Modern Stochastics: Theory and Applications 5 (1) (2018) 1–36
https://doi.org/10.15559/17-VMSTA93

A moment-distance hybrid method for estimating a mixture of two symmetric densities

David Källberg*, **Yuri Belyaev**, **Patrik Rydén**

Department of Mathematics and Mathematical Statistics, Umeå University, SE-901 87 Umeå, Sweden

david.kallberg@umu.se (D. Källberg), yuri.belyaev@umu.se (Y. Belyaev), patrik.ryden@umu.se (P. Rydén)

Received: 4 October 2017, Revised: 13 December 2017, Accepted: 18 December 2017, Published online: 18 January 2018

Abstract In clustering of high-dimensional data a variable selection is commonly applied to obtain an accurate grouping of the samples. For two-class problems this selection may be carried out by fitting a mixture distribution to each variable. We propose a hybrid method for estimating a parametric mixture of two symmetric densities. The estimator combines the method of moments with the minimum distance approach. An evaluation study including both extensive simulations and gene expression data from acute leukemia patients shows that the hybrid method outperforms a maximum-likelihood estimator in model-based clustering. The hybrid estimator is flexible and performs well also under imprecise model assumptions, suggesting that it is robust and suited for real problems.

Keywords Inference for mixtures, method of moments, minimum distance, model-based clustering

2010 MSC 62F07, 62F10, 62F35, 62P10, 92D10

1 Introduction

Mixture distributions are used in many fields of science for modeling data taken from different subpopulations. An important medical application is clustering of gene ex-

*Corresponding author.

pression data to discover novel subgroups of a disease. This is a high-dimensional problem and it is common to do a variable selection to obtain a subset of genes whose expression contribute in separating the subgroups. For two-class problems this may be carried out by fitting a univariate mixture distribution to each gene and single out variables for which the overlap between the component distributions is small enough [27]. There are also multivariate methods to do the variable selection, which are more computationally demanding but take into account the possible correlations between the genes and therefore reduce the loss of information in the univariate approach where each gene is modeled separately [9]. Further applications of mixtures can be found in image analysis [25], outlier detection [16], remote sensing [18], and epidemiology [24].

Karl Pearson [22] used the method of moments as a first attempt to estimate the parameters of a mixture distribution. Since then the computational difficulty of the problem and the increasing number of applications have sparked a vast amount of theoretical and applied research. Maximum likelihood inference was simplified with the introduction of the expectation-maximization (EM) algorithm in 1970s [7] and is up to now the most applied and studied approach, see [19] and references therein. Various modifications of the basic likelihood method have been proposed, aiming to overcome drawbacks resulting from the unbounded likelihood function and the sensitivity of outliers. For example, in [11] a family of divergences was introduced that is used as a generalization of the likelihood. There are also several variants of the EM-algorithm available, such as stochastic versions [4] and constrained formulations [15]. Minimum distance estimators is another family of parametric methods that has been applied extensively for mixtures, in particular due to its robustness against imprecise distributional assumptions [28, 5, 6]. Semiparametric techniques have also attracted much interest in this field, for example, estimation of location mixtures where only symmetry is imposed on the density components [2, 17]. For a comprehensive introduction to inference for mixtures, we refer to the monograph [26].

In this paper, we propose a hybrid approach for estimating five parameters of a mixture of two densities which are symmetric about their means. The approach combines the method of moments with a minimum distance estimator based on a quadratic measure of deviation between the fitted and empirical distribution functions. The motivation behind our approach is to develop a robust algorithm that produces accurate estimates also when the parametric shape of the mixture distribution is misspecified, which is common in practice.

The paper is organized as follows. In Section 2, we introduce the hybrid estimator and describe how it is obtained from empirical data. Section 3 is devoted to a simulation study where the proposed estimator is evaluated and compared to a conventional maximum likelihood estimator obtained via the EM-algorithm. We consider the methods' performance in estimating the unknown partition of a data set containing observations from two populations (model-based clustering), which is an important application of mixture distributions. We also evaluate the methods's ability to estimate the mixing proportion. In Section 4, we report the results of a case study where the methods are applied on gene expression data from patients with acute leukemia. In Section 5, we discuss the results and draw some conclusions.

2 The moment-distance hybrid method

In this section we present the novel moment-distance hybrid estimator (HM-estimator) and describe how it can be used for model-based clustering. We consider the problem where the real-valued random variable X has a two-component mixture distribution $F(\cdot)$ with density

$$f(x) = pf_1(x) + (1-p)f_2(x), \quad x \in \mathbb{R},$$

where $0 < p < 1$ is the mixing proportion and $f_i(\cdot) = f_i(\cdot|\mu_i, \sigma_i^2)$ is the density of a random variable X_i completely specified by its mean μ_i and variance $\sigma_i^2, i = 1, 2$. We assume that the third moment $E|X|^3$ is finite, and that the component densities $f_1(\cdot)$ and $f_2(\cdot)$ are *symmetric* about their means. A mixture of two bounded and symmetric densities has these properties, for example a two-component normal mixture. Let $\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ denote the parameter vector for $F(\cdot)$.

The HM-estimator, denoted $\hat{\theta}_{HM}$, is an estimator of θ that combines the method of moments and the minimum distance method. The method of moments is used to reduce the parameter space and the minimum distance approach, aiming to minimize the distance between the fitted model and the empirical distribution, is used to obtain the estimator $\hat{\theta}_{HM}$.

2.1 Definition of the HM-estimator

An estimate $\hat{\theta} = (\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ of θ based on a sample $x_1, \ldots, x_n$ is called *relevant* if

$$\begin{aligned} &0 < \hat{p} < 1, \\ &\hat{\sigma}_1^2, \hat{\sigma}_2^2 > 0, \\ &\min(x_1, \ldots, x_n) \le \hat{\mu}_i \le \max(x_1, \ldots, x_n), \quad i = 1, 2. \end{aligned}$$

Let Ω denote the set of all relevant estimates of θ. The method of moments is applied to reduce Ω to a subset Ω' of lower dimension. The first three moments of X can be expressed as

$$\nu_1 := E(X) = p\mu_1 + (1-p)\mu_2, \tag{1}$$

$$\nu_2 := E\big(X^2\big) = p\big(\sigma_1^2 + \mu_1^2\big) + (1-p)\big(\sigma_2^2 + \mu_2^2\big), \tag{2}$$

$$\nu_3 := E\big(X^3\big) = p\big(3\mu_1\sigma_1^2 + \mu_1^3\big) + (1-p)\big(3\mu_2\sigma_2^2 + \mu_2^3\big), \tag{3}$$

where the last equality relies on the symmetry of the component densities $f_1(\cdot)$ and $f_2(\cdot)$, see Appendix A.1 for details. Following the method of moments, we replace the parameters in (1)–(3) by their estimators while equating the theoretical moments $\{\nu_i\}$ with their sample counterparts

$$\hat{\nu}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{\nu}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \qquad \hat{\nu}_3 = \frac{1}{n}\sum_{i=1}^{n} x_i^3.$$

These sample moments can be highly variable so we suggest below to replace them with the more robust trimmed means, denoted $\{\hat{\nu}_k^*\}$, $k = 1, 2, 3$, see Section 3.1 for further details. We get the following undetermined system of equations:

$$\hat{\nu}_1^* = \hat{p}\hat{\mu}_1 + (1-\hat{p})\hat{\mu}_2, \tag{4}$$

$$\hat{\nu}_2^* = \hat{p}\big(\hat{\sigma}_1^2 + \hat{\mu}_1^2\big) + (1-\hat{p})\big(\hat{\sigma}_2^2 + \hat{\mu}_2^2\big), \tag{5}$$

$$\hat{\nu}_3^* = \hat{p}\big(3\hat{\mu}_1\hat{\sigma}_1^2 + \hat{\mu}_1^3\big) + (1-\hat{p})\big(3\hat{\mu}_2\hat{\sigma}_2^2 + \hat{\mu}_2^3\big). \tag{6}$$

The set $\Omega' \subseteq \Omega$ consists of all relevant estimates which solve the system (4)–(6). We define the HM-estimator as an element of Ω' with a minimum distance criteria as

$$\hat{\theta}_{HM} = \underset{\hat{\theta}\in\Omega'}{\arg\min}\, d\big(F(\cdot|\hat{\theta}), F_n(\cdot)\big), \tag{7}$$

where the function $d(F(\cdot|\hat{\theta}), F_n(\cdot))$ measures the distance between the fitted model distribution $F(\cdot|\hat{\theta})$ and the empirical distribution $F_n(\cdot)$ of the sample. In this paper we use an L_2-type measure given by

$$d\big(F(\cdot|\theta), F_n(\cdot)\big) = \frac{1}{n}\sum_{i=1}^{n}\big(F(x_i|\theta) - F_n(x_i)\big)^2 = \frac{1}{n}\sum_{i=1}^{n}\big(F(x_{(i)}|\theta) - i/n\big)^2,$$

where $x_{(1)}, \ldots, x_{(n)}$ is the ordered sample. It should be noted that this choice of distance is mainly due to its computational simplicity and that a number of different measures can be considered (see [26], chap. 4)

Next we describe how the HM-estimator is obtained in practice, via a reformulation of definition (7) that is more useful for computation.

2.2 *How to compute the HM-estimator*

In this subsection we describe how the HM-estimator (7) can be obtained in practice. To get a representation of the solutions of the system (4)–(6), we reparametrize the problem by introducing the proportion $\hat{r}$, defined by

$$\hat{\sigma}_2^2 = \hat{r}\hat{\sigma}_1^2. \tag{8}$$

Equations (4), (5), and (8) can be used to eliminate $\hat{\mu}_2$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ in equation (6), and as a result we obtain

$$\beta_3\hat{\mu}_1^3 + \beta_2\hat{\mu}_1^2 + \beta_1\hat{\mu}_1 + \beta_0 = 0, \tag{9}$$

where the coefficients are functions of $\hat{p}$ and $\hat{r}$ such that

$$\beta_0 = -\hat{\nu}_3^* + \frac{3\hat{\nu}_1^*\hat{\nu}_2^*\hat{r}}{\hat{p}+\hat{r}-\hat{p}\hat{r}} + \big(\hat{\nu}_1^*\big)^3\frac{3\hat{p} - 2(\hat{p}+\hat{r}-\hat{p}\hat{r})}{(1-\hat{p})^2(\hat{p}+\hat{r}-\hat{p}\hat{r})},$$

$$\beta_1 = 3\hat{\nu}_2^*\frac{\hat{r}-\hat{p}+\hat{r}-\hat{p}\hat{r}}{(1-\hat{p})(\hat{p}+\hat{r}-\hat{p}\hat{r})} - \big(\hat{\nu}_1^*\big)^2\frac{3\hat{p}(2\hat{p} - 2(\hat{p}+\hat{r}-\hat{p}\hat{r}) + 1)}{(1-\hat{p})^2(\hat{p}+\hat{r}-\hat{p}\hat{r})},$$

$$\beta_2 = 3\hat{\nu}_1^*\frac{\hat{p}(2\hat{p} - \hat{p}^2 + \hat{p}^2\hat{r} - \hat{r})}{(1-\hat{p})^2(\hat{p}+\hat{r}-\hat{p}\hat{r})},$$

$$\beta_3 = -\frac{\hat{p}(2\hat{p} - \hat{p}^2 + \hat{p}^2\hat{r} - \hat{r})}{(1-\hat{p})^2(\hat{p} + \hat{r} - \hat{p}\hat{r})}.$$

Furthermore, by combining (4), (5), and (8), we get that the estimated parameters $\hat{\mu}_2$ and $\hat{\sigma}_1^2$ are obtained as

$$\hat{\mu}_2 = \frac{\hat{\nu}_1^* - \hat{p}\hat{\mu}_1}{1-\hat{p}}, \tag{10}$$

$$\hat{\sigma}_1^2 = \frac{\hat{\nu}_2^* - \hat{p}\hat{\mu}_1^2 - (1-\hat{p})\hat{\mu}_2^2}{\hat{p} + (1-\hat{p})\hat{r}}. \tag{11}$$

If $\hat{p}$ and $\hat{r}$ are given, we see that (9) is a cubic equation for $\hat{\mu}_1$ and so the reparameterized system has at most three solutions that correspond to relevant estimates. Define M to be the set of all pairs $(\hat{p}, \hat{r})$ for which at least one relevant estimate exists, and let $T(\hat{p}, \hat{r})$ contain all relevant estimates corresponding to the pair $(\hat{p}, \hat{r}) \in M$. From the definitions of Ω', M, and $T(\hat{p}, \hat{r})$ it follows that

$$\min_{\hat{\theta} \in \Omega'} d\big(F(\cdot|\hat{\theta}), F_n(\cdot)\big) = \min_{(\hat{p},\hat{r}) \in M} g(\hat{p}, \hat{r}), \tag{12}$$

where the function

$$g(\hat{p}, \hat{r}) = \min_{\hat{\theta} \in T(\hat{p},\hat{r})} d\big(F(\cdot|\hat{\theta}), F_n(\cdot)\big), \quad (\hat{p}, \hat{r}) \in M,$$

is straightforward to compute since $T(\hat{p}, \hat{r})$ contains at most three elements and these are solutions of the polynomial equations (8)–(11). Equation (12) reformulates the problem of deriving the HM-estimator (7) to a minimization problem for the bivariate function $g(\hat{p}, \hat{r})$, $(\hat{p}, \hat{r}) \in M$. Let $(\hat{p}_{HM}, \hat{r}_{HM})$ denote the point that minimizes the function $g(\hat{p}, \hat{r})$, $g(\hat{p}, \hat{r}) \in M$. Then the estimator is given as

$$\hat{\theta}_{HM} = \operatorname*{arg\,min}_{\hat{\theta} \in T(\hat{p}_{HM}, \hat{r}_{HM})} d\big(F(\cdot|\hat{\theta}), F_n(\cdot)\big).$$

The minimization of $g(\hat{p}, \hat{r})$ can be obtained using a numerical optimization algorithm. Here we use the simplex algorithm in [21], which is implemented in the optim routine in `R` [23]. The starting value is found as the minimizer of $g(\hat{p}, \hat{r})$ over a finite grid of values. A schematic description of how the HM-estimator is computed is given in Figure 1.

