DISTRIBUTION MIXTURES, A REDUCED-BIAS ESTIMATION ALGORITHM

Nicolas Paul[1], Alexandre Girard[1], Michel Terré[2]

[1] EDF R&D département STEP
6, quai Watier, 78401 Chatou, France
[2] CNAM, équipe CEDRIC/Laetitia
292, rue Saint-Martin, 75003 Paris, France

ABSTRACT

We focus on the definition of a new optimization criteria for mixtures of distributions estimation based on an evolution of the *K-Product* criterion [5]. For the case of monovariate observations we show that the new proposed criterion does not have any local non-global minimizer. This property is also observed for multivariate observations. The relevance of the new *K-Product* criterion is theoretically studied and analyzed through simulations (in some monovariate cases). We show that for a mixture of three separate uniform components, the distance between the criterion unique minimizer and the true component expectations is less than half the components standard deviation.

Index Terms— *K-means*, *K-products*, Distribution mixtures

1. INTRODUCTION

The purpose of this paper is the unsupervised estimation of parameters for a mixture of distributions. Let $\{\mathbf{x}_n\}_{n\in[1,N]}$ be a set of N multivariate (D being the dimension) observations originating from a K-components mixture. The probability density function of the observations is given by

$$h(\mathbf{x}) = \sum_{k=1}^{K} p_k h_k(\mathbf{x}) \quad (1)$$

where h_k (resp. p_k) is the k^{th}-component's distribution (resp. weight). The number of components K is supposed to be known. Our purpose is to estimate the K expectations $\{\boldsymbol{a}_k\}$ given by

$$\mathbf{a}_k = \int_{-\infty}^{\infty} \mathbf{x} h_k(\mathbf{x}) d\mathbf{x} \quad (2)$$

In the general case, when the expectations coincide with the K principal modes of $h(\mathbf{x})$, non-parametric approaches can be used in order to estimate the K modes of observations' distribution. Those methods are not based on any hypothesis concerning the mixture components. The idea is to compute a partition of the observations' space $\mathbb{R}^D$ in hyper-cubes in order to estimate a sampling of the probability density function (pdf) $h(\mathbf{x})$.

A kernel function (such as a Gaussian) is associated to each observation. Then we estimate $h(\mathbf{x})$ by adding the contributions of all the kernels to each hyper-cube. The K principal modes of the estimated pdf are then calculated [1]. This method presents several drawbacks, such as the delicate adjustment of the hyper-cube size and the adjustment of the kernel functions' width.

If we don't have enough observations, the estimated pdf may contain modes that don't fit with the original distribution modes. The estimated pdf modes' convergence to the real pdf modes is studied in [2]. The most common parametric approach is to assume that the different components of the mixture can be modeled by parameterized functions, such as Gaussians, and to search all the parameters that maximize the likelihood of the observations.

The *EM* algorithm (Expectation-Maximization) [3], usually used to estimate missing data, is the most commonly used method to estimate parameters.

In the case of an equiprobable Gaussian mixture, where components' supports are separated and where components have the same variance, the log-likelihood function is very similar to the *K-means* criterion [4], frequently used in non-supervised classification and vectorial quantification

$$J_{\text{K-Means}}(\mathbf{u}_1, \dots, \mathbf{u}_K) = \sum_{n=1}^{N} \min_k \|\mathbf{x}_n - \mathbf{u}_k\|^2 \quad (3)$$

The main drawback of the *K-means* or *EM* algorithms is the potential convergence to some stationary points which are not global extrema, for example local non-global extrema.

A common example is presented in figure 1 where observations are from a mixture of three bi-dimensional (K=3, D=2) Gaussian distributions in a configuration relatively "simple": components are separated, equiprobable and have the same covariance. Even in these conditions, the *EM* algorithm or *K-means* don't necessarily converge to the correct solution. The results obtained with *EM* and *K-means* are relatively similar in this example, that's why only the results obtained with *K-means* are provided for the sake of clarity.

2. K-PRODUCT CRITERION

2.1. Introduction of the new criterion

Our approach is slightly different from the one defined in section 1. It consists in calculating the minimum of the *K-product* (*KP*) criterion defined by

$$J_\varepsilon(\mathbf{u}_1, \dots, \mathbf{u}_K) = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\varepsilon + \prod_{k=1}^{K}\|\mathbf{x}_n - \mathbf{u}_k\|^2} \quad (4)$$

where $\varepsilon > 0$ is a regularization parameter, allowing the criterion to have a continuous gradient in $\mathbb{R}^D$. When ε=0 the criterion is given by

$$J_0(\mathbf{u}_1, \dots, \mathbf{u}_K) = \frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K}\|\mathbf{x}_n - \mathbf{u}_k\| \quad (5)$$

This criterion is an evolution of the *K-product* criterion proposed in [5]

$$J_{\text{norm }2}(\mathbf{u}_1, \dots, \mathbf{u}_K) = \frac{1}{N}\sum_{n=1}^{N}\prod_{k=1}^{K}\|\mathbf{x}_n - \mathbf{u}_k\|^2 \quad (6)$$

Some of this criterion's applications are presented in [6].

