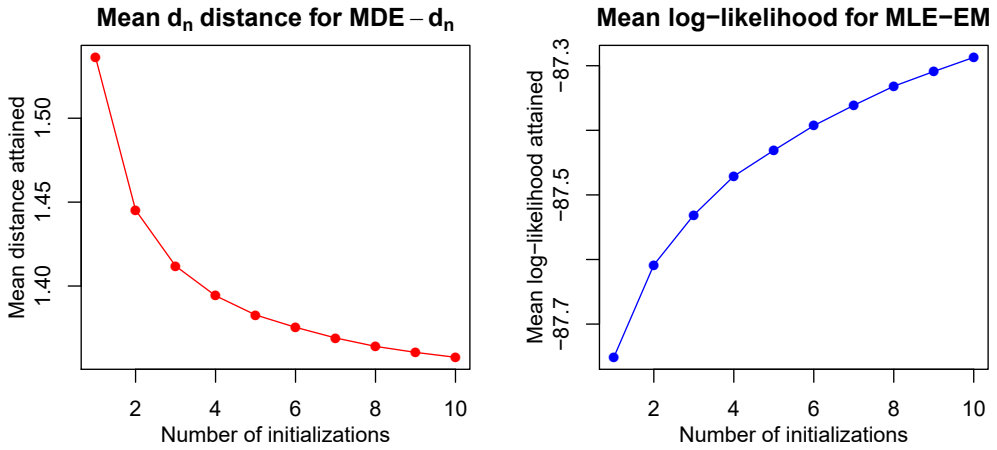

Figure B.14: Means of the respective distances obtained for the MDE-d_n (d_n distance) and MLE-EM (log-likelihood) estimators, for different numbers of K-means++ initializations, for the symmetric unimodal normal mixture distribution. Results are based on $N = 10^4$ samples of size $n = 100$.

distance. This behavior is not exclusive to the MDE-d_n estimator. The MLE-EM estimator exhibits the same behavior, but exhibits it even more pronounced, with weight parameters clustering at the extremes. Initializing more often and blindly selecting the estimate with the lowest distance is therefore not necessarily the best course of action for these estimators in this setting.

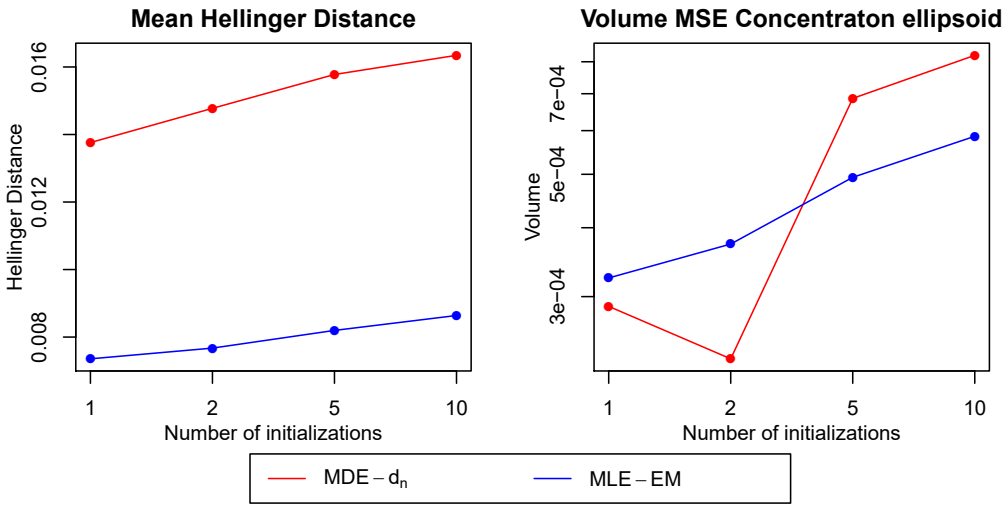

Figure B.15: Performance measures for different numbers of K-means++ initializations, for the symmetric unimodal normal mixture distribution. Results are based on $N = 10^4$ samples of size $n = 100$.