2.3 Model-based clustering

The mixture density $f(\cdot)$ is typically used to model a data set $x_1, \ldots, x_n$ where n_1 of the values are observations from component $f_1(\cdot)$ and the remaining $n_2 = n - n_1$ are observations from $f_2(\cdot)$. For such a sample, we can introduce a 0–1 vector $\mathbf{z} = (z_1, \ldots, z_n)$ that correctly assigns each observation to either $f_1(\cdot)$ (1's) or $f_2(\cdot)$ (0's). This vector defines a true partition of the observations with respect to the density components. Usually the components represent distinct subpopulations.

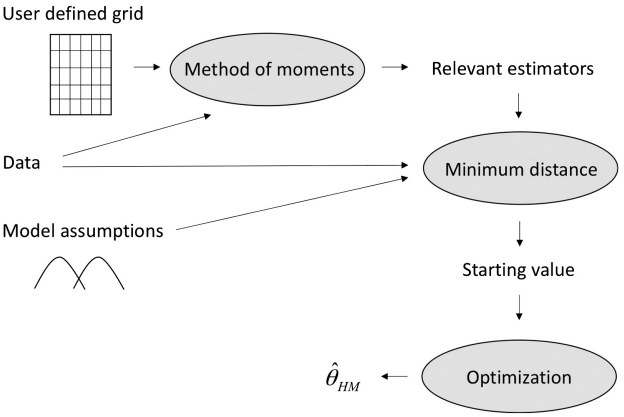


Fig. 1. A schematic description of how the HM-estimator is obtained. The user provides a grid with values $(\hat{p}, \hat{r})$. The method of moments is used to obtain all relevant estimates corresponding to the grid-points. The minimum distance approach is used to select the "best" of those estimates, which gives the starting point $(\hat{p}^{(0)}, \hat{r}^{(0)})$. Local grid optimization, minimizing the function $g(\hat{p}, \hat{r})$, with $(\hat{p}^{(0)}, \hat{r}^{(0)})$ as the starting value gives the HM-estimator. Note that the optimization step may be unnecessary if the grid is very dense

The true partition defined by $\mathbf{z} = (z_1, \ldots, z_n)$ is unobservable but can be estimated with the posterior membership probabilites, also known as the *responsibilites*, denoted by $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$, where

$$\hat{z}_i = \frac{\hat{p}\, f_1(x_i|\hat{\theta}_1)}{f(x_i|\hat{\theta})} = \frac{\hat{p}\, f_1(x_i|\hat{\theta}_1)}{\hat{p}\, \hat{f}_1(x_i|\hat{\theta}_1) + (1-\hat{p})\, \hat{f}_2(x_i|\hat{\theta}_2)}, \quad i = 1, \ldots, n. \tag{13}$$

The value $\hat{z}_i$ is in the interval [0, 1] and estimates the probability that x_i is an observation from component $f_1(\cdot)$. The vector $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_n)$ defines a so-called *soft* partition of the data $x_1, \ldots, x_n$ which serves as an approximation of the true partition given by $\mathbf{z} = (z_1, \ldots, z_n)$. The responsibilities in (13) can be obtained for any estimator $\hat{\theta}$ of θ, e.g. the proposed HM-estimator or the maximum likelihood (ML) estimator used in the numerical studies in Sections 3 and 4.

3 Simulation study

This section presents a simulation study where the proposed hybrid method (HM) is compared with a conventional ML-estimator derived via the EM-algorithm. We investigate the methods' performances in model-based clustering and their accuracy for estimating the mixing proportion. The consequences of calculating the estimators under incorrect model assumptions are getting particular attention.

3.1 Data and estimation

In the simulations, we restrict ourselves to the case where the component densities $f_1(\cdot)$ and $f_2(\cdot)$ belong to the same family of distributions. The estimators are calcu-

lated under the assumption that $f_1(\cdot)$ and $f_2(\cdot)$ are normal densities, which is a common assumption in practice. The data are generated from normal mixtures, for which the assumption is true, and also from mixtures of Laplace, logistic and contaminated Gaussian distributions (for details, see Appendix A.2). For the contaminated Gaussian distribution, we set the larger prior probability to $\alpha = 0.9$ and the variance proportion parameter to $\eta = 16$. The experiment thus includes both modest and large departures from the normal mixture assumption, allowing us to analyze the robustness of the methods with respect to imprecise model specifications.

Besides varying the family of the component densities, we consider six configurations of the parameter vector θ which correspond to a variety in shape of the mixture distributions. In addition to these configurations a negative control with a non-mixture distribution was added. The values are given in Table 1 and displayed graphically in Figure 2. Three sample sizes are considered; $n = 50$, 100, and 500.

Table 1. The configurations (i)–(vii) of the parameter vector θ used in the simulations

	Configuration						
Parameter	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)
μ_1	0	0	0	0	0	0	0
μ_2	2	3	3	4	4	3	0
σ_1^2	1	1	1	4	4	9	1
σ_2^2	1	1	1	1	1	1	1
p	0.50	0.50	0.25	0.50	0.25	0.50	0.10–0.50

The mixture data were generated as follows: first we simulated the (true) partition vector $\mathbf{z} = (z_1, \ldots, z_n)$ from the 0–1 variable Z with $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$. Then $x_1^{(1)}, \ldots, x_n^{(1)}$ and $x_1^{(2)}, \ldots, x_n^{(2)}$ were simulated from the component densities $f_1(\cdot)$ and $f_2(\cdot)$, respectively. Finally a sample $x_1, \ldots, x_n$ from the mixture density $f(\cdot)$ was obtained as

$$x_i = z_i x_i^{(1)} + (1 - z_i) x_i^{(2)}, \quad i = 1, \ldots, n.$$

We generated $N = 500$ data sets for each considered scenario (mixture family, parameter configuration, and sample size), and for each scenario we obtained 500 independent realizations $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(500)}$ of an estimator $\hat{\theta}$, which were used for statistical evaluation of the performance of the considered methods in clustering and in estimation of the mixing proportion.

Computing the estimators

The hybrid estimator $\hat{\theta}_{HM}$ was obtained as described in the previous section. The starting value for the simplex method was found as the minimizer of $g(\hat{p}, \hat{r})$ over a two-dimensional grid constructed from 10 values of $\hat{p}$ and 200 values of $\hat{r}$. The values of $\hat{p}$ and $\hat{r}$ in the grid were evenly distributed in the intervals (0, 1) and [0, 20], respectively. We used trimmed versions of the sample moments $\hat{\nu}_k$, $k = 1, 2, 3$. The 2.5% smallest and 2.5% largest of values in $x_1^k, \ldots, x_n^k$ were removed and the mean $\hat{\nu}_k^*$ of the resulting trimmed sample was used as the estimator of the kth moment ν_k.

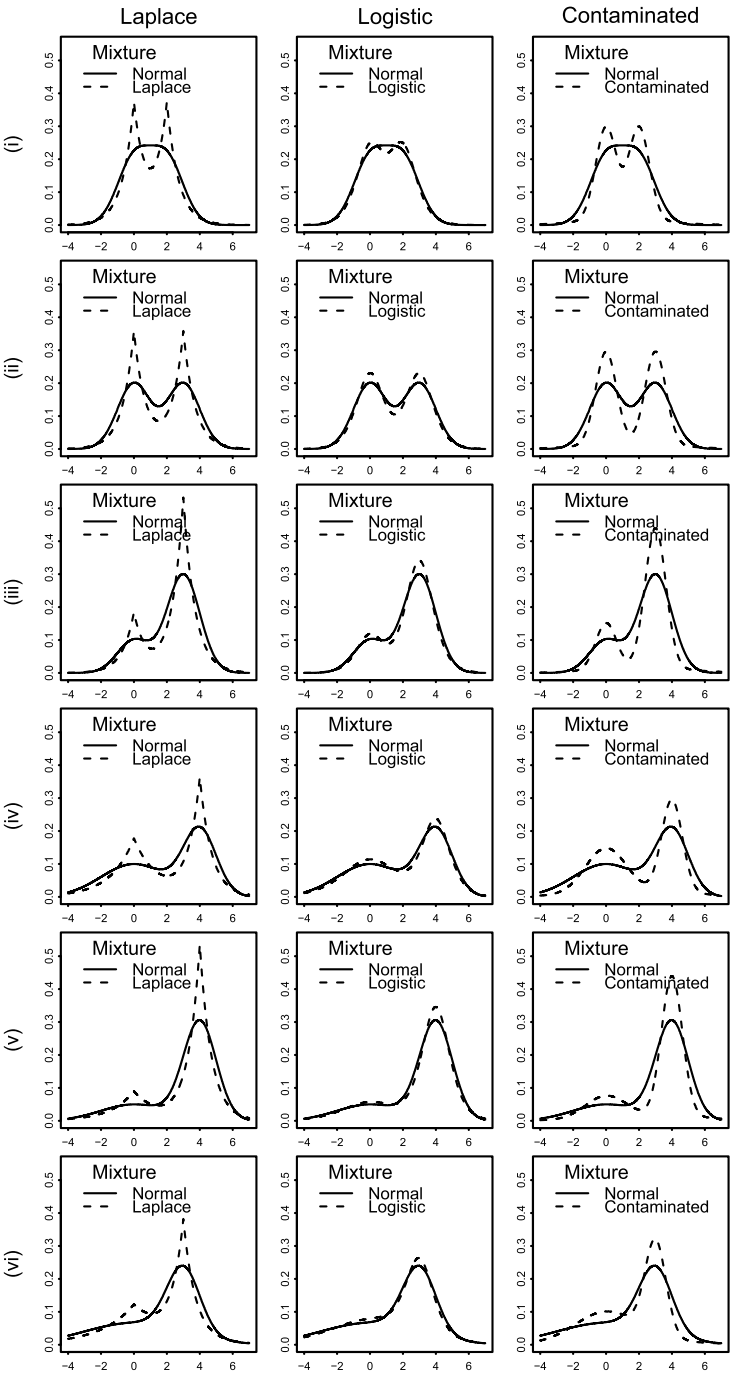


Fig. 2. The mixture distributions used in the simulations: four distribution families – mixtures of normal, logistic, Laplace, and contaminated normal distributions – and six parameter configurations (i)–(vi)

The maximum likelihood estimator $\hat{\theta}_{ML}$ was calculated with the EM-algorithm using the mixtools package in R [1]. The EM-algorithm converges at a local maximum of the likelihood function that depends on its starting value. We chose ten starting values randomly as described in [14], and the estimator $\hat{\theta}_{ML}$ was taken as the maximizer of the likelihood among the corresponding points of convergence.

3.2 Evaluation of the methods' clustering performance

For a simulated dataset $x_1, \ldots, x_n$ the true partition $\mathbf{z} = (z_1, \ldots, z_n)$ was known. The true partition was approximated by the soft partitions $\hat{\mathbf{z}}_{HM}$ and $\hat{\mathbf{z}}_{ML}$, calculated from the corresponding estimates $\hat{\theta}_{HM}$ and $\hat{\theta}_{ML}$, respectively, using Equation (13). To quantify the accuracy of an approximate partition, we used the *Fuzzy Adjusted Rand Index* (FARI) proposed in [3]. The FARI for $\mathbf{z}$ and its approximation $\hat{\mathbf{z}}$ – written as $\text{FARI}(\hat{\mathbf{z}}, \mathbf{z})$ – is a number in the interval $[-1, 1]$ measuring their closeness; the higher the index the better is the approximation $\hat{\mathbf{z}}$. A brief description of this index is given in Appendix A.3.

