ADAC (2007) 1:201–219
DOI 10.1007/s11634-007-0014-z

REGULAR ARTICLE

A global algorithm to estimate the expectations of the components of an observed univariate mixture

Nicolas Paul · Michel Terre · Luc Fety

Received: 29 March 2007 / Revised: 26 July 2007 / Accepted: 18 September 2007
Published online: 26 October 2007

Abstract This paper deals with the unsupervised classification of univariate observations. Given a set of observations originating from a K-component mixture, we focus on the estimation of the component expectations. We propose an algorithm based on the minimization of the "K-product" (KP) criterion we introduced in a previous work. We show that the global minimum of this criterion can be reached by first solving a linear system then calculating the roots of some polynomial of order K. The KP global minimum provides a first raw estimate of the component expectations, then a nearest-neighbour classification enables to refine this estimation. Our method's relevance is finally illustrated through simulations of various mixtures. When the mixture components do not strongly overlap, the KP algorithm provides better estimates than the Expectation-Maximization algorithm.

Keywords Univariate mixture · Component expectations estimation · Unsupervised classification

Mathematics Subject Classification (2000) 68T05

1 Introduction

In this paper we deal with unsupervised classification. Given a set of univariate observations originating from K possible components, we focus on the estimation of the component expectations. The number of components is supposed to be known and the component expectations are all different. One method consists in estimating

N. Paul (✉) · M. Terre · L. Fety
Conservatoire National des Arts et Metiers, Electronic and Communications,
292 rue Saint-Martin, 75003 Paris, France
e-mail: nicolas.paul@cnam.fr

the probability density function (pdf), a mixture of K pdf, by associating a kernel to each observation and adding the contributions of all the kernels (Parzen 1962). A search of the pdf modes then leads to the component expectations. The drawback of such a method is that it requires the selection of extra-parameters (kernel design, intervals for the mode search). An alternative method consists in using the Expectation-Maximization (EM) algorithm (Dempster et al. 1977). The EM algorithm is the most commonly used method when the mixture components belong to the same known parameterized family. It is an iterative algorithm that looks for the mixture parameters that maximize the likelihood of the observations. Each EM iteration consists of two steps. The Expectation step estimates the probability for each observation to come from each mixture component. Then, during the Maximization step, these estimated probabilities are used to update the estimation of the mixture parameters. This procedure converges to a maximum (local or global) of the likelihood (Dempster et al. 1977).

If the mixture components do not belong to a common and known parameterized family, the EM algorithm does not directly apply. Yet, if the component densities do not overlap too much, some clustering methods can be used to cluster the data and calculate the cluster means. In Fisher (1958) an algorithm is proposed to compute the K-partition of the N ordered observations which minimizes the sum of the squares within clusters. Instead of testing the $\binom{N-1}{K-1}$ possible partitions, some relationship between k-partitions and $(k+1)$-partitions is used to recursively compute the optimal K-partition. The main drawbacks of this method are a high sensitivity to potential differences between the cluster variances and a complexity in $\mathrm{O}(KN^2)$ (Fitzgibbon et al. 2000).

The K-means algorithm (Hartigan and Wong 1977) is one of the most popular clustering method. It is an iterative algorithm which groups the data into K clusters in order to minimize an objective function such as the sum of squared Euclidean point to cluster centroid distances. The main drawback of the popular K-means or EM algorithms is the potential convergence to some local non-global extrema of the criterion they use. Some solutions consist, for instance, in using smart initializations (see McLachlan and Peel (2000); Lindsay and Furman (1994) for EM, Bradley and Fayyad (1998) for K-means) or stochastic optimization in order to become less sensitive in the initialization (see Celeux et al. (1995), Pernkopf and Bouchaffra (2005) for EM, Krishna and Murty (1999) for K-means). Another drawback of these methods is the convergence speed which can be very slow. A survey of the clustering techniques can be found in Berkhin (2006) and in Xu and Wunsch (2005).

In this contribution, we propose a non-iterative algorithm which mainly consists in calculating the minimum of the "K-Product" (KP) criterion we first introduced in Paul et al. (2006): if $\{z_n\}_{n\in\{1,\ldots,N\}}$ is a set of N observations in $\mathbb{R}^1$ which originate from a K-component mixture and if $\{x_k\}_{k\in\{1,\ldots,K\}}$ is any vector of $\mathbb{R}^K$, we define the KP criterion as the sum of all the K-terms products $\prod_{k=1}^{K}(z_n - x_k)^2$ [see (2) below]. The main motivation for using such a criterion is that, though it provides a slightly biased estimation of the component expectations, its global minimum can be reached by first solving a system of K linear equations then calculating the roots of some polynomial of order K. Once these K roots have been obtained, a final clustering step assigns each observation to the closest root and calculates the means of the resulting clusters. Another advantage of the proposed method is that it does not require the specification of any extra-parameters.

The rest of the paper is organized as follow: In Sect. 2 the observation model is presented and the criterion is defined. In Sect. 3 the global minimum of the criterion is theoretically calculated. In Sect. 4 a K-product based algorithm for estimating the component expectations is described and the EM algorithm is recalled. Section 5 presents simulation results which illustrate the performance of the algorithm for different mixtures: mixtures of three, six and nine components have been simulated with various parameter configurations (common/different mixing proportions, common/different variances, Gaussian/non-Gaussian component densities). Finally, conclusions and perspectives are given in Sect. 6.

2 Observation model and criterion definition

Let $\{g_k : \mathbb{R}^1 \to \mathbb{R}^+ : z \to g_k(z)\}_{k\in\{1,\dots,K\}}$ be a set of K probability density functions with different expectations $a_k := \int_{z=-\infty}^{+\infty} z g_k(z)\mathrm{d}z \in \mathbb{R}^1$ and let $\{\pi_k\}_{k\in\{1,\dots,K\}}$ be a set of K nonnegative weights (prior probabilities) that sum up to one. The multimodal probability density function (pdf) of the random observable variable Z is a finite mixture given by:

$$f_Z(z) = \sum_{k=1}^{K} \pi_k g_k(z). \tag{1}$$

Note that the form of the densities g_k is usually not known by the statistician and that the g_k do not necessarily belong to the same parameterized family.