The difference lies in the choice of the norm used in the product presented in equations (5) and (6). The norm 2 is proposed in the original criterion (6), so called "*K-product norm 2*"in the sequel of the paper, the norm 1 (resp. norm 1 "regularized") is introduced in the new criterion (5) (resp. (4)) proposed is this paper.

The main drawback of the *K-product norm 2* criterion, is the bias between the minimizer and the components' expectations. When the number of observations tends to infinity, the minimizer presented in equation (6) does not tend to components' expectations. For example, in the case of a mixture with K=2 monovariate (D=1) and equiprobable components, with expectations $-a$ and $+a$ and having equal standard deviation σ, the minimizer of (6) tends to $\left\{a\sqrt{1+\sigma^2/a^2}, -a\sqrt{1+\sigma^2/a^2}\right\}$.

Figure 1 presents another example of the bias, with the *K*=3 mixture, mentioned in section 1.

2.2. Minimization of the new criterion

A relaxation algorithm with complexity O(*NKD*) can be used to minimize the criterion presented in equation (4). In the neighborhood of an estimation $\left\{\mathbf{u}_1^{ite-1}, \dots, \mathbf{u}_k^{ite-1}, \dots, \mathbf{u}_K^{ite-1}\right\}$ during the "*ite*" iteration, the gradient of $J_\varepsilon(\mathbf{u}_1, \dots, \mathbf{u}_K)$ with respect to the k^{th} component $\mathbf{u}_k$ is equivalent to

$$\frac{\partial J_\varepsilon}{\partial \mathbf{u}_k} \approx \frac{1}{N}\sum_{n=1}^{N} D_{n,k}(\mathbf{x}_n - \mathbf{u}_k) \quad (7)$$

where

$$D_{n,k} = \frac{C_{n,k}}{\left(\varepsilon + C_{n,k}\left\|\mathbf{x}_n - \mathbf{u}_k^{ite\text{-}1}\right\|^2\right)^{0.5}} \quad (8)$$

and

$$C_{\text{n,k}} = \prod_{l=1}^{l=k-1}\left\|\mathbf{x}_n - \mathbf{u}_l^{ite}\right\|^2 \prod_{l=k+1}^{l=K}\left\|\mathbf{x}_n - \mathbf{u}_l^{ite-1}\right\|^2 \quad (9)$$

$C_{\text{n,k}}$ can be computed using a recurrence on *k*, the update of k^{th} component of the $\mathbf{u}_k^{ite}$ vector is then obtained by solving

$$\mathbf{u}_k^{ite} = \sum_{n=1}^{N}\frac{D_{n,k}}{\sum_{m=1}^{N} D_{m,k}}\mathbf{x}_n \quad (10)$$

We finally obtain an algorithm with complexity O(*ND*).

2.3. Minimizers

Fig. 2 shows the criterion evolution obtained with different initializations based on the same 100 observations of the mixture described in Tab 1 (*D=4*, *K=3*).

Whatever the initialization is, the algorithm has converged to the same minimizer. This can be demonstrated in the monovariate case (*D=1*), using the function $\boldsymbol{w}(\mathbf{u})$ which associates to each vector $\mathbf{u}$ of $\mathbb{R}^K$, a vector containing the *K* Elementary Symmetric Polynomials (*ESP*) of $\mathbf{u} = (u_1, \dots, u_K)$

$$w_k(\mathbf{u}) = (-1)^{k+1}\sum_{j_1<\dots<j_k\le K} u_{j_1} \times \dots \times u_{j_k} \quad (11)$$

An important aspect of $\boldsymbol{w}(\mathbf{u})$ is that if two vectors $\mathbf{u}_A$ and $\mathbf{u}_B$ are defined such as $\boldsymbol{w}(\mathbf{u}_A) = \boldsymbol{w}(\mathbf{u}_B)$ then $\mathbf{u}_A$ and $\mathbf{u}_B$ are equal, modulo permutation.