Let

$$\Delta_{FARI} = \text{FARI}(\mathbf{z}, \hat{\mathbf{z}}_{HM}) - \text{FARI}(\mathbf{z}, \hat{\mathbf{z}}_{ML})$$

denote the difference between the indices. Note that a positive difference $\Delta_{FARI} > 0$ implies that the partition obtained via the HM-estimator was more accurate than the partition obtained via the ML-estimator.

To determine if there was a significant difference between the methods' clustering performance, we applied the t-test and the sign-test to the pairwise differences $\Delta_{FARI}^{(1)}, \ldots, \Delta_{FARI}^{(500)}$ for the 500 simulated samples. We also made a comparison of the methods given that a difference in FARI under a certain threshold was considered as negligible, which was achieved by applying the sign-test to the differences that satisfied $|\Delta_{FARI}^{(i)}| > 0.1$.

The considered scenarios corresponded to problems that were more or less difficult with respect to clustering and as part of our evaluations we quantified these difficulties. Here $\hat{\mathbf{z}}_{opt}$ denotes the optimal partition obtained when the true component densities and parameter values in (13) were used. The index

$$\text{FARI}_{opt} = \text{FARI}(\mathbf{z}, \hat{\mathbf{z}}_{opt})$$

corresponded to the clustering performance obtained under correct model assumptions and a perfect estimator of θ. For each scenario, we used the mean of $\text{FARI}_{opt}^{(1)}, \ldots, \text{FARI}_{opt}^{(500)}$ for the 500 simulated samples to measure the difficulty of the problem and as a reference value for the corresponding FARI obtained by the HM- and ML-estimators.

3.3 Evaluation of the methods' ability to estimate the mixing proportion

We compared the methods in terms of their accuracy for estimating the mixing proportion p. Details on how we defined the point estimators of p are given in Appendix A.4.

The following standard characteristics for evaluating an estimator $\hat{p}$ of p based on N simulations were used:

$$\begin{aligned} mean &= \frac{1}{N}\sum_{i=1}^{N}\hat{p}^{(i)}, \\ \hat{bias} &= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}^{(i)} - p\right), \\ \hat{MSE} &= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}^{(i)} - p\right)^2. \end{aligned} \tag{14}$$

These characteristics were calculated for the simulated estimates $\{\hat{p}_{HM}^{(i)}\}_{i=1}^{N}$ and $\{\hat{p}_{ML}^{(i)}\}_{i=1}^{N}$ obtained from the HM- and ML-estimators, respectively.

To determine if there was a significant difference in efficiency between the methods, we applied the t-test to the difference in estimated mean squared error $\Delta_{MSE} = \hat{MSE}_{ML} - \hat{MSE}_{HM}$, where the subscripts HM and ML refer to the methods used in (14). Note that a positive value of Δ_{MSE} suggested that the hybrid method was more efficient as an estimator of p.

3.4 Results

This section includes a detailed treatment of the results for sample size $n = 50$. The results for sample sizes $n = 100$ and $n = 500$ led to similar conclusions, and are given in Appendix A.5.

The HM- and ML-estimators were evaluated mainly by their ability to cluster the samples in agreement with the true partition vector $\mathbf{z}$ and their ability to estimate the proportion parameter p. The corresponding estimators for the difference in mean $\mu_2 - \mu_1$ were compared in a similar way as for the mixing proportion p. This was done for the case when the data were generated from a normal mixture distribution (i.e. under correct model assumptions), a logistic mixture distribution (modest violation of model assumptions), a Laplace mixture distribution and contaminated Gaussian distribution (serious violation of model assumptions). For each case six values of the parameter vector were evaluated, Table 1 and Figure 2.

3.4.1 Clustering performance

The relative performance of the methods was evaluated by considering the mean difference in FARI ($\overline{\Delta}_{FARI}$), where a positive value indicated an advantage of the HM-estimator, the proportion of samples that were more accurately clustered by the HM-estimator than by the ML-estimator ($prop_{HM}$), and the proportion of the observed considerable differences in FARI which were in favor of the HM-estimator ($propC_{HM}$), see Section 3.2 for further details.

Scenarios (i) and (vi) were hard clustering problems in the sense that the mean optimal FARI was low in all the cases: $mean_{opt} \in [0.30, 0.58]$, whereas the other scenarios (ii, iii, iv, v) corresponded to relatively easy clustering problems: $mean_{opt} \in [0.60, 0.81]$, Table 2.

Table 2. The average clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples with 50 observations each were generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) was obtained for each sample and estimator. The mean FARI was observed for each scenario, and the mean of the optimal FARI (opt.) obtained using the true mixture distribution serves as a reference

Mean of fuzzy adjusted Rand index, $n = 50$

Mixture		*HM*	*ML*	*opt.*
Normal	(i)	0.28	0.21	0.35
	(ii)	0.59	0.52	0.68
	(iii)	0.54	0.56	0.70
	(iv)	0.57	0.53	0.60
	(v)	0.64	0.65	0.70
	(vi)	0.29	0.27	0.30
Logistic	(i)	0.41	0.24	0.46
	(ii)	0.71	0.60	0.71
	(iii)	0.65	0.52	0.71
	(iv)	0.64	0.53	0.65
	(v)	0.67	0.57	0.73
	(vi)	0.34	0.26	0.39
Laplace	(i)	0.32	0.22	0.40
	(ii)	0.66	0.57	0.70
	(iii)	0.59	0.54	0.70
	(iv)	0.60	0.53	0.63
	(v)	0.67	0.64	0.72
	(vi)	0.30	0.27	0.33
Contaminated	(i)	0.44	0.26	0.58
	(ii)	0.78	0.64	0.80
	(iii)	0.74	0.65	0.81
	(iv)	0.73	0.60	0.71
	(v)	0.75	0.67	0.70
	(vi)	0.39	0.33	0.42

In the case where data were generated from normal mixtures most of the observed differences were significant but of varying magnitude: $\overline{\Delta}_{FARI} \in [-0.02, 0.08]$, $prop_{HM} \in [0.35, 0.68]$ and $propC_{HM} \in [0.40, 0.92]$. The HM-estimator performed significantly better than the ML-estimator for scenarios (i, ii, iv, vi) and significantly worse for (iii, v), but the differences were rather small for (iii, v, vi), Figure 3 and Table 3.

With the data generated from logistic mixtures, the HM-estimator outperformed the ML-estimator for all parameter configurations, and most of the observed differences and evaluation measures were significant; $\overline{\Delta}_{FARI} \in [0.03, 0.10]$, $prop_{HM} \in [0.50, 0.76]$, and $propC_{HM} \in [0.61, 0.94]$, Figure 4 and Table 3. The largest differences were observed for scenarios (i, ii, iv), whereas the differences in scenarios (iii, v) were quite moderate. The magnitude of the differences were similar to those obtained for normal mixture data, but in this case all of them indicated an advantage of the HM-estimator.

In the case where the data were simulated from a mixture of Laplace or contaminated Gaussian distributions the HM-estimator outperformed the ML-estimator

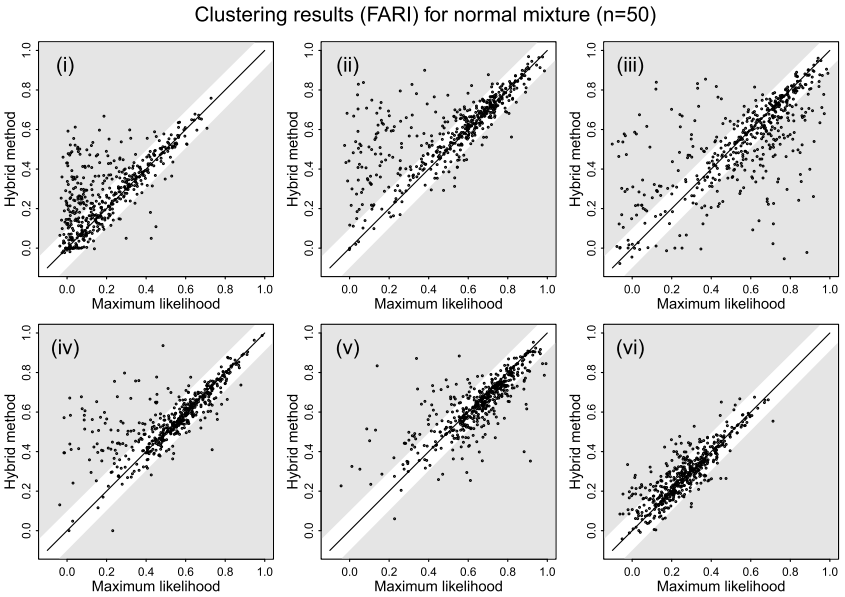


Fig. 3. The FARI observed for the hybrid and maximum likelihood methods. 500 samples each with 50 observations, are generated from normal mixture distributions with the parameter configurations (i)–(vi). Samples for which the hybrid method performs considerably better (worse) than the maximum likelihood estimator are in the upper (lower) shaded area. Points inside the white area mark samples that correspond to inconsiderable differences. A difference is regarded as considerable if the absolute difference in the methods' FARI exceeds 0.1

and all the observed differences were significant: $\overline{\Delta}_{FARI} \in [0.06, 0.17]$, $prop_{HM} \in [0.55, 0.85]$ and $propC_{HM} \in [0.72, 0.98]$, Figures 5–6 and Table 3. Overall the differences were more distinct than in the cases of normal and logistic mixtures.

Configuration (vii) defined a non-mixture distribution for which the desired result would be an average FARI value around zero and few high FARI values. Overall, both methods performed as expected and no clear differences between the methods were observed, with the exception that the ML method was more variable in the case $n = 500$, see Figures 8–10 in Appendix A.7.

3.4.2 Estimation of the proportion parameter p

The methods ability to estimate the mixing proportion p was evaluated using the mean, bias and mean squared error (MSE) for the corresponding point estimators of p, and their relative efficiency was analyzed via the estimated difference in MSE, see Section 3.3 for further details.

The HM-estimator (of the proportion parameter p) had lower MSE than the ML-estimator for almost all considered scenarios and most of the observed differences were significant, Table 4. The largest differences were observed when the model assumptions were seriously violated and the data were generated by a Laplace mixture and the smallest differences where observed for data generated by a normal mixture, Table 4. The observed MSE varied between the six parameter vectors, where scenar-

Table 3. The relative clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples each with 50 observations, are generated from four mixture distributions (normal, logistic, Laplace, and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) is observed for each sample and estimator. For each scenario we observe: the mean of the differences between the observed average FARI values for the HM- and ML-estimators ($\overline{\Delta}_{FARI}$) and the proportion of times the HM-estimator have a higher FARI value than the ML-estimator ($prop_{HM}$). A positive value of $\overline{\Delta}_{FARI}$ indicates a mean difference in favor of the HM-estimator. In the third column we observe the number of times n_{HM} (n_{ML}) the hybrid method performs considerably better (worse) than the maximum likelihood estimator. Here $propC_{HM}$ denotes the proportion of the samples with considerable differences for which the HM-estimator is superior. A difference is defined to be considerable if the distance between the methods' FARI is larger than 0.1. For each evaluation measure we test if the methods have the same average performance, the p-values relate to those tests

Comparison of HM and ML for soft clustering, $n = 50$

Mixture		$\overline{\Delta}_{FARI}$	p-value	$prop_{HM}$	p-value	$propC_{HM}$	p-value	n_{HM}	n_{ML}
Normal	(i)	0.07	0.00	0.67	0.00	0.92	0.00	143	13
	(ii)	0.08	0.00	0.68	0.00	0.86	0.00	126	20
	(iii)	−0.02	0.02	0.42	0.00	0.40	0.01	84	124
	(iv)	0.04	0.00	0.50	0.96	0.81	0.00	82	19
	(v)	−0.01	0.04	0.35	0.00	0.47	0.63	51	57
	(vi)	0.02	0.00	0.55	0.02	0.83	0.00	69	14
Logistic	(i)	0.10	0.00	0.76	0.00	0.94	0.00	200	14
	(ii)	0.08	0.00	0.68	0.00	0.91	0.00	127	12
	(iii)	0.05	0.00	0.50	0.89	0.61	0.00	138	89
	(iv)	0.07	0.00	0.65	0.00	0.94	0.00	136	8
	(v)	0.03	0.00	0.56	0.01	0.67	0.00	111	55
	(vi)	0.03	0.00	0.62	0.00	0.80	0.00	73	18
Laplace	(i)	0.17	0.00	0.85	0.00	0.98	0.00	268	5
	(ii)	0.12	0.00	0.72	0.00	0.97	0.00	147	4
	(iii)	0.13	0.00	0.62	0.00	0.79	0.00	184	49
	(iv)	0.12	0.00	0.73	0.00	0.98	0.00	175	4
	(v)	0.10	0.00	0.67	0.00	0.81	0.00	192	45
	(vi)	0.08	0.00	0.78	0.00	0.89	0.00	161	21
Contaminated	(i)	0.17	0.00	0.81	0.00	0.96	0.00	264	11
	(ii)	0.14	0.00	0.71	0.00	0.93	0.00	174	13
	(iii)	0.08	0.00	0.55	0.04	0.72	0.00	161	61
	(iv)	0.12	0.00	0.75	0.00	0.94	0.00	212	14
	(v)	0.08	0.00	0.66	0.00	0.85	0.00	182	31
	(vi)	0.06	0.00	0.65	0.00	0.82	0.00	157	34

ios (i) and (vi) had the highest MSE-values. Recall that these scenarios were the most difficult ones in terms of clustering. Furthermore, the advantage of the HM-estimator was most prominent for scenario (i) which also is in agreement with the clustering results. Investigating the precision of the methods via the magnitude of the observed bias revealed that the ML-estimator was more precise than the HM-estimator when the distributional assumption was valid and that the methods had similar precision when the assumptions were violated, Table 4. The results obtained for estimating the difference in mean $\mu_2 - \mu_1$ resembled the results obtained for the mixing proportion p, see Tables 11–13 in Appendix A.6.