Now let $\{z_n\}_{n\in\{1,\dots,N\}}$ be a set of N realizations of Z with pdf (1). In the following we always assume that N is greater than K and that the number of different realizations is greater than $K-1$. Our purpose is to estimate the K component expectations $a_1,\dots,a_K$ from the set of observations $\{z_n\}_{n\in\{1,\dots,N\}}$. The estimation of the component expectations is based on the minimization of the new K-product (KP) criterion $J(\mathbf{x})$ defined by:

$$J : \mathbb{R}^K \to \mathbb{R}^+ : \mathbf{x} \to \sum_{n=1}^{N}\prod_{k=1}^{K}(z_n - x_k)^2. \tag{2}$$

Note the difference with one form of the K-means algorithm that amounts to minimizing the criterion (3) defined by:

$$G : \mathbb{R}^K \to \mathbb{R}^+ : \mathbf{x} \to \sum_{n=1}^{N} \min_{k\in\{1,\dots,K\}} (z_n - x_k)^2. \tag{3}$$

The KP criterion (2) is nonnegative for any vector $\mathbf{x}$. The first intuitive motivation for defining this criterion is its behavior in the limit case when all the variances of the pdfs g_k are null. In this limit case, all the observations are equal to one of the a_k.

Therefore, if $\mathbf{a} := (a_1, a_2, \ldots, a_K)^t$ is the vector of the component expectations, we have $J(\mathbf{a}) = 0$. $J(\mathbf{x})$ will be minimum if and only if $\mathbf{x}$ is equal to $\mathbf{a}$ or any of its $K!$ permutations. In the general case, when the component variances are strictly positive, the KP minimum remains a useful approximation of the component expectations. A refined estimation of the component expectations, based on the K-product minimum, is detailed in Sect. 4. Our second motivation for defining the KP criterion (2) is that it will have $K!$ minima that are the $K!$ permutations of one single vector which can be reached by solving a linear system of equation, then finding the roots of some polynomial of order K. This is shown in Sect. 3.

3 Global minimum of the KP criterion

We first present in Sect. 3.1 some useful definitions which are needed in Sect. 3.2 to calculate the global minimum of J.

3.1 Some useful definitions

To any observation z_n in $\mathbb{R}^1$ we associate the vector $\mathbf{z}_n$ defined by:

$$\mathbf{z}_n := (z_n^{K-1}, z_n^{K-2}, \ldots, 1)^t \in \mathbb{R}^K. \tag{4}$$

The vector $\mathbf{z}$ and the Hankel matrix $\mathbf{Z}$ are then respectively defined by:

$$\mathbf{z} := \sum_{n=1}^{N} z_n^K \mathbf{z}_n \in \mathbb{R}^K \tag{5}$$

$$\mathbf{Z} := \sum_{n=1}^{N} \mathbf{z}_n \mathbf{z}_n^t = \left(\sum_{n=1}^{N} z_n^{2K-j-l} \right)_{j,l=1,\ldots,K} \in \mathbb{R}^{K \times K}. \tag{6}$$

The matrix $\mathbf{Z}$ is regular if the number of different observations in $\{z_1, \ldots, z_K\}$ is greater than $K - 1$ (an explanation is given in Appendix A).

Now let $\mathbf{y} = (y_1, \ldots, y_K)^t$ be a vector of $\mathbb{R}^K$. We define the polynomial $q_{\mathbf{y}}(\alpha)$ of order K as:

$$q_{\mathbf{y}}(\alpha) := \alpha^K - \sum_{k=1}^{K} \alpha^{K-k} y_k \qquad \alpha \in \mathbb{R}^1. \tag{7}$$

If $\mathbf{r} = (r_1, \ldots, r_K)^t$ is the vector of $\mathbb{C}^K$ containing the K roots $r_1, \ldots, r_K$ of $q_{\mathbf{y}}(\alpha)$ the factorial form of $q_{\mathbf{y}}(\alpha)$ is:

$$\begin{aligned}
q_{\mathbf{y}}(\alpha) &= \prod_{k=1}^{K}(\alpha - r_k) \\
&= \alpha^K - (r_1 + \cdots + r_K)\alpha^{K-1} + \cdots + (-1)^K(r_1 \times r_2 \times \cdots \times r_K) \\
&= \alpha^K - \sum_{k=1}^{K} \alpha^{K-k} w_k(\mathbf{r}),
\end{aligned}$$

where $w_k(\mathbf{r})$ is the Elementary Symmetric Polynomial (ESP) in the variables $r_1, \ldots, r_K$ defined by:

$$w_k(\mathbf{r}) := (-1)^{k+1} \sum_{\substack{\{j_1,\ldots,j_k\} \in \{1\cdots K\}^k \\ 1 \leqslant j_1 < \cdots < j_k \leqslant K}} r_{j_1} \cdot r_{j_2} \cdot \cdots \cdot r_{j_k}. \tag{8}$$

For instance, for $K = 3$, we have:

$$\begin{aligned}
w_1(\mathbf{r}) &= r_1 + r_2 + r_3 \\
w_2(\mathbf{r}) &= -(r_1 r_2 + r_2 r_3 + r_1 r_3) \\
w_3(\mathbf{r}) &= r_1 r_2 r_3.
\end{aligned}$$

If we introduce the ESP vector of $\mathbf{r}$, $\mathbf{w}(\mathbf{r})$, defined by:

$$\mathbf{w}(\mathbf{r}) := (w_1(\mathbf{r}), \ldots, w_K(\mathbf{r}))^t, \tag{9}$$

the relationship between the roots and coefficients of $q_{\mathbf{y}}(\alpha)$ becomes:

$$\mathbf{y} = \mathbf{w}(\mathbf{r}) \Leftrightarrow \; q_{\mathbf{y}}(r_k) = 0 \quad \forall k \in \{1, \ldots, K\}. \tag{10}$$

3.2 The minimum of the KP criterion

The global minimum of J is given by Theorem 1:

Theorem 1 *If $\boldsymbol{y}_{\min}$ is the solution of $\mathbf{Z}\boldsymbol{y}_{\min} = \boldsymbol{z}$ [where $\boldsymbol{z}$ and $\mathbf{Z}$ have been defined in* (5) *and* (6)] *and if $\boldsymbol{x}_{\min}$ is a vector containing, in any order, the K roots of $q_{\boldsymbol{y}_{\min}}(\alpha)$* [*defined in* (7)], *then $\boldsymbol{x}_{\min}$ belongs to $\mathbb{R}^K$ and $\boldsymbol{x}_{\min}$ yields the global minimum of $\boldsymbol{J}$.*

The proof is given in Appendix B.

4 Estimation of the component expectations

In Sect. 4.1 we describe a K-product based algorithm to estimate the expectations of the K components of an observed univariate mixture. The classical EM algorithm is then described in Sect. 4.2.