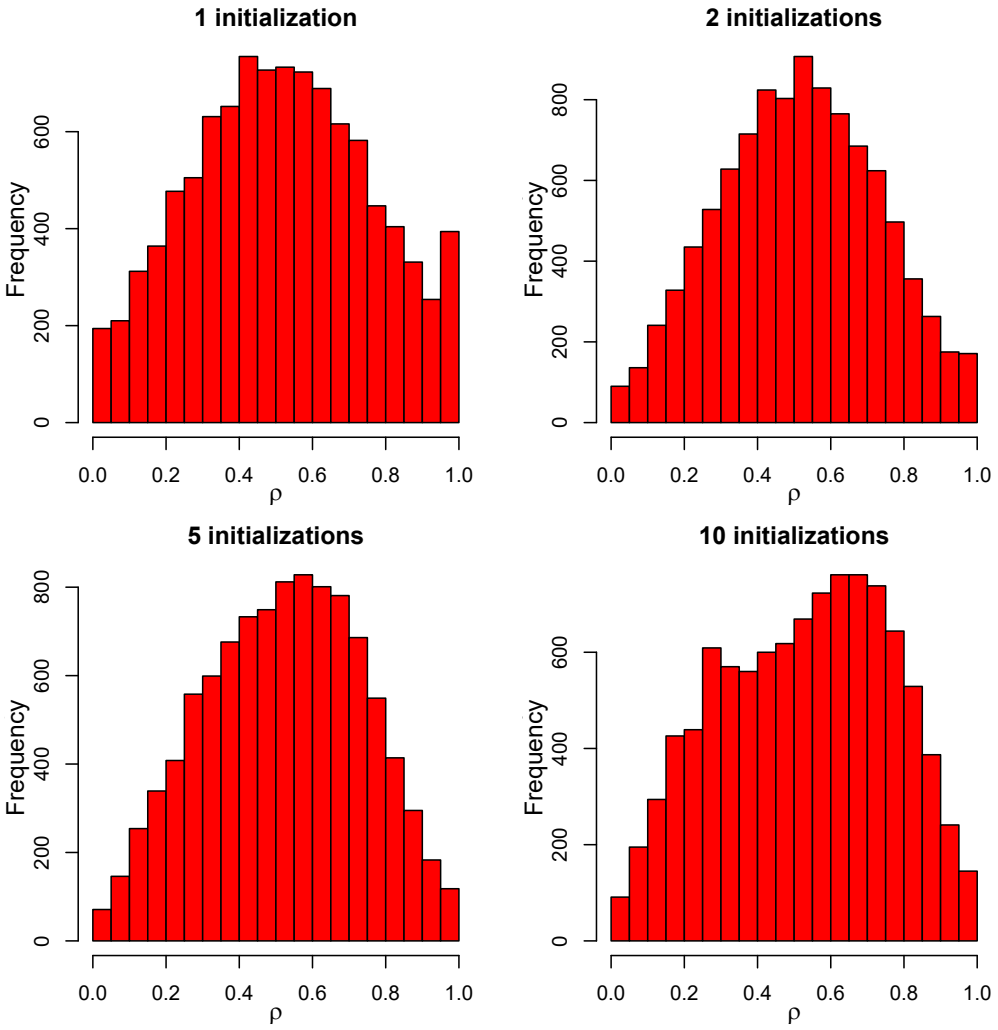

Figure B.16: Histograms of the MDE-d_n estimates of the weight parameter for different numbers of K-means++ initializations, for the symmetric unimodal normal mixture distribution. Results are based on $N = 10^4$ samples of size $n = 100$.