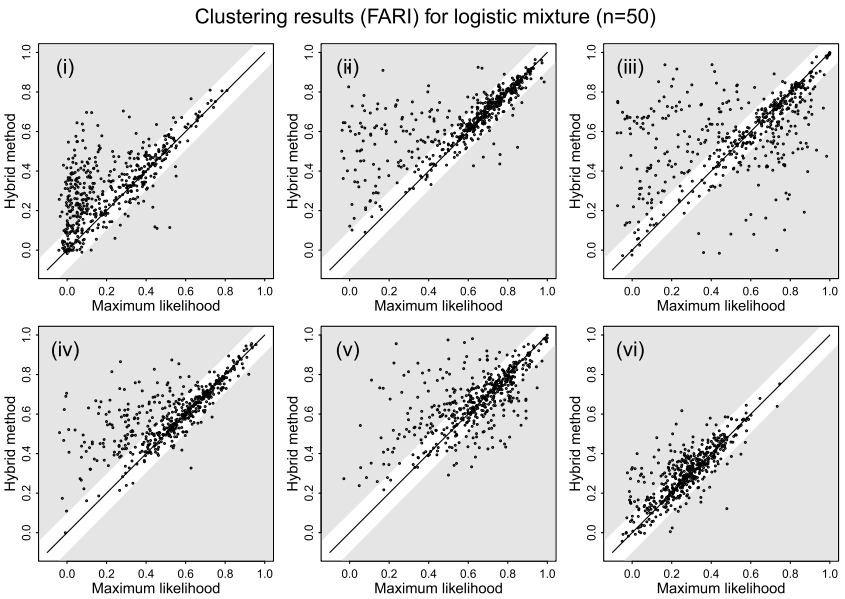


Fig. 4. The FARI observed for the hybrid and maximum likelihood methods. 500 samples each with 50 observations, are generated from logistic mixture distributions with the parameter configurations (i)–(vi). Samples for which the hybrid method performs considerably better (worse) than the maximum likelihood estimator are in the upper (lower) shaded area. Points inside the white area mark samples that correspond to inconsiderable differences. A difference is regarded as considerable if the absolute difference in the methods' FARI exceeds 0.1

4 Case study: clustering of acute leukemia data

4.1 Description of the data

In [13] a microarray experiment on human mRNA samples for measuring gene expression levels in two subtypes of acute leukemia is described, namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The experiment contained $n = 72$ samples, of which 47 were of type ALL and 25 were of type AML, and the expression levels of 6,817 genes were observed. In this study we used the preprocessed and filtered version of this data considered in [8], which contains the expression values of 3,571 genes.

4.2 Identification of differentially expressed genes

We applied a supervised procedure to get a subset of the 3,571 genes which were expressed differently with respect to the ALL/AML grouping. For each gene i we calculated the normal mixture clustering $\hat{\mathbf{z}}^{(i)}$ in (13) with parameter estimates obtained under known group memberships (i.e. the sample means and variances). Then we used (the measure) $\mathrm{FARI}(\mathbf{z}, \hat{\mathbf{z}}^{(i)})$, where $\mathbf{z}$ is the true ALL/AML grouping expressed as a 0–1 vector, to quantify the extent to which the mean expression of gene i differs between the groups. The 342 genes that met the criterion

$$\mathrm{FARI}\big(\mathbf{z}, \hat{\mathbf{z}}^{(i)}\big) \geq 0.1$$

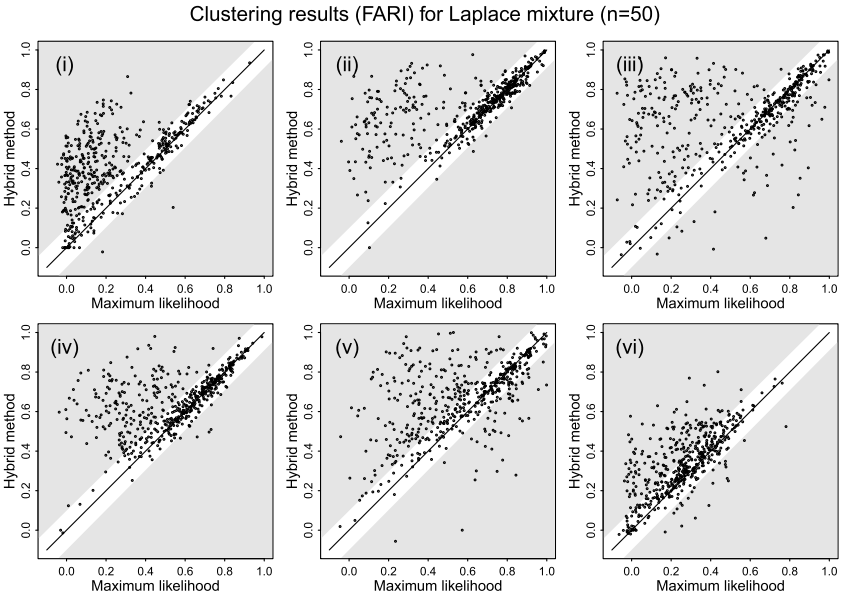


Fig. 5. The FARI observed for the hybrid and maximum likelihood methods. 500 samples each with 50 observations, are generated from Laplace mixture distributions with the parameter configurations (i)–(vi). Samples for which the hybrid method performs considerably better (worse) than the maximum likelihood estimator are in the upper (lower) shaded area. Points inside the white area mark samples that correspond to inconsiderable differences. A difference is regarded as considerable if the absolute difference in the methods' FARI exceeds 0.1

were regarded as truly differently expressed genes and included in our test set. Alternatively, we could have used the two-sample t-test to make this selection.

We applied the HM and ML clustering methods to each of the 342 test variables to compare their ability to divide the 72 cancer samples into the ALL and AML groups. The analysis was carried out as described in Section 3.2.

4.3 Results

Independent of method the overall performance was rather poor; most of the clusterings had a FARI between 0 and 0.4, Figure 7. This implies that it is hard to cluster the samples accurately based on single genes.

The observed differences between the HM and ML methods were all significant and in favor of the hybrid method. For 213 of the 342 test genes the HM method clustered the samples more accurately than the ML method (i.e. $prop_{HM} = 0.623$, p-value $< 10^{-5}$). The mean difference in FARI (i.e. $\overline{\Delta}_{FARI}$) was 0.020 (p-value $< 10^{-5}$). Moreover, of the 38 genes for which there was a *considerable difference* between the methods (i.e. an absolute difference larger than 0.1), the HM clustering was superior over the ML clustering in 32 of the cases (i.e. $propC_{HM} = 0.842$, p-value $< 10^{-4}$). For notation, see Section 3.4.1.

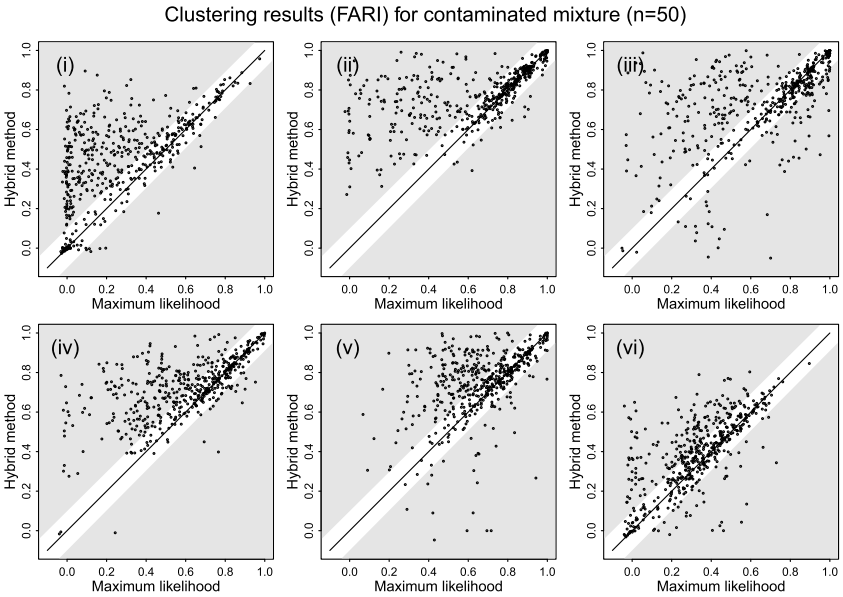


Fig. 6. The FARI observed for the hybrid and maximum likelihood methods. 500 samples each with 50 observations, are generated from contaminated Gaussian mixture distributions with the parameter configurations (i)–(vi). Samples for which the hybrid method performs considerably better (worse) than the maximum likelihood estimator are in the upper (lower) shaded area. Points inside the white area mark samples that correspond to inconsiderable differences. A difference is regarded as considerable if the absolute difference in the methods' FARI exceeds 0.1

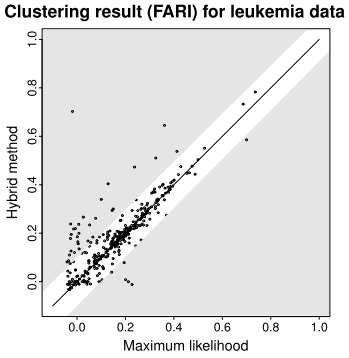


Fig. 7. Clustering results for the cancer data. The fuzzy adjusted Rand indices (FARI) observed for the hybrid and maximum likelihood methods. Data were taken from a microarray experiment on gene expression levels in two types of acute leukemia: ALL and AML. 342 genes were measured across 72 samples. Genes for which the hybrid method performed considerably better (worse) than the maximum likelihood estimator are in the upper (lower) shaded area. Here a difference was defined to be considerable if the absolute difference in FARI between the methods was larger than 0.1

Table 4. The accuracy of the HM- and ML-estimators with regard to estimating the proportion parameter p. 500 samples each with 50 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of p

Estimation of the mixing proportion, $n = 50$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		p	HM	ML	HM	ML	HM	ML		
Normal	(i)	0.50	0.543	0.483	0.043	−0.017	0.052	0.084	0.033	0.000
	(ii)	0.50	0.517	0.508	0.017	0.008	0.017	0.036	0.019	0.000
	(iii)	0.25	0.331	0.280	0.081	0.030	0.028	0.027	−0.000	0.889
	(iv)	0.50	0.471	0.481	−0.029	−0.019	0.014	0.023	0.009	0.000
	(v)	0.25	0.238	0.246	−0.012	−0.004	0.011	0.013	0.002	0.104
	(vi)	0.50	0.362	0.423	−0.138	−0.077	0.040	0.041	0.001	0.617
Logistic	(i)	0.50	0.553	0.518	0.053	0.018	0.043	0.081	0.037	0.000
	(ii)	0.50	0.504	0.503	0.004	0.003	0.013	0.029	0.016	0.000
	(iii)	0.25	0.314	0.338	0.064	0.088	0.022	0.040	0.017	0.000
	(iv)	0.50	0.488	0.528	−0.012	0.028	0.012	0.024	0.012	0.000
	(v)	0.25	0.254	0.297	0.004	0.047	0.010	0.020	0.011	0.000
	(vi)	0.50	0.400	0.489	−0.100	−0.011	0.034	0.035	0.001	0.627
Laplace	(i)	0.50	0.533	0.511	0.033	0.011	0.029	0.074	0.044	0.000
	(ii)	0.50	0.484	0.498	−0.016	−0.002	0.011	0.025	0.015	0.000
	(iii)	0.25	0.295	0.383	0.045	0.133	0.014	0.048	0.034	0.000
	(iv)	0.50	0.491	0.561	−0.009	0.061	0.011	0.025	0.014	0.000
	(v)	0.25	0.280	0.371	0.030	0.121	0.012	0.035	0.024	0.000
	(vi)	0.50	0.445	0.536	−0.055	0.036	0.029	0.050	0.021	0.000
Contaminated	(i)	0.50	0.517	0.491	0.017	−0.009	0.046	0.083	0.037	0.000
	(ii)	0.50	0.488	0.502	−0.012	0.002	0.008	0.025	0.016	0.000
	(iii)	0.25	0.276	0.362	0.026	0.112	0.011	0.036	0.024	0.000
	(iv)	0.50	0.491	0.563	−0.009	0.063	0.011	0.024	0.014	0.000
	(v)	0.25	0.257	0.359	0.007	0.109	0.009	0.024	0.014	0.000
	(vi)	0.50	0.426	0.527	−0.074	0.027	0.033	0.040	0.006	0.051

5 Discussion and conclusion

We consider a univariate cluster problem, which arises in many applications, where the data are generated from a mixture distribution with two components and where the aim is to group samples of the same type. This problem is commonly solved using the EM-algorithm based on the assumption that the observations are generated by a mixture of two normal densities. Although this is a powerful method it is also sensitive to incorrectly specified distributions. Furthermore, the assumption that data approx-

imately follow a normal mixture is rather restrictive, which makes the EM-approach unfeasible in many applications.