Table 1 KP algorithm and complexities of single steps

Step 1: Calculate a minimum of J	
Calculate $\mathbf{Z}$ and $\mathbf{z}$	$\mathrm{O}(NK)$
Calculate $\mathbf{y}_{\min}$ by solving $\mathbf{Z}\mathbf{y}_{\min} = \mathbf{z}$	$\mathrm{O}(K^2)$
Calculate the roots $(x_{1,\min}, \ldots, x_{K,\min})$ of $q_{\mathbf{y}_{\min}}(\alpha)$	$\mathrm{O}(K^2)$
Step 2: Clustering and cluster mean estimation	
Assign each z_n to the closest $x_{k,\min}$	$\mathrm{O}(NK)$
Calculate the K centroids of the resulting clusters	$\mathrm{O}(N)$

4.1 K-product algorithm

The proposed algorithm for estimating the component expectations consists of two steps. In the first step, the minimum of the function $J(\mathbf{x})$ [see (2)], $\mathbf{x}_{\min} = (x_{1,\min}, \ldots, x_{K,\min})^t$, is calculated, giving a first raw estimate of the set of component expectations. This first estimate is slighly biased: let us consider, for instance, a Gaussian mixture with $K = 2$ balanced components ($\pi_1 = \pi_2 = 0.5$) with expectations $-a$ and a and with a common standard deviation σ. When the number of observations, N, tends to infinity, the (asymptotic) solution of $\mathbf{Z}\mathbf{y}_{\min} = \mathbf{z}$ is $\mathbf{y}_{\min} = (0, a^2 + \sigma^2)^t$ and the roots of $q_{\mathbf{y}_{\min}}(\alpha)$ are:

$$\mathbf{x}_{\min} = (x_{1,\min}, x_{2,\min}) = \left(-a\sqrt{1+\frac{\sigma^2}{a^2}}, a\sqrt{1+\frac{\sigma^2}{a^2}}\right) \neq (-a, a).$$

Therefore, in a second step, each observation z_n is assigned to the nearest $x_{k,\min}$, K clusters are formed, and the cluster means are calculated. These cluster means provide the final estimation of the component expectations. The steps of the algorithm and their complexities are summarized in Table 1. Some implementation of the KP algorithm can be designed with complexity $\mathrm{O}(NK + K^2)$, which is equivalent to $\mathrm{O}(NK)$ since N is greater than K.

A free version of a Matlab 7.0.4 implementation of the KP algorithm is available on request to the authors. In the first step, the linear system solving is based on the Cholesky factorization of the symmetrical matrix $\mathbf{Z}$. The complexity is $\mathrm{O}(K^3)$, but could be reduced to $\mathrm{O}(K^2)$ by using the Hankel property of $\mathbf{Z}$ [see Bojanczyk et al. (1995) for instance]. The roots of $q_{\mathbf{y}_{\min}}(\alpha)$ are then calculated with Matlab function "roots", which builds the companion matrix of $q_{\mathbf{y}_{\min}}(\alpha)$ then finds its eigenvalues with a QR factorization method. The complexity is $\mathrm{O}(K^3)$, but could be reduced to $\mathrm{O}(K^2)$ by using some faster roots finding algorithms [see Uhlig (1999) for instance].

4.2 EM algorithm

Assuming that the mixture components have a Gaussian pdf, the standard EM algorithm proceeds as follows (Dempster et al. 1977): if $\hat{\beta}_{n,k}^{(\mathrm{ite})}$ is the estimated probability that z_n comes from component k and if $\hat{\pi}_k^{(\mathrm{ite})}$, $\hat{a}_k^{(\mathrm{ite})}$ and $\hat{\sigma}_k^{2,(\mathrm{ite})}$ are respectively, the

estimated prior, expectation and variance of component k at iteration ite, then the estimates at iteration ite + 1 are provided by the two following steps:
Expectation step:

$$\hat{\beta}_{n,k}^{(\text{ite}+1)} = \frac{\frac{\hat{\pi}_k^{(\text{ite})}}{\sqrt{2\pi}\hat{\sigma}_k^{(\text{ite})}} \exp\left(-\frac{1}{2}\left(\frac{z_n - \hat{a}_k^{(\text{ite})}}{\hat{\sigma}_k^{(\text{ite})}}\right)^2\right)}{\sum_{k=1}^{K} \frac{\hat{\pi}_k^{(\text{ite})}}{\sqrt{2\pi}\hat{\sigma}_k^{(\text{ite})}} \exp\left(-\frac{1}{2}\left(\frac{z_n - \hat{a}_k^{(\text{ite})}}{\hat{\sigma}_k^{(\text{ite})}}\right)^2\right)}.$$

Maximization step:

$$\hat{\pi}_k^{(\text{ite}+1)} = \frac{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)}}{N}$$

$$\hat{a}_k^{(\text{ite}+1)} = \frac{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)} z_n}{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)}}$$

$$\hat{\sigma}_k^{2,(\text{ite}+1)} = \frac{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)} \left(z_n - \hat{a}_k^{(\text{ite}+1)}\right)^2}{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)}}.$$

This iterative procedure converges to a (local or global) maximum of the likelihood function $\prod_{n=1}^{N} \left\{\sum_{k=1}^{k} \frac{\pi_k}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2}\left(\frac{z_n - a_k}{\sigma_k}\right)^2\right)\right\}$. The main drawback of the EM algorithm is the potential convergence to some non-global maximum of the likelihood. In particular, this version of EM may converge to very unbalanced parameters, with, on the one hand, some small classes with a small variance and a small prior probability, and, on the other hand, some large classes with a large variance and a large prior probability. One possibility to decrease the risk of bad convergence consists in assuming a common variance σ^2 and a common prior probability $\pi_k = \frac{1}{K}$ to all the mixture components. In this case only $K + 1$ parameters have to be estimated, the K component expectations $a_1, \ldots, a_K$ and the common variance σ^2. In this "constrained" version of the EM algorithm the update of the estimated mixture parameters is provided by the two following steps:
Expectation step:

$$\hat{\beta}_{n,k}^{(\text{ite}+1)} = \frac{\exp\left(-\frac{1}{2}\left(\frac{z_n - \hat{a}_k^{(\text{ite})}}{\hat{\sigma}^{(\text{ite})}}\right)^2\right)}{\sum_{k=1}^{K} \exp\left(-\frac{1}{2}\left(\frac{z_n - \hat{a}_k^{(\text{ite})}}{\hat{\sigma}^{(\text{ite})}}\right)^2\right)}.$$

Maximization step:

$$\hat{a}_k^{(\text{ite}+1)} = \frac{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)} z_n}{\sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)}}$$

$$\hat{\sigma}^{2,(\text{ite}+1)} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} \hat{\beta}_{n,k}^{(\text{ite}+1)} \left(z_n - \hat{a}_k^{(\text{ite}+1)}\right)^2.$$

5 Simulations

5.1 Simulation scenarios and performance criterion

In our simulations several types of Gaussian mixtures have first been considered. The number K of components is three (scenario A), six (scenario B), and nine (scenario C). In scenario A, the set of component expectations is equal to $\{0, 1, 2\}$. In scenario B, the set of component expectations is equal to $\{0, 1, 2, 4, 5, 6\}$ (i.e. with differences 1 and 2). In scenario C the set of component expectations is equal to $\{0, 1, 2, 4, 5, 6, 8, 9, 10\}$ (i.e. with differences 1 and 2). For each scenario "X", four cases have been studied: common variance and common mixing weight (scenario X1), different variances and common mixing weight (scenario X2), common variance and different mixing weights (scenario X3) and different variances and different mixing weights (scenario X4). A summary of all the scenarios is given in Tables 2, 3 and 4. The number of observations (N) per simulation run is 100 in scenario A, 200 in scenario B and 300 in scenario C. A non-Gaussian case has also been investigated in scenario B_{bis} described in Table 5: The expectations, variances and prior probabilities are the same than in scenario B, but three components have a uniform density (with expectations 0,2,5) and three components have a Laplace density (with expectations 1,4,6).