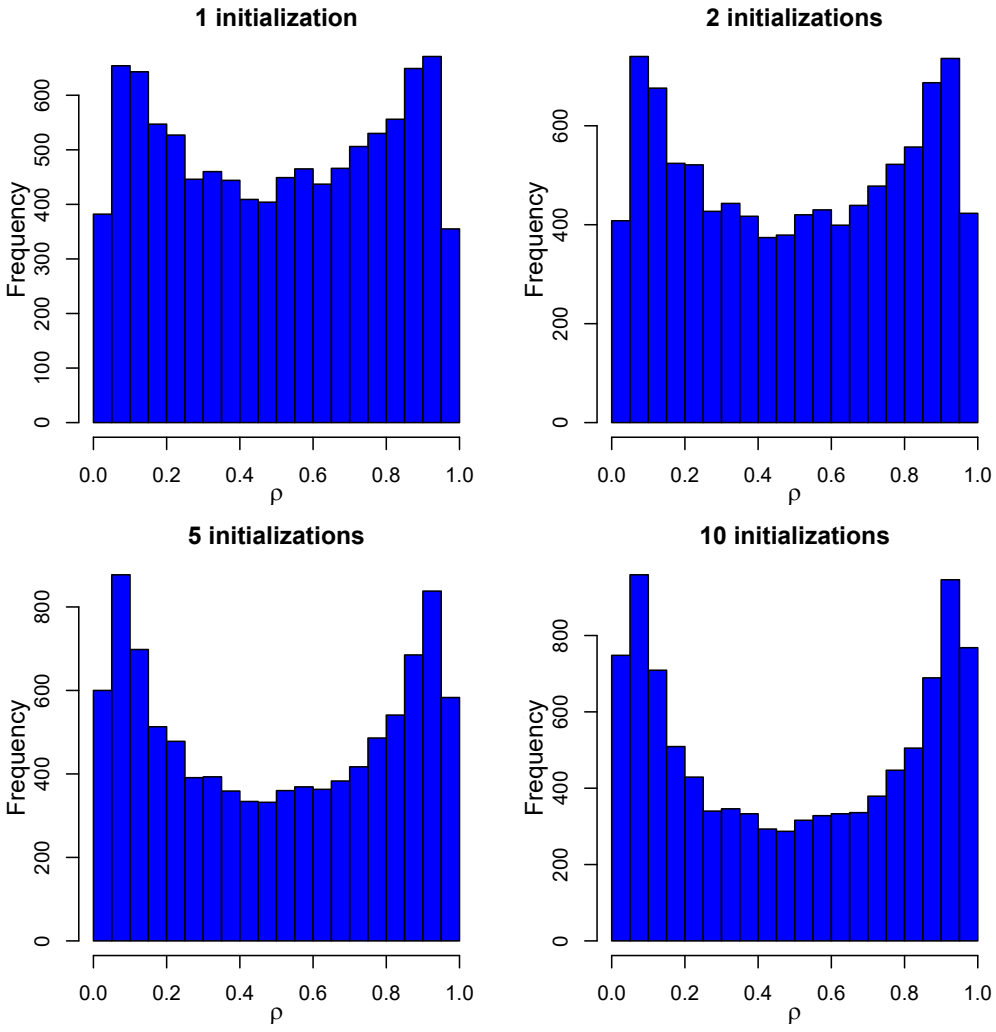

Figure B.17: Histograms of the MLE-EM estimates of the weight parameter for different numbers of K-means++ initializations, for the symmetric unimodal normal mixture distribution. Results are based on $N = 10^4$ samples of size $n = 100$.

Appendix C

Code

This appendix contains R code used for simulations throughout the main text.