The use of hybrid methods in mixture problems is, to the best of our knowledge, rather unexplored. The variant we propose can be motivated as follows: the method of moments is general in the sense that the parametric family can be left unspecified, it is enough to assume that the component densities are symmetric and have finite moments, and the minimum distance method is robust against symmetric departures from the assumed normal mixture distribution.

The results suggest that the proposed HM-estimator has a considerably better ability to cluster the samples than the ML-estimator, in particular if the assumption of a normal mixture is incorrect. This result is observed for both simulated and real data, and holds independently of the sample size. A slight advantage for the HM-estimator is also observed in the case where the Gaussian mixture assumption is valid.

We also consider estimation of the mixing proportion p, a problem that has attracted much interest in the literature [20]. Our results show that the HM-estimator is more robust and efficient than the ML-estimator for estimating p for a wide range of mixture distributions and sample sizes. This is consistent with several related studies on minimum distance inference for p [6, 5].

It should be noted that the HM-estimator can easily be adapted to any parametric mixture of symmetric densities, not just the normal mixture distribution. Furthermore, we can consider a less restrictive assumption that allows the components distributions to be of several types. For example, we may use the composite assumption that the data are generated by a mixture of two normal distributions, the mixture of two Laplace distributions, or the mixture of one normal distribution and one Laplace distribution. In this case parameter estimates can be obtained via the proposed hybrid method, either by extending the distance function, or by deriving the HM-estimator for each assumed mixture distribution and take the estimator with the best fit to the empirical data to be used as the final estimator. Further studies are needed to show that this approach is reasonable.

A general drawback with the method of moments is that the estimating equations sometimes lack solutions, and our variant is not an exception. However, this problem is usually overlooked and does not seem to be of practical importance, see [26] for a discussion. Another concern in our case is when there are no relevant estimates close to the true parameter vector. For example, there are no solutions of the moment equations with $\hat{p} = 1/2$ and $\hat{\sigma}_1 = \hat{\sigma}_2$, regardless of the data values. This issue did not seem to have a major impact in our simulation study where cases with $p = 1/2$ and $\sigma_1 = \sigma_2$ are included, but should be considered in future studies.

We use the FARI to evaluate the performance of the clustering methods, the reason for this is that FARI has a higher resolution than the ordinary adjusted Rand index, and is therefore better to separate approaches for which the clustering performance is relatively similar.

We propose to robustify the hybrid estimator by using trimmed (5% removed) versions of the sample moments. This is to enable high performance also in the presence of outliers, which are often encountered in real datasets and modeled here by the Laplace and contaminated Gaussian mixtures. For some of our simulations we have applied the HM method without trimming. Overall the results are usually better

when we apply the 5% trimming, but there are some exceptions (data not shown). Moreover, one could argue that the ML-estimator may perform better if some of the extreme observations are removed prior to the estimation. The 5% trimming used in our simulations is merely for illustration and should not be taken as a general recommendation; how to choose the trimming level on the basis of data is a topic for future research.

In most applications several variables are observed and the common practice is to base the clustering on all, or at least several, of the observed variables. For high-dimensional genomic data this type of approaches has been shown to be difficult and non-informative variables need to be removed in order to have success [10]. It would be interesting to derive a variable selection procedure that utilizes the robustness of the hybrid approach for selecting informative variables in high-dimensional unsupervised classification problems. An interesting generalization of this work is to adopt it to the case were several variables are observed.

To conclude, the proposed moment-distance hybrid method has good clustering performance, is robust against incorrect model assumptions and can easily be applied to a wide range of problems.

Funding

This work was supported by grants from the Swedish Research Council (P.R.), Dnr 340-2013-5185 (P.R.), the Kempe Foundations (D.K., P.R.), Dnr JCK-1315, and the Faculty of Science and Technology, Umeå University (P.R.).

Appendix

A.1 Theoretical moments of the mixture

If $f_i(\cdot)$ is the density of a random variable $X_i, i = 1, 2$, the theoretical moments of a random variable X with mixture density $f(\cdot) = pf_1(\cdot) + (1 - p)f_2(\cdot)$ can be expressed as

$$E\big(X^k\big) = pE\big(X_1^k\big) + (1 - p)E\big(X_2^k\big), \quad k = 1, 2, \ldots.$$

This combined with the trivial relations $E(X_i) = \mu_i$ and $E(X_i^2) = \sigma_i^2 + \mu_i^2$ leads to the first two moments of X, (1) and (2). For the cubic moment, we use the symmetry of the density $f_i(\cdot)$ around its mean μ_i, which yields a vanishing third central moment, i.e $E(X_i - \mu_i)^3 = 0$. From this and some algebra, it follows that

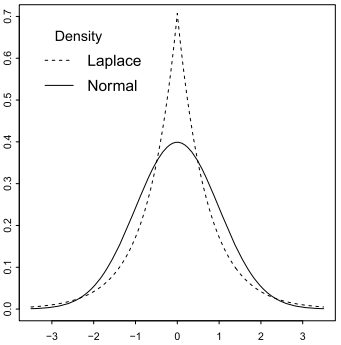
$$\begin{aligned} E\big(X_i^3\big) &= E\big((X_i - \mu_i + \mu_i)^3\big) \\ &= E\big((X_i - \mu_i)^3\big) + 3E\big((X_i - \mu_i)^2\big)\mu_i + 3E(X_i - \mu_i)\mu_i^2 + \mu_i^3 \\ &= 3\sigma_i^2\mu_i + \mu_i^3. \end{aligned}$$

A.2 The Laplace, logistic, and contaminated Gaussian distributions

The **Laplace distribution** with mean μ and shape parameter $b > 0$ has the density

$$g(x|\mu, b) = \frac{1}{2b} e^{-|x-\mu|/b}, \quad x \in \mathbb{R}.$$

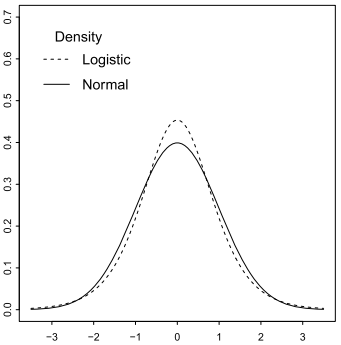
The variance is $\sigma^2 = b/\sqrt{2}$.



The density of the **Logistic distribution** with mean μ and shape parameter $b > 0$ is given by

$$g(x|\mu, b) = \frac{e^{\frac{x-\mu}{b}}}{b(1 + e^{-\frac{x-\mu}{b}})^2}, \quad x \in \mathbb{R}.$$

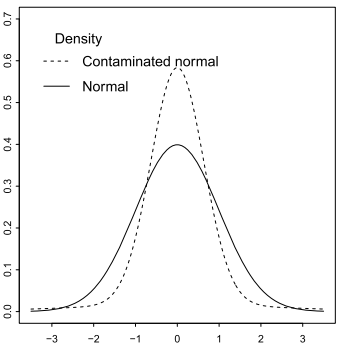
The variance is $\sigma^2 = b/\sqrt{2}$.



The **contaminated Gaussian distribution** is a two-component normal variance mixture where one of the components, with mean μ and variance σ^2 has a large prior probability, denoted $\alpha \in (0, 1)$, and represents "good" observations, and the other component has the same mean but $\eta > 1$ times larger variance and represents "bad" observations [12]. The density is given by

$$g\big(x; \mu, \sigma^2, \alpha, \eta\big) = \alpha\phi\big(x; \mu, \sigma^2\big) + (1 - \alpha)\phi\big(x; \mu, \eta\sigma^2\big),$$

where $\phi(x; \mu, \sigma^2), x \in \mathbb{R}$, is the normal density. The variance is $[\alpha + \eta(1 - \alpha)]\sigma^2$.



A.3 The Rand indices for measuring similarity of partitions

The material in this section is based on the paper [3], to which we refer for more details.

Hard partitions

Let $E = \{e_1, \ldots, e_n\}$ be a set of n elements that are to be partitioned in two groups. A partition can be identified by a labeling vector $\mathbf{z} = (z_1, \ldots, z_n)$, where z_i is either 0 or 1, $i = 1, \ldots, n$. Two elements e_i and e_j are assigned to the same group by the partition $\mathbf{z}$ if $z_i = z_j$.

Let $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ be two partitions of E. Often it is of interest to quantify their similarity, for example, when $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are obtained by two different clustering methods. Two common measures of closeness between partitions $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are defined via the following characteristics:

a = the number of pairs in E assigned to the same group both by $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$.
b = the number of pairs in E assigned to different groups both by $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$.
c = the number of pairs in E assigned to the same group by $\mathbf{z}^{(1)}$ but to different groups by $\mathbf{z}^{(2)}$.
d = the number of pairs in E assigned to different groups by $\mathbf{z}^{(1)}$ but to the same group by $\mathbf{z}^{(2)}$.

The Rand index (RI) for $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ is defined as

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}. \tag{15}$$

It lies in the interval [0, 1], where 0 indicates that $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ do not agree on any pair of elements and 1 means that $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ coincide. The adjusted rand index (ARI) is a corrected version of RI that has an expected value of 0 for randomly sampled partitions:

$$ARI = \frac{2(ad-bc)}{c^2b^2 + 2ad + (a+d)(c+b)}. \tag{16}$$

The ARI attains values in the interval $[-1, 1]$.

Next we give a more formal definition of the numbers a, b, c, and d. Two elements e_i and e_j are said to be *bonding* in a partition $\mathbf{z}$ that puts them in the same group. To each partition $\mathbf{z}$ there is a bonding matrix $\mathbf{B}$ with elements

$$B_{ij} = \begin{cases} 1, & \text{if } e_i \text{ and } e_j \text{ are bonding in } \mathbf{z} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

For a n-dimensional square matrix $\mathbf{X}$, we introduce the function

$$h(\mathbf{X}) = \frac{1}{2}\sum_{i,j} X_{ij},$$

Now let $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ be the bonding matrices for the partitions $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$, respectively. Then, if we let $\times$ denote element-wise multiplication between matrices, the numbers a, b, c, and d can be expressed in terms of the bonding matrices as

$$\begin{aligned} a &= h\big(\mathbf{B}^{(1)} \times \mathbf{B}^{(2)}\big) - \frac{n}{2}, \\ b &= h\big(\big(1 - \mathbf{B}^{(1)}\big) \times \big(1 - \mathbf{B}^{(2)}\big)\big), \\ c &= h\big(\mathbf{B}^{(1)} \times \big(1 - \mathbf{B}^{(2)}\big)\big), \\ d &= h\big(\big(1 - \mathbf{B}^{(1)}\big) \times \mathbf{B}^{(2)}\big). \end{aligned}$$

Fuzzy partitions

The partitioning considered in the previous section is called *hard* (or *crisp*) and is a special case of a more general concept. A *fuzzy* (or *soft*) partition of the set $E = \{e_1, \ldots, e_n\}$ into two groups is represented by a vector $\mathbf{z} = (z_1, \ldots, z_n)$ with $0 \leq z_i \leq 1$. The pair $(z_i, 1 - z_i)$ gives the degree to which the element e_i is a member of the two groups. Note that a hard partition is also fuzzy.

Next we introduce indexes of similarity between two fuzzy partitions $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ that give the same results as RI and ARI whenever $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ are hard. An extended definition of the bonding matrix $\mathbf{B}$ will be used. Let us measure the degree of bonding B_{ij} between two elements e_i and e_j in a fuzzy partition $\mathbf{z}$ with the *cosine similarity* between the vectors $(z_i, 1 - z_i)$ and $(z_j, 1 - z_j)$:

$$B_{ij} = \frac{z_i z_j + (1 - z_i)(1 - z_j)}{\sqrt{(z_i^2 + (1 - z_i)^2)(z_j^2 + (1 - z_j)^2)}}.$$

Let $\mathbf{B}$ denote the corresponding bonding matrix. It is easy to check that this coincides with definition (17) of the bonding matrix for a hard partition. Now we use representation (17) of a, b, c, and d with the extended definition of a bonding matrix. The generalizations of (15) and (16) follow directly and are called *the fuzzy Rand index* (FRI) and *fuzzy adjusted Rand index* (FARI), respectively. The FRI and FARI measure similarity between two fuzzy partitions, and attain values in the intervals [0, 1] and [−1, 1], respectively.