Table 2 Simulations, scenario A: expectations, probabilities, and variances of the Gaussian mixture components

Expectation	Scenario A1		Scenario A2		Scenario A3		Scenario A4	
	Variance	Prior	Variance	Prior	Variance	Prior	Variance	Prior
0	σ^2	$\frac{1}{3}$	σ^2	$\frac{1}{3}$	σ^2	0.4	σ^2	0.4
1	σ^2	$\frac{1}{3}$	$\frac{\sigma^2}{2}$	$\frac{1}{3}$	σ^2	0.4	$\frac{\sigma^2}{2}$	0.4
2	σ^2	$\frac{1}{3}$	σ^2	$\frac{1}{3}$	σ^2	0.2	σ^2	0.2

Table 3 Simulations, scenario B: expectations, probabilities, and variances of the Gaussian mixture components

Expectation	Scenario B1		Scenario B2		Scenario B3		Scenario B4	
	Variance	Prior	Variance	Prior	Variance	Prior	Variance	Prior
0	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
1	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
2	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.1	σ^2	0.1
4	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
5	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
6	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.1	$\frac{\sigma^2}{2}$	0.1

Table 4 Simulations, scenario C: expectations, probabilities, and variances of the Gaussian mixture components

Expectation	Scenario C1		Scenario C2		Scenario C3		Scenario C4	
	Variance	Prior	Variance	Prior	Variance	Prior	Variance	Prior
0	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{2}{15}$	σ^2	$\frac{2}{15}$
1	σ^2	$\frac{1}{9}$	$\frac{\sigma^2}{2}$	$\frac{1}{9}$	σ^2	$\frac{2}{15}$	$\frac{\sigma^2}{2}$	$\frac{2}{15}$
2	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{15}$	σ^2	$\frac{1}{15}$
4	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{15}$	σ^2	$\frac{1}{15}$
5	σ^2	$\frac{1}{9}$	$\frac{\sigma^2}{2}$	$\frac{1}{9}$	σ^2	$\frac{3}{15}$	$\frac{\sigma^2}{2}$	$\frac{3}{15}$
6	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{15}$	σ^2	$\frac{1}{15}$
8	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{2}{15}$	σ^2	$\frac{2}{15}$
9	σ^2	$\frac{1}{9}$	$\frac{\sigma^2}{2}$	$\frac{1}{9}$	σ^2	$\frac{2}{15}$	$\frac{\sigma^2}{2}$	$\frac{2}{15}$
10	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{9}$	σ^2	$\frac{1}{15}$	σ^2	$\frac{1}{15}$

Table 5 Simulations, scenario $\mathrm{B_{bis}}$: expectations, probabilities, variances and density forms of the non-Gaussian mixture components

Density	Expectation	$\mathrm{B1_{bis}}$		$\mathrm{B2_{bis}}$		$\mathrm{B3_{bis}}$		$\mathrm{B4_{bis}}$	
		Variance	Prior	Variance	Prior	Variance	Prior	Variance	Prior
Uniform	0	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
Laplace	1	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
Uniform	2	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.1	σ^2	0.1
Laplace	4	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.2	$\frac{\sigma^2}{2}$	0.2
Uniform	5	σ^2	$\frac{1}{6}$	σ^2	$\frac{1}{6}$	σ^2	0.2	σ^2	0.2
Laplace	6	σ^2	$\frac{1}{6}$	$\frac{\sigma^2}{2}$	$\frac{1}{6}$	σ^2	0.1	$\frac{\sigma^2}{2}$	0.1

In each simulation run, we compare the performance of our algorithm to the standard EM and the constrained EM algorithms described in Sect. 4.2. Yet we only present here the results of the constrained EM since, in the investigated scenarios, the performances of the contrained EM are always superior to the performances of the standard EM, even when the true mixture components have different variances and/or different prior probabilities. To initialize the EM algorithms in our simulations, the K initial component expectations $\hat{a}_k^{(0)}$ were randomly chosen from a uniform distribution in the range $[\min(z_n),\ \max(z_n)]$. For each n, $\hat{\beta}_{n,k}^{(1)}$ is set to one if $\hat{a}_k^{(0)}$ is the closest component expectation to the observation z_n, and $\hat{\beta}_{n,k}^{(1)}$ is set to zero otherwise. This initialization is repeated until each cluster contains at least one observation. Then the EM starts with a maximization step. The algorithm is stopped if all estimated parameters do not change between two EM steps or if a maximal number of 100 iterations is reached.

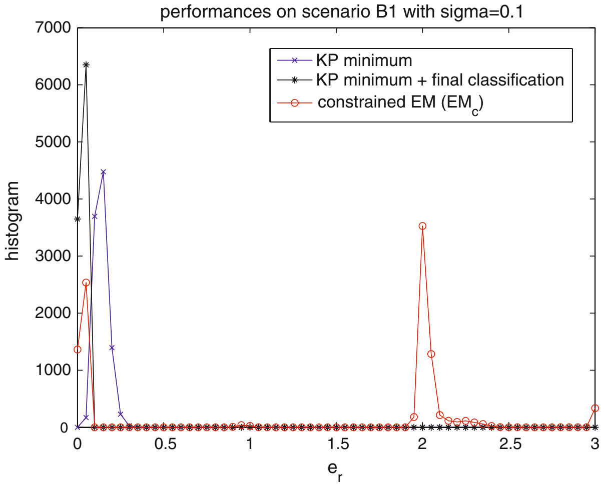


Fig. 1 Estimation performances on scenario B1 with $\sigma = 0.1$. Ten thousand simulation runs have been performed. For each simulation run, 200 observations have been generated. e_r is the maximal distance between the ordered vector of true component expectations and the ordered vector of estimated component expectations

To get rid of the permutation ambiguity, the estimation performance is evaluated as follows: If $\mathbf{a}$ is the vector of the true component expectations and $\hat{\mathbf{a}}_r$ is the vector of the estimated component expectations at simulation run r, the performance criterion e_r is defined as the maximal absolute distance between the true and estimated ordered vector of component expectations:

$$e_r := \left\| \mathrm{sort}(\mathbf{a}) - \mathrm{sort}(\hat{\mathbf{a}}_r) \right\|_\infty ,$$

where $\mathrm{sort}(\mathbf{x})$ is the ordered permutation of $\mathbf{x}$ and $\|\cdot\|_\infty$ is the infinity norm in $\mathbb{R}^K$.