C.1 Computation d_n

```
#' Compute the dn statistic for an iid univariate
   uniform(0,1) sample
#'
#' @param x The uniform sample for which the statistic
   will be computed, which does not need to be sorted.
#'
#' @return The value of the statistic dn for the sample
   x
dn <- function(x){
  #Save the data length in n
  n <- length(x)

  #Compute the order statistics
  u <- sort(x)

  #Save places just before and after discontinuities in
     Fn, which are the datapoints
  #and just before the datapoints
  epsilon <- 10^-16
  #The vector which will save the coordinates of
     discontinuities
  before.and.after.jumps <- numeric(2*n)
```

```
#Fill odd places with coordinates just before
   discontinuity
before.and.after.jumps[seq(1,2*n,2)] <- u-epsilon
#Fill even places with coordinates of the datapoint
before.and.after.jumps[seq(2,2*n,2)] <- u
#Note that before.and.after.jumps is already sorted,
   as u was sorted.

#Save the ECDF of the sample
Fn <- ecdf(u)

##Precompute spacings and their squared sum
u0 <- c(0,u)
t <- diff(u0) #t will contain u_1 - 0, u_2 - u_1, u_3
   - u_2, ... , u_n - u_{n-1}

#Compute the sum of squared spacings of all spacings
   up until and including u_{i-1}, so:
#sum.of.squared.spacings[1] = 0,
#sum.of.squared.spacings[2] = (u1 - u0)^2
#sum.of.squared.spacings[3] = (u1 - u0)^2 + (u2 - u1)
   ^2
#sum.of.squared.spacings[i] = (u1 - u0)^2 + ... + (u{i
   -1} - u{i-2})^2.
sum.of.squared.spacings <- c(0,cumsum(t^2))

#Define the function In(s), dependent on sorted sample
    u and the sum of squared spacings
In <- function(s){
  #Compute the maximum value of i such that u_i < s
  max.i <- round(n*Fn(s)) #Must use rounding to
     prevent numerical errors

  #Compute the area as the sum of squared spacings and
      the remainder incomplete squared spacing
  #The sum of complete spacings must be until u_i, so
     we need to retrieve the value at [max.i + 1],
  #as sum.of.squared.spacings[i] includes spacings
     until u_{i-1}.
  #The incomplete final spacing must be between u_{max
     .i} and s, and u0 has a leading 0 in case s < u1,
  #so we retrieve u_{max.i} at u0[max.i + 1]
```

```
    area <- ((n+1)/2) * ( sum.of.squared.spacings[max.i
       +1] + (s - u0[max.i+1])^2 )
    return(area)
  }

  #Returns the absolute difference between In and Fn (
     which is the argument for optimization),
  #|In - Fn|, at point t
  diff.argument <- function(t){
    return(abs(In(t) - Fn(t)))
  }

  #Find sup_{0<=t<=1} of |In - Fn|, which must occur at
     or before the points of discontinuity, precomputed
  sup.diff <- max(diff.argument(before.and.after.jumps))

  #Return the value of the statistic by multiplying the
     maximum by 2*sqrt(n)
  d.n <- 2*sqrt(n+1)*sup.diff

  return(d.n)
}
```

C.2 Computation o_n

```
#' This function computes the statistic on of a uniform
   sample x
#'
#' @param x The uniform sample for which the statistic
   will be computed, which does not need to be sorted.
#'
#' @return The value of the statistic on for the sample
   x
on <- function(x){

  #Save the data length in n
  n <- length(x)

  #Sort the data and prepend/append 0 and 1 at the
     endpoints
  u <- sort(x)
  u <- c(0,u,1)
```

```
  #Compute the spacings between each datapoint
  t <- diff(u)

  #Compute the values of S_i. As the first value of S_i
     is 0, prepend a 0
  S <- c(0, ((n+1)/2)*cumsum(t^2))
  S <- S[1:(n+1)] #The last value of S_i is never used,
     so remove it to prevent errors with different-sized
      vectors later

  #Compute the values of the sequences (i-1)/n and (i-1)
     /3
  q <- (1:(n+1) - 1) / n
  q3 <- (1:(n+1) - 1) / 3

  #Compute the terms for each i inside the sum
  #Note that vectors in R start at 1
  #For the first term, S should be 0, t should be the
     first spacing (u1-0), q should be 0 and q3 should
     be 0.
  #For the second term, S should be (n/2)*(u1-0)^2, t
     should be (u2-u1), q should be 1/n and q3 should be
      1/3
  #The length of all vectors should be (n+1)
  a <- S^2*t - 2*q*S*t + (n+1)/3*S*t^3 + q^2*t - ((n+1)/
     n)*q3*t^3 + (((n+1)^2)/20)*t^5

  #Compute the sum of all the terms a, and return them
  return(4*(n+1)*sum(a))
}
```