A.4 Point estimator of the mixing proportion

Despite the notation $\hat{\theta} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p})$ for an estimator of θ, we have to specify what we mean by a point estimator. Note that since $\theta = \theta^{(1)} = (\mu_1, \mu_2, \sigma_2^2, \sigma_1^2, p)$ and $\theta^{(2)} = (\mu_2, \mu_1, \sigma_2^2, \sigma_1^2, 1 - p)$ define the same distribution when the components $f_1(\cdot)$ and $f_2(\cdot)$ belong to the same family, it cannot be said which of $\theta^{(1)}$ and $\theta^{(2)}$ is estimated by $\hat{\theta}$. This implies that $\hat{\theta}$ is not a point estimator in the strict sense, and that it is unclear whether $\hat{p}$ estimates p or $1 - p$. To resolve this ambiguity we assume without loss of generality that $\mu_1 < \mu_2$ and let the estimator $\hat{\theta}$ satisfy $\hat{\mu}_1 < \hat{\mu}_2$, meaning that $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p})$ is replaced with $(\hat{\mu}_2, \hat{\mu}_1, \hat{\sigma}_2^2, \hat{\sigma}_1^2, 1 - \hat{p})$ whenever $\hat{\mu}_1 > \hat{\mu}_2$. We claim that this approach of defining point estimators of θ and p is reliable when the mean separation $|\mu_1 - \mu_2|$ between the components is not too small.

A.5 Results for sample sizes $n = 100$ and $n = 500$

A.5.1 Clustering performance

Table 5. The average clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples with 100 observations each were generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) was obtained for each sample and estimator. The mean FARI was observed for each scenario, and the mean of the optimal FARI (opt.) obtained using the true mixture distribution serves as a reference

Mean of fuzzy adjusted Rand index, $n = 100$

Data		*HM*	*ML*	*opt.*
Normal	(i)	0.27	0.19	0.35
	(ii)	0.64	0.61	0.69
	(iii)	0.56	0.60	0.71
	(iv)	0.57	0.55	0.60
	(v)	0.67	0.68	0.71
	(vi)	0.29	0.27	0.29
Logistic	(i)	0.43	0.19	0.45
	(ii)	0.73	0.59	0.70
	(iii)	0.68	0.47	0.71
	(iv)	0.65	0.49	0.65
	(v)	0.67	0.53	0.73
	(vi)	0.35	0.26	0.40
Laplace	(i)	0.33	0.19	0.39
	(ii)	0.69	0.62	0.70
	(iii)	0.65	0.58	0.71
	(iv)	0.60	0.52	0.63
	(v)	0.69	0.65	0.72
	(vi)	0.32	0.28	0.33
Contaminated	(i)	0.48	0.20	0.59
	(ii)	0.81	0.64	0.81
	(iii)	0.78	0.61	0.80
	(iv)	0.76	0.54	0.72
	(v)	0.78	0.64	0.70
	(vi)	0.40	0.30	0.42

Table 6. The relative clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples each with 100 observations, are generated from four mixture distributions (normal, logistic, Laplace, and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) is observed for each sample and estimator. For each scenario we observe: the mean of the differences between the observed average FARI values for the HM- and ML-estimators ($\overline{\Delta}_{FARI}$) and the proportion of times the HM-estimator have a higher FARI value than the ML-estimator ($prop_{HM}$). A positive value of $\overline{\Delta}_{FARI}$ indicates a mean difference in favor of the HM-estimator. In the third column we observe the number of times n_{HM} (n_{ML}) the hybrid method performs considerably better (worse) than the maximum likelihood estimator. Here $propC_{HM}$ denotes the proportion of the samples with considerable differences for which the HM-estimator is superior. A difference is defined to be considerable if the distance between the methods' FARI is larger than 0.1. For each evaluation measure we test if the methods have the same average performance, the p-values relate to those tests

Comparison of HM and ML for soft clustering, $n = 100$

Dist.		$\overline{\Delta}_{FARI}$	p-value	$prop_{HM}$	p-value	$propC_{HM}$	p-value	n_{HM}	n_{ML}
Normal	(i)	0.09	0.00	0.74	0.00	0.95	0.00	187	10
	(ii)	0.03	0.00	0.67	0.00	0.77	0.00	58	17
	(iii)	−0.04	0.00	0.36	0.00	0.32	0.00	69	146
	(iv)	0.02	0.00	0.53	0.20	0.98	0.00	57	1
	(v)	−0.01	0.03	0.37	0.00	0.57	0.41	30	23
	(vi)	0.02	0.00	0.64	0.00	0.91	0.00	40	4
Logistic	(i)	0.24	0.00	0.94	0.00	0.99	0.00	349	2
	(ii)	0.14	0.00	0.82	0.00	0.99	0.00	144	1
	(iii)	0.21	0.00	0.77	0.00	0.93	0.00	246	20
	(iv)	0.16	0.00	0.84	0.00	1.00	0.00	258	0
	(v)	0.14	0.00	0.77	0.00	0.94	0.00	263	17
	(vi)	0.10	0.00	0.90	0.00	0.98	0.00	193	3
Laplace	(i)	0.14	0.00	0.83	0.00	0.95	0.00	256	12
	(ii)	0.08	0.00	0.73	0.00	0.93	0.00	90	7
	(iii)	0.07	0.00	0.54	0.05	0.66	0.00	142	73
	(iv)	0.07	0.00	0.71	0.00	0.96	0.00	157	7
	(v)	0.04	0.00	0.52	0.30	0.80	0.00	114	28
	(vi)	0.04	0.00	0.75	0.00	1.00	0.00	80	0
Contaminated	(i)	0.28	0.00	0.92	0.00	0.99	0.00	375	5
	(ii)	0.17	0.00	0.84	0.00	1.00	0.00	192	0
	(iii)	0.17	0.00	0.74	0.00	0.92	0.00	279	25
	(iv)	0.23	0.00	0.91	0.00	1.00	0.00	366	2
	(v)	0.14	0.00	0.84	0.00	0.96	0.00	277	13
	(vi)	0.10	0.00	0.84	0.00	0.93	0.00	174	14

Table 7. The average clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples with 500 observations each were generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) was obtained for each sample and estimator. The mean FARI was observed for each scenario, and the mean of the optimal FARI (opt.) obtained using the true mixture distribution serves as a reference

Mean of fuzzy adjusted Rand index, $n = 500$

Data		*HM*	*ML*	*opt.*
Normal	(i)	0.27	0.25	0.35
	(ii)	0.67	0.67	0.68
	(iii)	0.54	0.69	0.71
	(iv)	0.57	0.59	0.60
	(v)	0.69	0.71	0.71
	(vi)	0.32	0.30	0.30
Logistic	(i)	0.45	0.07	0.45
	(ii)	0.75	0.66	0.70
	(iii)	0.74	0.34	0.71
	(iv)	0.64	0.41	0.65
	(v)	0.67	0.48	0.73
	(vi)	0.38	0.28	0.39
Laplace	(i)	0.33	0.16	0.39
	(ii)	0.71	0.70	0.70
	(iii)	0.65	0.62	0.71
	(iv)	0.59	0.52	0.63
	(v)	0.69	0.63	0.72
	(vi)	0.34	0.30	0.33
Contaminated	(i)	0.48	0.04	0.59
	(ii)	0.84	0.62	0.81
	(iii)	0.81	0.49	0.80
	(iv)	0.76	0.49	0.71
	(v)	0.77	0.62	0.71
	(vi)	0.45	0.36	0.42

Table 8. The relative clustering performance of the hybrid method (HM) and the maximum likelihood (ML) method. 500 samples each with 500 observations, are generated from four mixture distributions (normal, logistic, Laplace, and contaminated Gaussian) with the parameter configurations (i)–(vi). The fuzzy adjusted Rand index (FARI) is observed for each sample and estimator. For each scenario we observe: the mean of the differences between the observed average FARI values for the HM- and ML-estimators ($\overline{\Delta}_{FARI}$) and the proportion of times the HM-estimator have a higher FARI value than the ML-estimator ($prop_{HM}$). A positive value of $\overline{\Delta}_{FARI}$ indicates a mean difference in favor of the HM-estimator. In the third column we observe the number of times n_{HM} (n_{ML}) the hybrid method performs considerably better (worse) than the maximum likelihood estimator. Here $propC_{HM}$ denotes the proportion of the samples with considerable differences for which the HM-estimator is superior. A difference is defined to be considerable if the distance between the methods' FARI is larger than 0.1. For each evaluation measure we test if the methods have the same average performance, the p-values relate to those tests

Comparison of HM and ML for soft clustering, $n = 500$

Data		$\overline{\Delta}_{FARI}$	p-value	$prop_{HM}$	p-value	$propC_{HM}$	p-value	n_{HM}	n_{ML}
Normal	(i)	0.02	0.00	0.47	0.14	0.70	0.00	112	48
	(ii)	0.00	0.01	0.58	0.00	0.40	1.00	2	3
	(iii)	−0.15	0.00	0.05	0.00	0.02	0.00	7	364
	(iv)	−0.01	0.00	0.21	0.00	0.80	0.38	4	1
	(v)	−0.01	0.00	0.18	0.00	0.00	1.00	0	1
	(vi)	0.02	0.00	0.90	0.00	–	–	0	0
Logistic	(i)	0.38	0.00	0.99	0.00	1.00	0.00	474	0
	(ii)	0.09	0.00	0.85	0.00	1.00	0.00	68	0
	(iii)	0.40	0.00	0.91	0.00	0.99	0.00	409	2
	(iv)	0.22	0.00	0.97	0.00	1.00	0.00	438	0
	(v)	0.18	0.00	0.99	0.00	1.00	0.00	464	0
	(vi)	0.10	0.00	1.00	0.00	1.00	0.00	229	0
Laplace	(i)	0.17	0.00	0.80	0.00	0.96	0.00	314	12
	(ii)	0.01	0.00	0.69	0.00	0.83	0.22	5	1
	(iii)	0.03	0.00	0.39	0.00	0.49	0.90	122	125
	(iv)	0.08	0.00	0.73	0.00	0.99	0.00	206	1
	(v)	0.06	0.00	0.79	0.00	1.00	0.00	156	0
	(vi)	0.05	0.00	1.00	0.00	1.00	0.01	8	0
Contaminated	(i)	0.44	0.00	0.99	0.00	1.00	0.00	495	0
	(ii)	0.22	0.00	0.95	0.00	1.00	0.00	230	0
	(iii)	0.32	0.00	0.96	0.00	1.00	0.00	461	2
	(iv)	0.26	0.00	1.00	0.00	1.00	0.00	496	0
	(v)	0.15	0.00	1.00	0.00	1.00	0.00	416	0
	(vi)	0.09	0.00	1.00	0.00	1.00	0.00	136	0

A.5.2 Estimation of the proportion parameter p

Table 9. The accuracy of the HM- and ML-estimators with regard to estimating the proportion parameter p. 500 samples each with 100 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of p

Estimation of the mixing proportion, $n = 100$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		p	HM	ML	HM	ML	HM	ML		
Normal	(i)	0.50	0.594	0.524	0.094	0.024	0.049	0.088	0.040	0.000
	(ii)	0.50	0.524	0.503	0.024	0.003	0.010	0.015	0.006	0.000
	(iii)	0.25	0.358	0.287	0.108	0.037	0.026	0.020	−0.007	0.005
	(iv)	0.50	0.489	0.488	−0.011	−0.012	0.006	0.011	0.005	0.000
	(v)	0.25	0.257	0.258	0.007	0.008	0.005	0.007	0.001	0.041
	(vi)	0.50	0.365	0.438	−0.135	−0.062	0.029	0.025	−0.004	0.080
Logistic	(i)	0.50	0.548	0.478	0.048	−0.022	0.020	0.094	0.074	0.000
	(ii)	0.50	0.496	0.505	−0.004	0.005	0.005	0.022	0.018	0.000
	(iii)	0.25	0.290	0.427	0.040	0.177	0.011	0.064	0.053	0.000
	(iv)	0.50	0.509	0.610	0.009	0.110	0.006	0.026	0.020	0.000
	(v)	0.25	0.284	0.404	0.034	0.154	0.008	0.039	0.031	0.000
	(vi)	0.50	0.485	0.592	−0.015	0.092	0.016	0.042	0.027	0.000
Laplace	(i)	0.50	0.584	0.506	0.084	0.006	0.033	0.090	0.057	0.000
	(ii)	0.50	0.511	0.500	0.011	−0.000	0.007	0.020	0.013	0.000
	(iii)	0.25	0.307	0.344	0.057	0.094	0.014	0.037	0.023	0.000
	(iv)	0.50	0.509	0.556	0.009	0.056	0.006	0.016	0.010	0.000
	(v)	0.25	0.264	0.308	0.014	0.058	0.005	0.014	0.009	0.000
	(vi)	0.50	0.412	0.516	−0.088	0.016	0.020	0.020	0.000	0.979
Contaminated	(i)	0.50	0.548	0.505	0.048	0.005	0.025	0.108	0.083	0.000
	(ii)	0.50	0.494	0.508	−0.006	0.008	0.004	0.022	0.018	0.000
	(iii)	0.25	0.266	0.398	0.016	0.148	0.006	0.038	0.032	0.000
	(iv)	0.50	0.499	0.610	−0.001	0.110	0.005	0.031	0.026	0.000
	(v)	0.25	0.259	0.381	0.009	0.131	0.005	0.025	0.020	0.000
	(vi)	0.50	0.469	0.555	−0.031	0.055	0.018	0.043	0.026	0.000