5.2 Simulation results

The distribution of e_r is displayed in Fig. 1 and summarized in Table 6 for the scenario B1 ($K = 6$) with $\sigma = 0.1$ and 10,000 simulation runs. The KP minimum yields a biased estimation: e_r is greater than 0.1 for 86% of the runs. Then the full KP

Table 6 Histogram of e_r on scenario B1 with $\sigma = 0.1$

Method	Value of e_r					
	[0, 0.1]	[0.1, 0.2]	[0.2, 0.3]	[0.3, 0.5]	[0.5, 1]	> 1
KP minimum	14%	79%	7%	0%	0%	0%
Full KP	100%	0%	0%	0%	0%	0%
Constrained EM	39%	0%	0%	0%	1%	60%

algorithm (calculation of the KP minimum followed by nearest-neighbour classification) provides a perfect set of estimates (e_r is always less than 0.1). On the contrary the constrained EM algorithm converges to a wrong set of modes for 61% of the runs. In this case, the constrained EM gets stuck at a non-global maximum of the likelihood. Typically one estimated component expectation is located in the middle of two true component expectations, while two other estimated component expectations are close to the same true component expectation.

In Fig. 2 we present the constrained EM and KP performances for scenario A ($K = 3$) with different values of σ. The two methods have equivalent performances, with a slight superiority of the constrained EM when the true mixture components have the same prior probabilities (scenarios A1 and A2) and a slight superiority of KP otherwise. Note that the KP performances are almost independent on the configuration of the prior probabilities (common/different) and on the configuration of the mixture component variances (common/different).

The results in scenarios B ($K = 6$) and C ($K = 9$) are displayed in Figs. 3 and 4. For small values of σ, the KP algorithm yields more accurate estimates then the constrained EM algorithm: when σ is greater than 0, there is a risk that the constrained EM converges to a wrong set of estimated component expectations. On the contrary, the KP algorithm provides perfect estimates for some non-null values of σ: In scenario B (resp. scenario C), e_r always remains less than 0.1 if σ is less than 0.1 (resp. 0.05). Yet, if the mixture components strongly overlap ($\sigma > 0.25$), the EM algorithm has a small but non-null probability to converge to the correct set of component expectations. In such a situation several restarts of the constrained EM with different initializations will finally provide a correct set of component expectations. On the contrary, when the component densities strongly overlap, the bias of the KP minimum is too large and the final nearest-neighbour classification step fails to separate the observations which originate from different components.

The results for the non-Gaussian scenario B_{bis} are finally displayed in Fig. 5: the relative performance of the two methods is roughly the same than in the Gaussian case. More simulations and theoretical studies are required to conclude on the non-Gaussian case, but we expect the KP algorithm to provide correct estimates for any form of the component densities as soon as the component densities do not strongly overlap. Indeed, the definition of the KP criterion does not make any assumption on the form of the component densities.

In all the investigated scenarios the KP algorithm appears to be an appropriate tool to estimate component expectations if the component densities do not strongly overlap. It does not need several restarts or stochastic optimization procedure and it does not involve any extra-parameters. This makes the KP algorithm an efficient method for any on-line and/or complexity constrained applications.

6 Conclusion

Given a set of observations originating from a K-component univariate mixture, we focused on the estimation of the component expectations when the number K

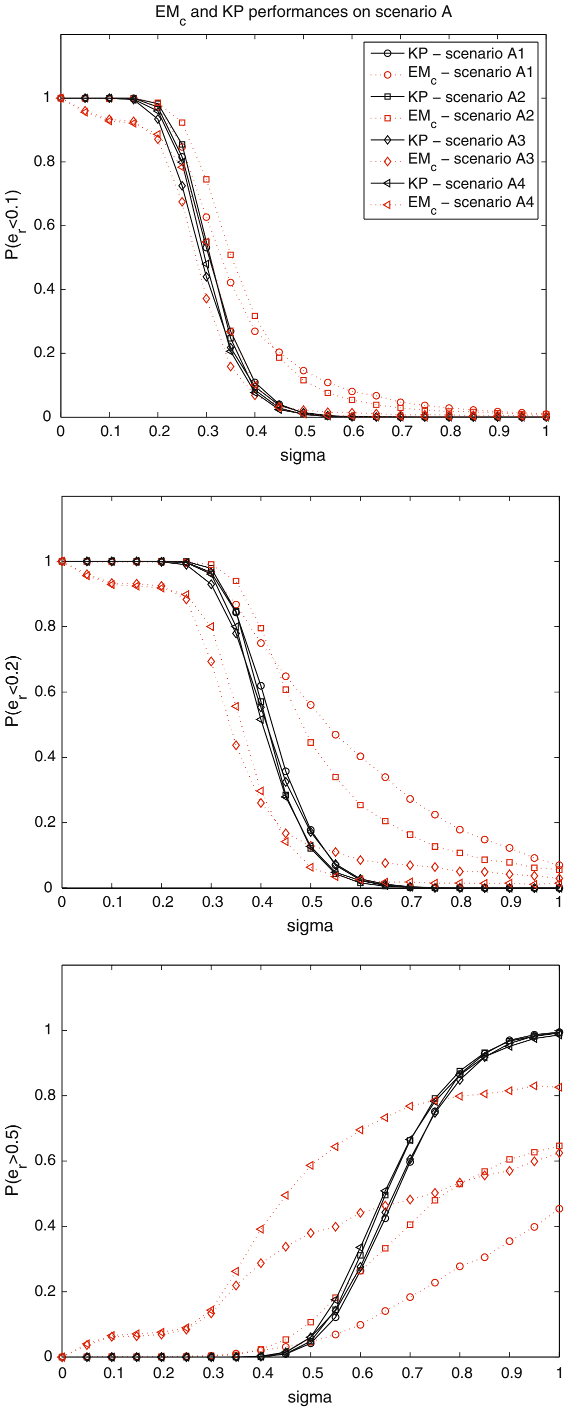
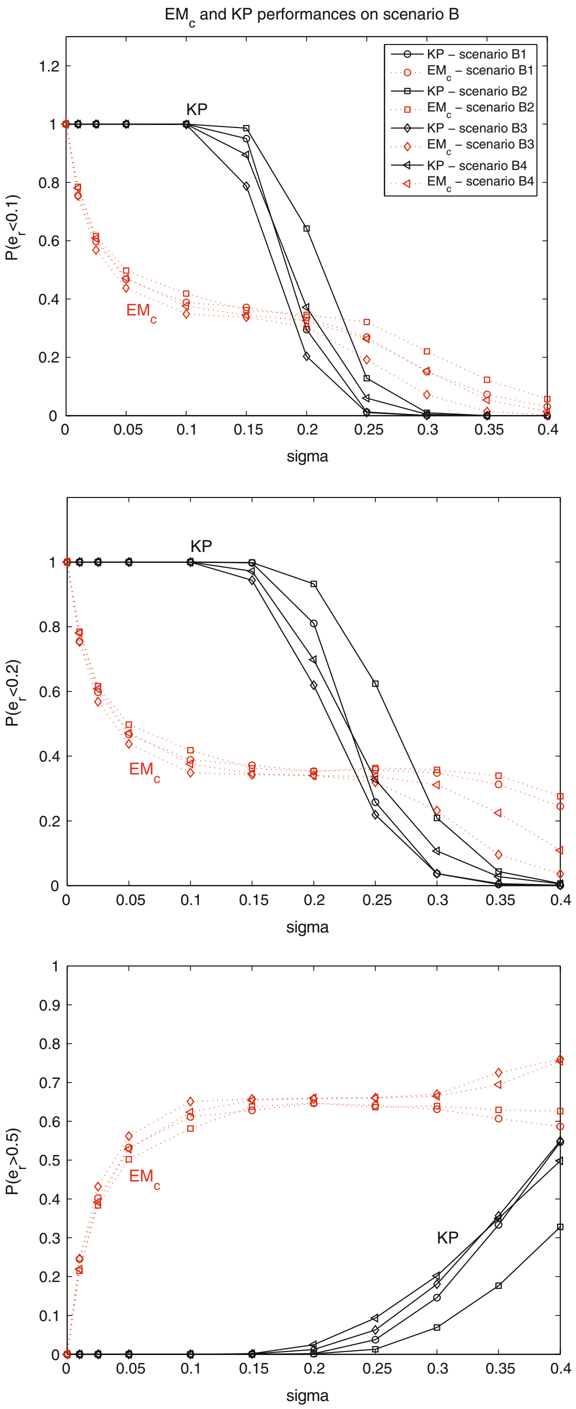