Lets $H_\varepsilon(\mathbf{v})$ be the function strictly convex defined by

$$H_\varepsilon(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\varepsilon + (x_n^K - (x_n^{K-1}, \dots, x_n^0)^{\mathbf{t}}\mathbf{v})^2} \quad (12)$$

For all $\mathbf{u}$ in $\mathbb{R}^K$, we have $J_\varepsilon(\mathbf{u}) = H_\varepsilon(\boldsymbol{w}(\mathbf{u}))$. We demonstrate that if $\mathbf{u}_{\min}$ is a minimizer of $J_\varepsilon(\mathbf{u})$ then $\boldsymbol{w}(\mathbf{u}_{\min})$ is the unique minimizer of $H_\varepsilon(\mathbf{v})$. All the minimizer of $J_\varepsilon(\mathbf{u})$ are then equal, modulo a permutation.

3. THEORETICAL PERFORMANCE

The relevance of the *K-product* minimizer for estimating mixtures expectations can be evaluated by simulation (multivariate case), and theoretically (univariate case).

In particular we show in this section that for a mixture of three separated uniform distributions the distance between the three elements of the single minimizer of *K-product* and the three expectations is bounded by the standard deviation of the components.

In the univariate case, the optimization of asymptotic criterion $J_0^\infty(u_1, \dots, u_K) = \int_{-\infty}^{\infty}\prod_{k=1}^{K}|x - u_k|h(x)dx$ leads to valid integral relations for any mixture of distributions. For $K = 3$ components, cancellation of combinations like $\frac{\partial J_0^\infty}{\partial u_k} - \frac{\partial J_0^\infty}{\partial u_l}$ and $u_k\frac{\partial J_0^\infty}{\partial u_k} - u_l\frac{\partial J_0^\infty}{\partial u_l}$ eliminates the terms in u_k in

the integrals and thus we obtain the following optimality conditions for k = 0, 1, 2

$$\int_{u_1}^{u_2} x^k \mathrm{h}(x)\mathrm{d}x + \int_{u_3}^{\infty} x^k \mathrm{h}(x)\mathrm{d}x = \frac{\mathrm{E}_h\{X^k\}}{2} \qquad (13)$$

The dependence of relation (13) between u_i appears only on bounds of integrals. These relationships also help to establish the results directly on the conditions of uniqueness of local minima.

These relations are simplified in the case of uniform distributions mixtures with separated supports $[a_k - b \;\; a_k + b]$ with common width $2b$, with standard deviation $\sigma = b/\sqrt{3}$ and with expectation a_k.

Going back to (13) the optimal solution $\{u_1, u_2, u_3\}$ solves

$$\begin{gathered} -u_1 + u_2 - u_3 = -a_1 + a_2 - a_3 \\ -u_1^2 + u_2^2 - u_3^2 = -a_1^2 + a_2^2 - a_3^2 - b^2 \\ -u_1^3 + u_2^3 - u_3^3 = -a_1^3 + a_2^3 - a_3^3 + 3b^2\left(-a_1 + a_2 - a_3\right) \end{gathered} \qquad (14)$$

After development, this system leads to the following relations for the estimation errors $v_k = u_k - a_k$

$$v_2 = \frac{b^2}{2 + \frac{b^2}{(a_2 - a_1)(a_3 - a_2)}} \left(\frac{1}{a_2 - a_1} - \frac{1}{a_3 - a_2}\right)$$

$$v_2 = v_1 + v_3$$

$$v_1 v_3 + v_1(a_2 - a_1) - v_3(a_3 - a_2) + \frac{b^2}{2} = 0 \qquad (15)$$

We can bound the estimation errors with the system (15). For separated components, the first equation leads to $0 < \frac{1}{(a_k - a_l)} < 1/2b$ if $k > l$ then $|v_2| < b/4$ by the first relation. Moving to the second equation, we obtain that if v_1 and v_3 have the same sign then their absolute values are lower than $b/4$. If they have opposite signs, the third equation of (15) leads to $v_1 < 0$ and $v_3 > 0$ and $|v_1|(a_2 - a_1) + |v_3|(a_3 - a_2) < b^2/2$ then $|v_{1,3}| < b^2/2 \times 1/2b = b/4$. Then, for k = 1, 2 et 3

$$|u_k - a_k| < \frac{b}{4} < \frac{\sigma}{2} \qquad (16)$$

The difference between the minimizer of the asymptotic *K-product norm 1* criterion and expectations components is less than half the standard deviation of the components. This property, as demonstrated here for the simplified case of a univariate mixture of uniform components, is also observed in the case of multivariate Gaussian mixtures.

4. SIMULATION RESULTS

Simulations are also used to verify the accuracy of the criterion in more complex cases, particularly in the case already mentioned and presented in figure 1 or in the case of higher dimensions (Table 1). These results clearly show the interest of our approach "*K-product norm 1*" compared to existing methods.