Table 10. The accuracy of the HM- and ML-estimators with regard to estimating the proportion parameter p. 500 samples each with 500 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of p

Estimation of the mixing proportion, $n = 500$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		p	HM	ML	HM	ML	HM	ML		
Normal	(i)	0.50	0.671	0.499	0.171	−0.001	0.037	0.047	0.009	0.001
	(ii)	0.50	0.530	0.497	0.030	−0.003	0.003	0.002	−0.001	0.002
	(iii)	0.25	0.403	0.260	0.153	0.010	0.026	0.004	−0.023	0.000
	(iv)	0.50	0.524	0.501	0.024	0.001	0.002	0.003	0.001	0.001
	(v)	0.25	0.255	0.253	0.005	0.003	0.001	0.002	0.001	0.000
	(vi)	0.50	0.386	0.490	−0.114	−0.010	0.016	0.003	−0.013	0.000
Logistic	(i)	0.50	0.579	0.500	0.079	0.000	0.009	0.140	0.130	0.000
	(ii)	0.50	0.502	0.491	0.002	−0.009	0.001	0.012	0.011	0.000
	(iii)	0.25	0.267	0.537	0.017	0.287	0.002	0.101	0.099	0.000
	(iv)	0.50	0.523	0.659	0.023	0.159	0.002	0.029	0.027	0.000
	(v)	0.25	0.294	0.439	0.044	0.189	0.003	0.038	0.034	0.000
	(vi)	0.50	0.507	0.651	0.007	0.151	0.003	0.027	0.024	0.000
Laplace	(i)	0.50	0.641	0.505	0.141	0.005	0.025	0.098	0.073	0.000
	(ii)	0.50	0.520	0.501	0.020	0.001	0.001	0.002	0.000	0.323
	(iii)	0.25	0.331	0.335	0.081	0.085	0.010	0.028	0.018	0.000
	(iv)	0.50	0.535	0.583	0.035	0.083	0.002	0.012	0.010	0.000
	(v)	0.25	0.272	0.345	0.022	0.095	0.001	0.013	0.011	0.000
	(vi)	0.50	0.443	0.556	−0.057	0.056	0.006	0.005	−0.001	0.019
Contaminated	(i)	0.50	0.600	0.496	0.100	−0.004	0.013	0.175	0.162	0.000
	(ii)	0.50	0.491	0.494	−0.009	−0.006	0.001	0.020	0.020	0.000
	(iii)	0.25	0.268	0.473	0.018	0.223	0.002	0.054	0.053	0.000
	(iv)	0.50	0.513	0.654	0.013	0.154	0.001	0.027	0.025	0.000
	(v)	0.25	0.278	0.399	0.028	0.149	0.002	0.023	0.022	0.000
	(vi)	0.50	0.487	0.580	−0.013	0.080	0.003	0.016	0.014	0.000

A.6 Estimation of the difference in mean $\mu_2 - \mu_1$

Table 11. The accuracy of the HM- and ML-estimators with regard to estimating the difference in mean parameter $\mu_2 - \mu_1$. 500 samples each with 50 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of $\mu_2 - \mu_1$

Estimation of the difference in mean $\mu_2 - \mu_1$, $n = 50$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		$\mu_2 - \mu_1$	HM	ML	HM	ML	HM	ML		
Normal	(i)	2	2.105	2.216	0.105	0.216	0.226	0.345	0.119	0.001
	(ii)	3	2.902	2.984	−0.098	−0.016	0.113	0.185	0.072	0.002
	(iii)	3	2.486	2.870	−0.514	−0.130	0.711	0.539	−0.172	0.006
	(iv)	4	3.953	4.166	−0.047	0.166	0.429	0.538	0.109	0.065
	(v)	4	3.645	4.196	−0.355	0.196	0.816	0.908	0.092	0.215
	(vi)	3	3.889	3.887	0.889	0.887	2.191	2.788	0.596	0.011
Logistic	(i)	2	2.099	2.048	0.099	0.048	0.248	0.481	0.234	0.000
	(ii)	3	2.917	2.878	−0.083	−0.122	0.108	0.294	0.185	0.000
	(iii)	3	2.566	2.561	−0.434	−0.439	0.627	1.004	0.376	0.000
	(iv)	4	3.877	3.925	−0.123	−0.075	0.305	0.607	0.302	0.000
	(v)	4	3.572	3.686	−0.428	−0.314	1.137	1.375	0.238	0.047
	(vi)	3	3.521	3.323	0.521	0.323	1.679	2.378	0.699	0.039
Laplace	(i)	2	2.075	1.815	0.075	−0.185	0.309	0.687	0.378	0.000
	(ii)	3	3.011	2.855	0.011	−0.145	0.080	0.347	0.267	0.000
	(iii)	3	2.673	2.342	−0.327	−0.658	0.624	1.332	0.708	0.000
	(iv)	4	3.973	3.725	−0.027	−0.275	0.424	0.931	0.507	0.008
	(v)	4	3.557	3.162	−0.443	−0.838	1.260	2.042	0.782	0.000
	(vi)	3	3.245	3.257	0.245	0.257	1.300	3.624	2.325	0.000
Contaminated	(i)	2	2.336	2.029	0.336	0.029	1.634	1.365	−0.269	0.364
	(ii)	3	2.950	2.799	−0.050	−0.201	0.064	0.551	0.486	0.000
	(iii)	3	2.753	2.508	−0.247	−0.492	0.566	1.220	0.653	0.000
	(iv)	4	3.950	3.715	−0.050	−0.285	0.542	1.360	0.818	0.005
	(v)	4	3.744	3.076	−0.256	−0.924	1.158	1.800	0.643	0.000
	(vi)	3	3.407	3.145	0.407	0.145	2.256	4.548	2.292	0.012

Table 12. The accuracy of the HM- and ML-estimators with regard to estimating the difference in mean parameter $\mu_2 - \mu_1$. 500 samples each with 100 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of $\mu_2 - \mu_1$

Estimation of the difference in mean $\mu_2 - \mu_1$, $n = 100$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		$\mu_2 - \mu_1$	HM	ML	HM	ML	HM	ML		
Normal	(i)	2	2.075	2.116	0.075	0.116	0.146	0.267	0.121	0.001
	(ii)	3	2.880	2.962	−0.120	−0.038	0.072	0.095	0.023	0.076
	(iii)	3	2.446	2.854	−0.554	−0.146	0.501	0.414	−0.087	0.056
	(iv)	4	3.770	4.033	−0.230	0.033	0.262	0.280	0.018	0.684
	(v)	4	3.511	4.055	−0.489	0.055	0.547	0.543	−0.005	0.914
	(vi)	3	3.805	3.486	0.805	0.486	1.361	1.347	−0.015	0.911
Logistic	(i)	2	2.054	2.031	0.054	0.031	0.160	0.486	0.326	0.000
	(ii)	3	2.941	2.935	−0.059	−0.065	0.051	0.164	0.113	0.000
	(iii)	3	2.609	2.605	−0.391	−0.395	0.383	0.770	0.387	0.000
	(iv)	4	3.837	3.761	−0.163	−0.239	0.210	0.464	0.255	0.000
	(v)	4	3.510	3.493	−0.490	−0.507	0.680	1.216	0.536	0.000
	(vi)	3	3.503	2.993	0.503	−0.007	1.021	1.118	0.098	0.744
Laplace	(i)	2	2.104	1.768	0.104	−0.232	0.179	0.779	0.599	0.000
	(ii)	3	2.997	2.803	−0.003	−0.197	0.033	0.373	0.339	0.000
	(iii)	3	2.776	2.085	−0.224	−0.915	0.340	1.656	1.316	0.000
	(iv)	4	3.901	3.512	−0.099	−0.488	0.159	0.841	0.681	0.000
	(v)	4	3.563	2.899	−0.437	−1.101	0.828	2.116	1.288	0.000
	(vi)	3	2.993	2.840	−0.007	−0.160	0.571	3.014	2.443	0.000
Contaminated	(i)	2	2.141	2.094	0.141	0.094	0.825	2.103	1.277	0.000
	(ii)	3	2.993	2.807	−0.007	−0.193	0.029	0.458	0.428	0.000
	(iii)	3	2.822	2.194	−0.178	−0.806	0.217	1.208	0.991	0.000
	(iv)	4	3.948	3.448	−0.052	−0.552	0.136	1.706	1.569	0.000
	(v)	4	3.741	2.818	−0.259	−1.182	0.562	2.015	1.453	0.000
	(vi)	3	3.146	3.039	0.146	0.039	2.391	5.900	3.509	0.071

Table 13. The accuracy of the HM- and ML-estimators with regard to estimating the difference in mean parameter $\mu_2 - \mu_1$. 500 samples each with 500 observations, are generated from four mixture distributions (normal, logistic, Laplace and contaminated Gaussian) with the parameter configurations (i)–(vi). The parameter p is estimated for each sample. For each scenario we observe: the true parameter value (p), the average estimate (mean), the average deviation from the true value (bias), the Mean Squared Error (MSE), the difference between the ML-MSE and HM-MSE values ($\Delta_{\hat{MSE}}$) and a p-value for the test that this difference is significant. A positive (negative) value of $\Delta_{\hat{MSE}}$ indicates that the HM-estimator is more (less) efficient than the ML-estimator of $\mu_2 - \mu_1$

Estimation of the difference in mean $\mu_2 - \mu_1$, $n = 500$

	Data	True	Mean		$\hat{bias}$		$\hat{MSE}$		$\Delta_{\hat{MSE}}$	p-val
		$\mu_2 - \mu_1$	HM	ML	HM	ML	HM	ML		
Normal	(i)	2	1.962	2.044	−0.038	0.044	0.016	0.078	0.063	0.008
	(ii)	3	2.876	3.005	−0.124	0.005	0.027	0.012	−0.015	0.000
	(iii)	3	2.231	2.986	−0.769	−0.014	0.633	0.040	−0.593	0.000
	(iv)	4	3.657	3.984	−0.343	−0.016	0.161	0.068	−0.093	0.000
	(v)	4	3.351	3.983	−0.649	−0.017	0.450	0.181	−0.269	0.000
	(vi)	3	3.677	3.055	0.677	0.055	0.615	0.166	−0.449	0.000
Logistic	(i)	2	1.974	1.858	−0.026	−0.142	0.014	0.586	0.571	0.000
	(ii)	3	2.915	3.033	−0.085	0.033	0.015	0.021	0.006	0.247
	(iii)	3	2.516	2.722	−0.484	−0.278	0.302	0.494	0.192	0.000
	(iv)	4	3.699	3.609	−0.301	−0.391	0.133	0.320	0.187	0.000
	(v)	4	3.296	3.154	−0.704	−0.846	0.526	1.051	0.524	0.000
	(vi)	3	3.222	2.639	0.222	−0.361	0.164	0.203	0.040	0.040
Laplace	(i)	2	2.050	1.340	0.050	−0.660	0.011	1.178	1.167	0.000
	(ii)	3	2.979	2.957	−0.021	−0.043	0.007	0.140	0.134	0.000
	(iii)	3	2.850	1.709	−0.150	−1.291	0.054	2.159	2.105	0.000
	(iv)	4	3.817	3.185	−0.183	−0.815	0.067	0.793	0.726	0.000
	(v)	4	3.281	2.431	−0.719	−1.569	0.652	2.577	1.924	0.000
	(vi)	3	2.836	2.312	−0.164	−0.688	0.076	0.529	0.453	0.000
Contaminated	(i)	2	1.956	1.123	−0.044	−0.877	0.012	2.306	2.294	0.000
	(ii)	3	2.980	2.748	−0.020	−0.252	0.005	0.269	0.264	0.000
	(iii)	3	2.799	1.771	−0.201	−1.229	0.087	1.674	1.587	0.000
	(iv)	4	3.848	3.171	−0.152	−0.829	0.047	0.888	0.840	0.000
	(v)	4	3.463	2.617	−0.537	−1.383	0.357	1.977	1.620	0.000
	(vi)	3	2.955	2.440	−0.045	−0.560	0.078	0.466	0.389	0.000