Fig. 2 Performances of the constrained EM (EM$_c$, *dotted lines*) and KP (*full lines*) algorithms on scenario A for different values of σ. For each value of σ and for each sub-scenario 10,000 simulation runs have been performed. For each simulation run, 100 observations have been generated. e_r is the maximal distance between the ordered vector of true component expectations and the ordered vector of estimated component expectations. The performance criteria are the observed frequencies for e_r to be smaller than 0.1 (*top*), smaller than 0.2 (*middle*) and greater than 0.5 (*bottom*)

Fig. 3 Performances of the constrained EM (EM$_c$, *dotted lines*) and KP (*full lines*) algorithms on scenario B for different values of σ. For each value of σ and for each sub-scenario 10,000 simulation runs have been performed. For each simulation run, 200 observations have been generated. e_r is the maximal distance between the ordered vector of true component expectations and the ordered vector of estimated component expectations. The performance criteria are the observed frequencies for e_r to be smaller than 0.1 (*top*), smaller than 0.2 (*middle*) and greater than 0.5 (*bottom*)



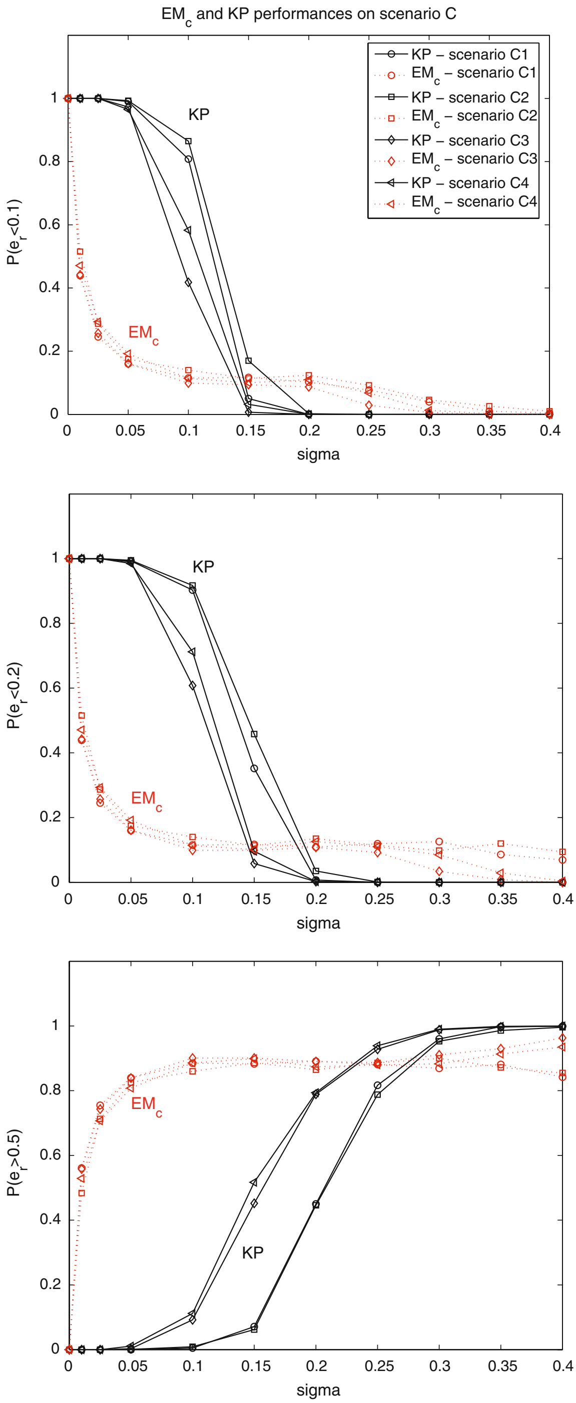
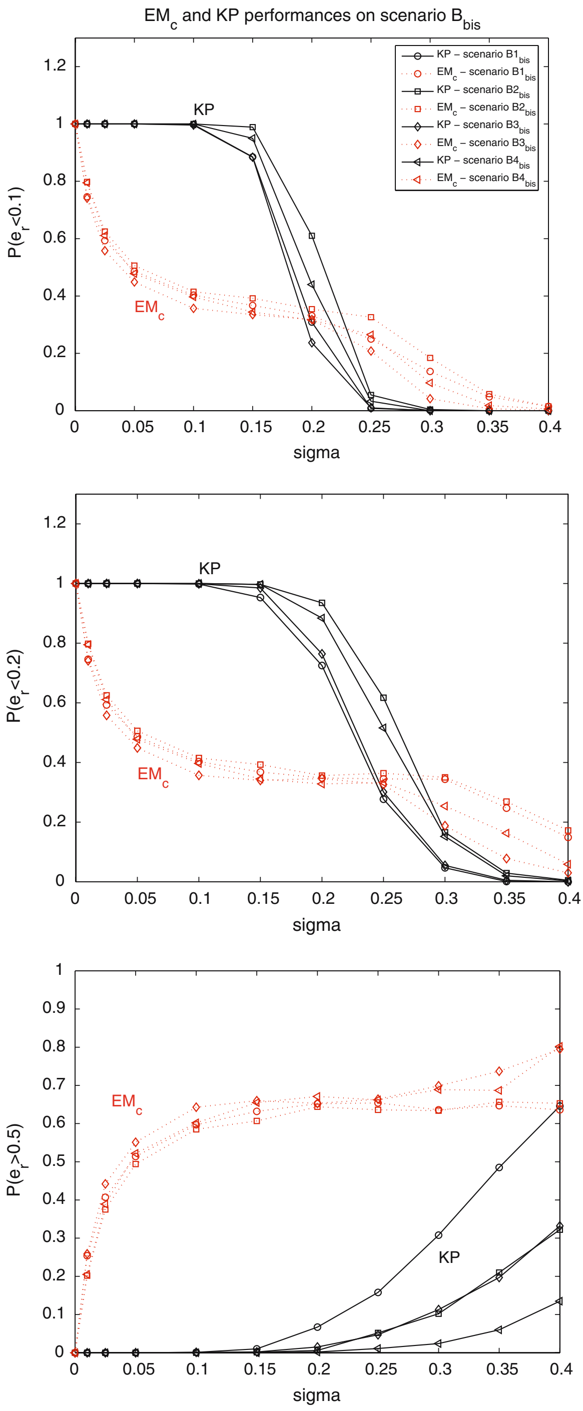