C.3 MDE Exponential Distribution

This is an R implementation of the MDE for the exponential distribution. For other distributions, the code is similar, except for the null distribution and, in case of more than one parameter, the initialization.

```
#' Minimum-distance estimator for exponential
   distribution
#'
#' @param x The data
#' @param stat The distance statistic used in the MDE
#' @param interval The interval for which the MDE should
    search for the minimum value of the parameter
```

```
#'
#' @return A list with components $est with the
   estimation and $dist with the minimum distance found
MDE.exp <- function(x,stat,interval){
  #The exponential distribution function as a function
     of the data and the parameters
  F0 <- function(x,t){
    return(pexp(x,t))
  }

  #Compute the value of the statistic as a function of
     the parameters. Make sure that the parameters
  #are in the feasable region, otherwise set value to
     infinity.
  minimization.statistic <- function(t){
    if(t <= 0){
      return(Inf)
    }else{
      return(stat(F0(x,t)))
    }
  }

  #Call an appropriate minimization subroutine to
     minimize the statistics
  optimum <- optimize(f=minimization.statistic,interval=
     interval)

  #Return the minimum (the estimate) and the minimum
     distance found (in case comparison between
     estimates
  #need to be made following different initializations).
  return(list("est"=optimum$minimum,"dist"=optimum$
     objective))
}
```

C.4 MSP Exponential Distribution

This is an R implementation of the MSP for the exponential distribution. For other distributions, the code is similar, except for the null distribution and, in case of more than one parameter, the initialization.

For large values of the rate parameter, the difference between the transformed

samples become so small that machine precision is no longer accurate, and the difference is set to 0. This causes the logarithm to evaluate to negative infinity. While this is not always a problem, if the optimization region is too broad, the region where the distance evaluates to infinity become too large as well, and the optimization does not work properly. Therefore, we implement a procedure where, if the estimate has associated distance value of infinity, we shorten the right-hand-side of the interval with a factor.

```
#' Maximum-product of spacings estimator for exponential
    distribution
#'
#' @param x The data
#' @param stat The distance statistic used in the MSP
#' @param interval The interval for which the MDE should
    search for the minimum value of the parameter
#'
#' @return A list with components $est with the
   estimation and $dist with the minimum distance found
MSP.exp <- function(x,interval,reduceratio = 0.5){
  #The exponential distribution function as a function
     of the data and the parameters
  F0 <- function(x,t){
    return(pexp(x,t))
  }

  #Compute the sum of the logs of the uniform spacings
  sum.log.spacings <- function(t){
    #Compute the probability integral transform as a
       function of theta to obtain ordered uniform
       spacings
    u <- c(0,sort(F0(x,t)),1)
    #Compute the sum of the logs of spacings, and return
        -1*sum so minimizing this maximizes spacings. If
    #spacings are too close, we get NA values, in which
       case we return Infinity.
    sum.log.spacings <- sum(log(diff(u)))
    if(is.infinite(sum.log.spacings)){
      return(Inf)
    }else{
      return(-1*sum.log.spacings)
    }
  }
```

```
  #Compute the value of the statistic as a function of
     the parameters. Make sure that the parameters
  #are in the feasable region, otherwise set value to
     infinity.
  minimization.statistic <- function(t){
    if(t <= 0){
      return(Inf)
    }else{
      return(sum.log.spacings(t))
    }
  }

  #Call an appropriate minimization subroutine to
     minimize the statistics
  optimum <- list()
  optimum$objective <- Inf
  while(is.infinite(optimum$objective)){
    interval <- c(interval[1],reduceratio*interval[2])
    optimum <- optimize(f=minimization.statistic,
       interval=interval)
  }


  #Return the minimum (the estimate) and the minimum
     distance found (in case comparison between
     estimates
  #need to be made following different initializations).
  return(list("est"=optimum$minimum,"dist"=optimum$
     objective))
}
```