A.7 Clustering results for non-mixture distributions

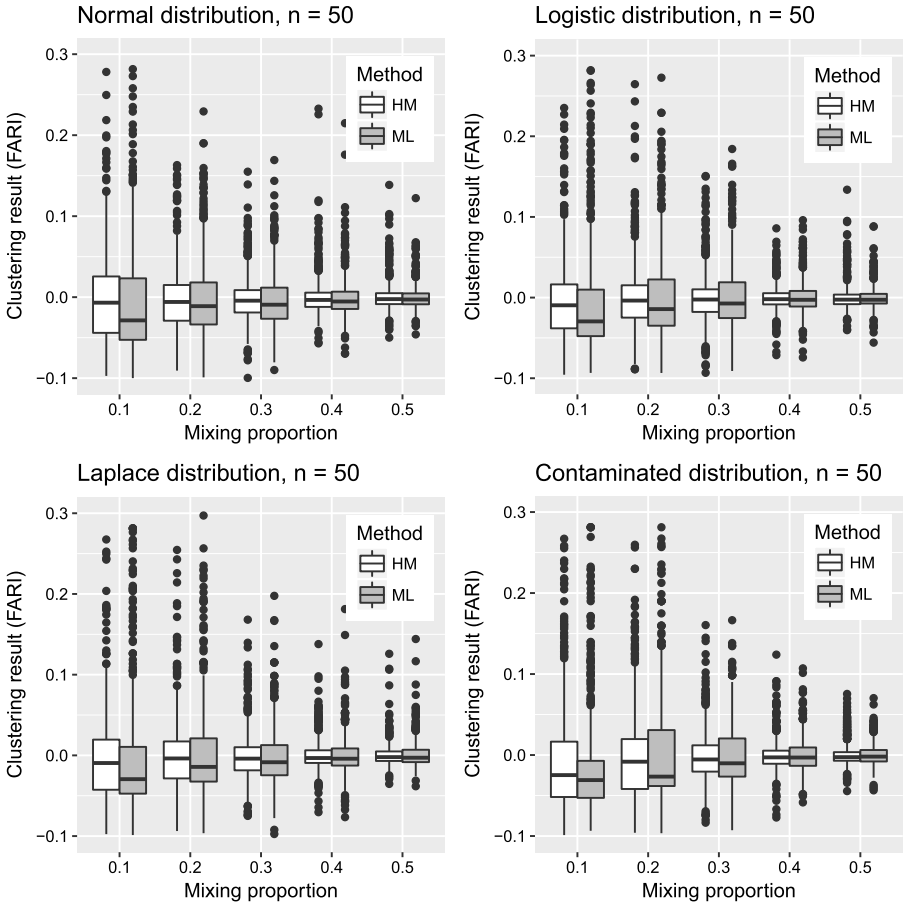


Fig. 8. Clustering results in terms of fuzzy adjusted Rand index (FARI) for data containing no information about the class labels, i.e. the mixture components coincide. For each distribution (normal, logistic, Laplace, and contaminated normal), 5 values of the mixing proportion were considered. 500 samples each with 50 observations were generated

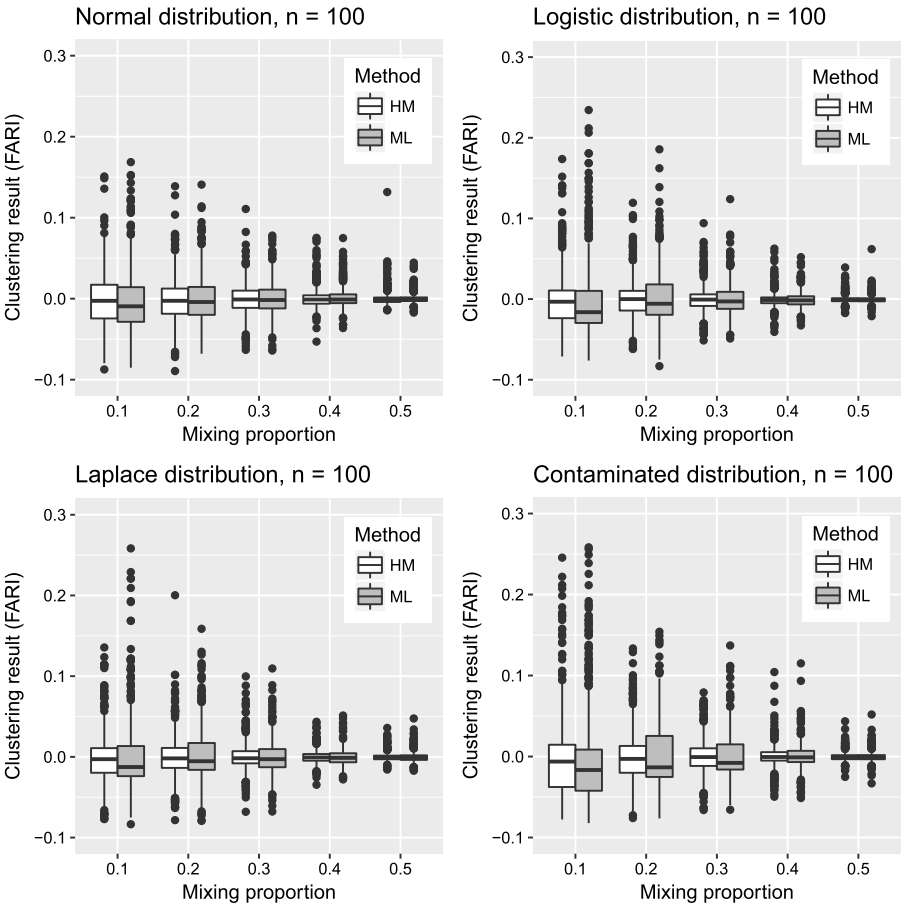


Fig. 9. Clustering results in terms of fuzzy adjusted Rand index (FARI) for data containing no information about the class labels, i.e. the mixture components coincide. For each distribution (normal, logistic, Laplace, and contaminated normal), 5 values of the mixing proportion were considered. 500 samples each with 100 observations were generated

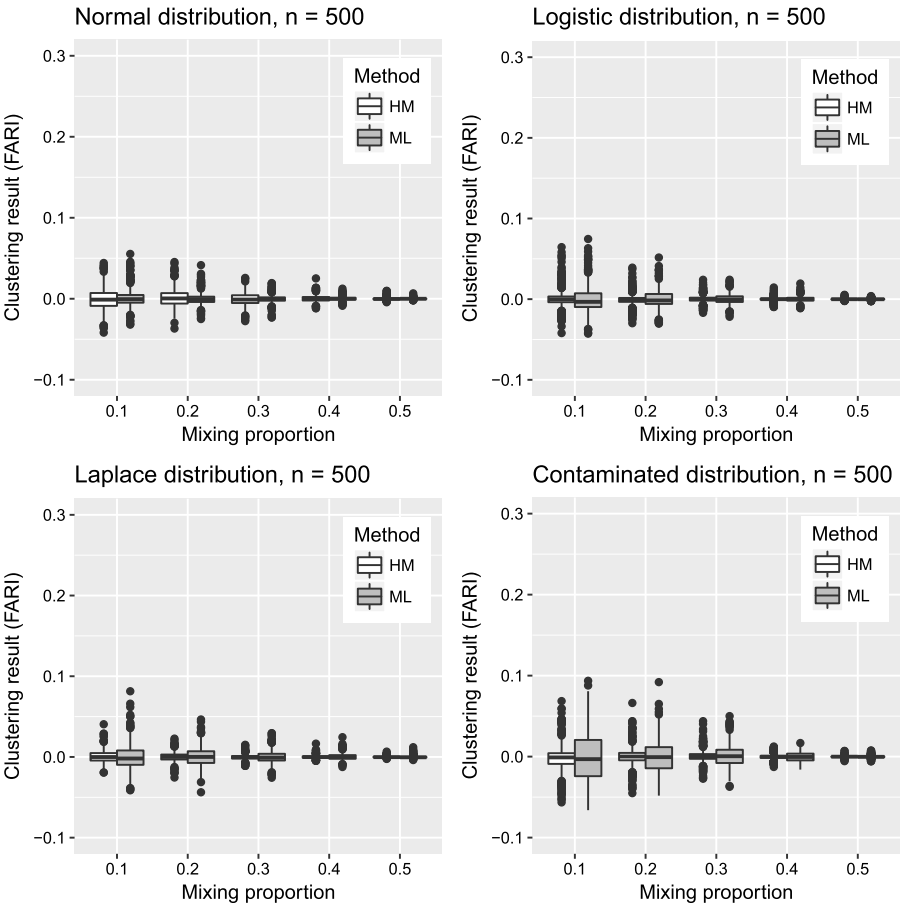


Fig. 10. Clustering results in terms of fuzzy adjusted Rand index (FARI) for data containing no information about the class labels, i.e. the mixture components coincide. For each distribution (normal, logistic, Laplace, and contaminated normal), 5 values of the mixing proportion were considered. 500 samples each with 500 observations were generated

References

[1] Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.: mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software **32**(6), 1–29 (2009)

[2] Bordes, L., Mottelet, S., Vandekerkhove, P., et al.: Semiparametric estimation of a two-component mixture model. The Annals of Statistics **34**(3), 1204–1232 (2006). MR2278356

[3] Brouwer, R.K.: Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems **32**(3), 213–235 (2009)

[4] Celeux, G., Chauveau, D., Diebolt, J.: Stochastic versions of the em algorithm: an experimental study in the mixture case. Journal of Statistical Computation and Simulation **55**(4), 287–314 (1996)

[5] Clarke, B., Heathcote, C.: Robust estimation of k-component univariate normal mixtures. Annals of the Institute of Statistical Mathematics **46**(1), 83–93 (1994). MR1272750

[6] Cutler, A., Cordero-Braña, O.I.: Minimum hellinger distance estimation for finite mixture models. Journal of the American Statistical Association **91**(436), 1716–1723 (1996)

[7] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38 (1977)

[8] Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association **97**(457), 77–87 (2002)

[9] Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Statistica Sinica **20**(1), 101 (2010). MR2640659

[10] Freyhult, E., Landfors, M., Önskog, J., Hvidsten, T.R., Rydén, P.: Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. BMC Bioinformatics **11**(1), 503 (2010)

[11] Fujisawa, H., Eguchi, S.: Robust estimation in the normal mixture model. Journal of Statistical Planning and Inference **136**(11), 3989–4011 (2006)

[12] Gleason, J.R.: Understanding elongation: The scale contaminated normal family. Journal of the American Statistical Association **88**(421), 327–337 (1993)

[13] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)

[14] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer **27**(2), 83–85 (2005)

[15] Hathaway, R.J.: A constrained formulation of maximum-likelihood estimation for normal mixture distributions. The Annals of Statistics, 795–800 (1985)

[16] Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review **22**(2), 85–126 (2004)

[17] Hunter, D.R., Wang, S., Hettmansperger, T.P.: Inference for mixtures of symmetric distributions. The Annals of Statistics, 224–251 (2007)

[18] Ju, J., Kolaczyk, E.D., Gopal, S.: Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. Remote Sensing of Environment **84**(4), 550–560 (2003)

[19] McLachlan, G., Peel, D.: Finite Mixture Models. John Wiley & Sons (2004)

[20] McLachlan, G.J., Basford, K.E.: Mixture models: Inference and applications to clustering. Applied Statistics (1988)

[21] Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal **7**(4), 308–313 (1965). MR3363409

[22] Pearson, K.: Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 71–110 (1894)

[23] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2017). https://www.R-project.org/

[24] Schlattmann, P., Böhning, D.: Mixture models and disease mapping. Statistics in Medicine **12**(19–20), 1943–1950 (1993)

[25] Sfikas, G., Nikou, C., Galatsanos, N.: Robust Image Segmentation with Mixtures of Student's t-Distributions. In: IEEE International Conference on Image Processing, 2007. ICIP 2007, vol. 1, p. 273. IEEE (2007)

[26] Titterington, D., Smith, A., Makov, U.: Statistical Analysis of Finite Mixture Models. Wiley, Chichester, UK (1985)

[27] Wolf, D.M., Lenburg, M.E., Yau, C., Boudreau, A., van 't Veer, L.J.: Gene co-expression modules as clinically relevant hallmarks of breast cancer diversity. PloS ONE **9**(2), 88309 (2014)

[28] Woodward, W.A., Parr, W.C., Schucany, W.R., Lindsey, H.: A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. Journal of the American Statistical Association **79**(387), 590–598 (1984)