Fig. 4 Performances of the constrained EM (EM$_c$, *dotted lines*) and KP (*full lines*) algorithms on scenario C for different values of σ. For each value of σ and for each sub-scenario 1,000 simulation runs have been performed. For each simulation run, 300 observations have been generated. e_r is the maximal distance between the ordered vector of true component expectations and the ordered vector of estimated component expectations. The performance criteria are the observed frequencies for e_r to be smaller than 0.1 (*top*), smaller than 0.2 (*middle*) and greater than 0.5 (*bottom*)

Fig. 5 Performances of the constrained EM (EM$_c$, *dotted lines*) and KP (*full lines*) algorithms on scenario B$_{bis}$ for different values of σ. For each value of σ and for each sub-scenario 1,000 simulation runs have been performed. For each simulation run, 200 observations have been generated. e_r is the maximal distance between the ordered vector of true component expectations and the ordered vector of estimated component expectations. The performance criteria are the observed frequencies for e_r to be smaller than 0.1 (*top*), smaller than 0.2 (*middle*) and greater than 0.5 (*bottom*)



of components is known. We proposed a method based on the minimization of the "K-product" criterion first introduced in Paul et al. (2006). We have shown that the global minimum of this criterion can be reached with a linear least square minimization followed by a roots finding algorithm. This minimum is used to get a first raw estimate of the component expectations, and a final nearest-neighbour classification enables to refine the estimation. The proposed method is not iterative, its complexity is $\mathrm{O}(NK)$ and it does not require the specification of any extra parameter. Simulations have illustrated the performance and superiority of the KP algorithm in comparison with the EM algorithm when the mixture component densities do not strongly overlap.

We focused on the univariate case and our current research deals with the multivariate case. If the observations $\mathbf{z}_n$ belong to $\mathbb{R}^d$, if $\{\mathbf{x}_k\}_{k\in\{1,\ldots,K\}}$ is any set of K vectors of $\mathbb{R}^d$, the KP criterion is now defined as the sum of all the K-terms products $\prod_{k=1}^{K}\|\mathbf{z}_n-\mathbf{x}_k\|_{\mathbb{R}^d}^2$. The minima of such a criterion and some algorithms to find these minima are currently being studied. We also investigate the case of an unknown number of components. For this we are currently studying the location of the KP minimum when the assumed number of components (K_{test}) is different from the true number of components in the observed mixture.

Acknowledgments The authors would like to thank B. Scherrer, M. Bellanger, G. Saporta, JP. Nakache, the ADAC editor and the anomynous reviewers for their constructive comments that helped in improving this manuscript.

Appendix A: Non-singularity of Z

In Appendix A we explain why the matrix $\mathbf{Z}$ of size $K \times K$, defined in (6) is regular if the number of different observations is greater than $K-1$. $\mathbf{Z}$ can be written as the following matrix product:

$$\mathbf{Z} = \mathbf{V}\mathbf{V}^t,$$

where $\mathbf{V}$ is a $K \times N$ Vandermonde Matrix defined by:

$$\mathbf{V} \triangleq (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N),$$

and $\mathbf{z}_n$ has been defined in (4). Let us assume that the K first observations are different. The determinant of the $K \times K$ Vandermonde matrix $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ is equal to $\prod_{1\le i<j\le K}(z_j - z_i)$, which is different from zero. The rank of $\mathbf{V}$ is then equal to K, so the rank of $\mathbf{Z}$ is equal to K and $\mathbf{Z}$ is regular. □

Appendix B: Proof of Theorem 1

In Appendix B we prove Theorem 1. Let F be the function defined by:

$$F : \mathbb{C}^K \to \mathbb{R}^+ : \mathbf{x} \to \sum_{n=1}^{N}\prod_{k=1}^{K}\|z_n - x_k\|_{\mathbb{C}}^2.$$

The restriction of F to $\mathbb{R}^K$ is the function J from (2) since the observations z_n are real:

$$\forall \mathbf{x} \in \mathbb{R}^K : \quad F(\mathbf{x}) = J(\mathbf{x}). \tag{11}$$

Now let H be the function defined by:

$$H : \mathbb{C}^K \to \mathbb{R}^+ : \mathbf{y} \to \sum_{n=1}^{N} \left\| z_n^K - \mathbf{z}_n^t \mathbf{y} \right\|_{\mathbb{C}}^2.$$

We show that the function H applied to the ESP of a vector $\mathbf{x}$ in $\mathbb{C}^k$ is equal to the function F applied to $\mathbf{x}$: Consider, for $\mathbf{x} \in \mathbb{C}^K$:

$$F(\mathbf{x}) = \sum_{n=1}^{N} \left\| \prod_{k=1}^{K} (z_n - x_k) \right\|_{\mathbb{C}}^2. \tag{12}$$

Developping (12) by using definition (8) and including definitions (4) and (9) leads to:

$$F(\mathbf{x}) = \sum_{n=1}^{N} \left\| z_n^K - \sum_{k=1}^{K} z_n^{K-k} w_k(\mathbf{x}) \right\|_{\mathbb{C}}^2 = \sum_{n=1}^{N} \left\| z_n^K - \mathbf{z}_n^t \mathbf{w}(\mathbf{x}) \right\|_{\mathbb{C}}^2 = H(\mathbf{w}(\mathbf{x})). \tag{13}$$

The global minimum of $H(\mathbf{y})$ is the linear least square solution $\mathbf{y}_{\text{min}}$ given by:

$$\mathbf{y}_{\text{min}} = \underset{\mathbf{y} \in \mathbb{C}^K}{\operatorname{argmin}} \left\{ \sum_{n=1}^{N} \left\| z_n^K - \mathbf{z}_n^t \mathbf{y} \right\|_{\mathbb{C}}^2 \right\}. \tag{14}$$

Developping (14) by using definitions (5) and (6) and remembering that the coefficients of $\mathbf{Z}$ and $\mathbf{z}$ are real:

$$\begin{aligned} \mathbf{y}_{\text{min}} &= \underset{\mathbf{y} \in \mathbb{C}^K}{\operatorname{argmin}} \left\{ \mathbf{y}^H \mathbf{Z} \mathbf{y} - 2\text{Re}\{\mathbf{y}^H\} \mathbf{z} \right\} \\ \mathbf{Z} \mathbf{y}_{\text{min}} &= \mathbf{z}, \quad \mathbf{y}_{\text{min}} \in \mathbb{R}^K. \end{aligned} \tag{15}$$

The Hankel matrix $\mathbf{Z}$ is regular since the number of different observations is greater than $K-1$ (Appendix A). System (15) therefore has exactly one solution. Since $\mathbf{Z}$ belongs to $\mathbb{R}^{K\times K}$ and $\mathbf{z}$ belongs to $\mathbb{R}^K$, $\mathbf{y}_{\text{min}}$ belongs to $\mathbb{R}^K$. Now let $\mathbf{x}_{\text{min}}=(x_{1,\text{min}}, \ldots, x_{K,\text{min}})^t$ be a vector containing, in any order, the K (potentially complex) roots of $q_{\mathbf{y}_{\text{min}}}(\alpha)$. One can show that the following holds:

(i) $\mathbf{x}_{\text{min}}$ is a global minimum of F
(ii) $\mathbf{x}_{\text{min}} \in \mathbb{R}^K$
(iii) $\mathbf{x}_{\text{min}}$ is a global minimum of J

Property (i) is a direct consequence of (13) since, for each $\mathbf{x} \in \mathbb{C}^K$:

$$F(\mathbf{x}) = H(\mathbf{w}(\mathbf{x})) \geq \min\{H(\mathbf{y})\} = H(\mathbf{y}_{\min}).$$

According to (10), $\mathbf{y}_{\min} = \mathbf{w}(\mathbf{x}_{\min})$ and we have, for all $\mathbf{x} \in \mathbb{C}^K$:

$$\begin{aligned} &F(\mathbf{x}) \geq H(\mathbf{w}(\mathbf{x}_{\min})), \text{ thus} \\ &F(\mathbf{x}) \geq F(\mathbf{x}_{\min}), \end{aligned}$$

which proves (i). Property (ii) can be shown by contradiction: If $\mathbf{x}_{\min}$ does not belong to $\mathbb{R}^K$, then for one of the $x_{k,\min}$ we have $x_{k,\min} \neq \text{Re}\{x_{k,\min}\}$ and, since all the observations z_n are real:

$$\forall n \in \{1, \ldots, N\} : \quad \left\| z_n - x_{k,\min} \right\|_{\mathbb{C}} > \left\| z_n - \text{Re}\{x_{k,\min}\} \right\|_{\mathbb{C}},$$

which leads to:

$$F(\mathbf{x}_{\min}) > F(\text{Re}\{\mathbf{x}_{\min}\}).$$

This is impossible since $\mathbf{x}_{\min}$ is a global minimum of F. This proves property (ii). We finally have to prove (iii): since $\mathbf{x}_{\min} \in \mathbb{R}^K$ we have, using (11):

$$F(\mathbf{x}_{\min}) = J(\mathbf{x}_{\min}). \tag{16}$$

Furthermore, according to (11), for all $\mathbf{x} \in \mathbb{R}^K$:

$$J(\mathbf{x}) = F(\mathbf{x}) \geq \min_{\mathbf{x}}\{F(\mathbf{x})\} = F(\mathbf{x}_{\min}) = J(\mathbf{x}_{\min}),$$

according to property (i) and (16). This proves (iii), and properties (ii) and (iii) directly lead to Theorem 1. □

References

Berkin P (2006) A survey of clustering data mining techniques. In: Kogan J, Nicholas C, Teboulle M (eds) Grouping multidimensional data: recent advances in clustering. Springer, Berlin, pp 25–71

Bradley PS, Fayyad UM (1998) Refining initial points for K-means clustering. In: Proceedings of the 15th international conference on Machine Learning. Morgan Kaufmann, San-Fransisco, pp 91–99

Bojanczyk AW, Brent RP, Hoog FR (1995) Stability analysis of a general Toeplitz system solver. Numer Algorithms 10:225–244

Celeux G, Chauveau D, Diebolt J (1995) On stochastic versions of the EM algorithm. INRIA research report no 2514, available http://www.inria.fr/rrrt/rr-2514.html

Dempster A, Laird N, Rubin D (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Fisher WD (1958) On grouping for maximum homogeneity. J Am Stat Assoc 53(284):789–798

Fitzgibbon LJ, Allison L, Dowe DL (2000) Minimum message length grouping of ordered data. In: Arimura H, Jain S (eds) Proceedings of the 11th international conference on algorithmic learning theory, Sydney, Australia. LNAI, Springer, Berlin, pp 56–70

Hartigan J, Wong M (1979) A K-means clustering algorithm. J Appl Stat 28:100–108

Krishna K, Narasimha Murty M (1999) Genetic K-means algorithm. IEEE Trans Syst Man Cybern B Cybern 29(3):433–439

Lindsay B, Furman D (1994) Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. Comput Stat Data Anal 17(5):493–507

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Parzen E (1962) On estimation of a probability density function and mode. Ann Math Stat 33:1065–1076

Paul N, Terre M, Fety L (2006) The K-product criterion for Gaussian mixture estimation. In: Proceedings of the 7th Nordic signal processing symposium, Reykjavik, Iceland, pp 334–337, doi:10.1109/NORSIG.2006.275248

Pernkopf F, Bouchaffra D (2005) Genetic-based EM algorithm for learning Gaussian mixture models. IEEE Trans Pattern Anal Mach Intell 27(8):1344–1348

Uhlig F (1999) General polynomial roots and their multiplicities in O(n) memory and O(n^2) time. Linear Multilinear Algebra 46(4):327–359

Xu R, Wunsch DII (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16(3):645–678