In these cases relatively "simple" (K = 3 identical, equiprobable, separated components) conventional algorithms such as *EM* or *K-means* don't necessarily converge to the correct solution, but may converge to a non-global optimizer, with a vector which "covers" the two closest components and the remaining two vectors which are distributed in the remaining class. These algorithms need then to be restarted several times with different initializations, in order to be sure to get a proper solution.

The behavior of the relaxation algorithm presented in section 2.2 is illustrated in figures 3 to 6. It is shown that the algorithm converges in less of 50 iterations and the bias of *K-product norm 2* with respect to the reduced bias of *K-product norm 1* is highlighted.

4. CONCLUSION

An evolution of the standard *K-product* originally proposed in [5] is introduced to estimate the expectations of a mixture of distributions when the number of components is known. The relevance criterion is studied theoretically. We show in the univariate case, that the criterion doesn't admit any non-global local minimizer, and that for a mixture of three separated uniform distributions with common supports width, the gap between the single minimizer and expectations that we sought to be estimated is less than half of the standard deviation of the components.

The validity of these results in the multivariate case, as suggested by the simulations, is being studied, as well as the extension to the case where the number of components is not known.

2. REFERENCES

[1] Duda R., Hart P., Stock D., Pattern Classification. John Wiley and Sons, New-York (2001)

[2] Parzen E., "On Estimation of a Probability Density Function and Mode," *Annals of Mathematical Statistics*, vol. 33, 1962, pp. 1065-1076.

[3] Dempster A., Laird N., Rubin D., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, vol. B 39, 1977, pp.:1-38.

[4] Hartigan J., Wong M., "A K-Means Clustering Algorithm," *Journal of Applied Statistics*, vol. 28, 1979, pp.100-108.

[5] Paul N., Terre M., Fety L., "A Global Algorithm to Estimate the Expectations of the Components of an Observed Univariate Mixture," *Advances in Data Analysis and Classification* vol. 1(3), 2007, pp.201-219.

[6] Terre M, Fety L., Paul N., "K-Produit : un critère de classification pour le traitement du signal," *Traitement du Signal*, vol. 27/2, 2010, pp. 221-239.

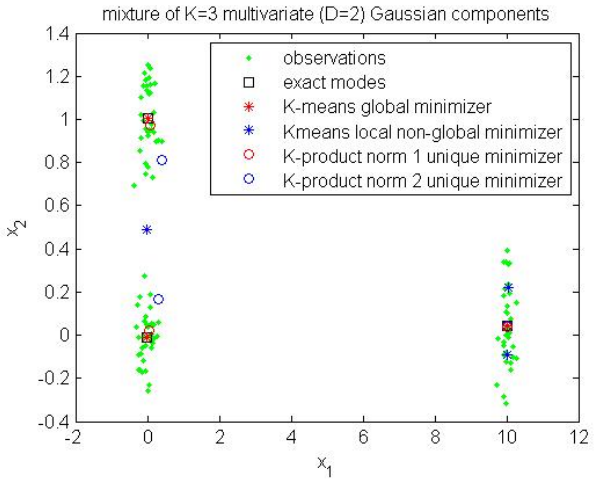

Figure 1: 100 observations of a mixture of K=3 Gaussian multivariate components (D=2) and locations of optimizers corresponding to different criterion. For the class on the right all solutions are superposed on the exact solution, except for the non-global *K-means*. For the classes on the left part of the figure, only the global *K-means* minimizer and the unique minimizer of the "*K-product norm 1*" lead to a pertinent solution. We have to be careful with this figure because horizontal and vertical scales are different.

Exact Means	*K-Means*, global minimizer	*K-Means*, non global minimizer	*K-Product norm 1*, unique minimizer
{0, 0, 0, 0}	{-0.02,-0.04, 0.00,-0.05}	{0.47, -0.02, -0.48, -0.04}	{0.03, -0.07, 0.05, 0.01}
{1, 0, -1, 0}	{1.01, 0.00, -1.01, -0.01}	{-0.05, -9.96, 0.20, 9.93}	{0.94 ,-0.06, -0.91, 0.06}
{1, -10, -1, 10}	{1.01, -10.04, 0.00, 10.03}	{0.02, -10.07, 0.11, 10.09}	{0.01, -10.04, -0.00, 10.05}

Table 1: Minimizers of *K-means* and *K-product norm 1* over 100 observations of a K=3 components of a Gaussian mixture (D=4). Covariance matrix being equals to $0{,}2^2\,\mathrm{I}$. All components of the mixture have the same probability.

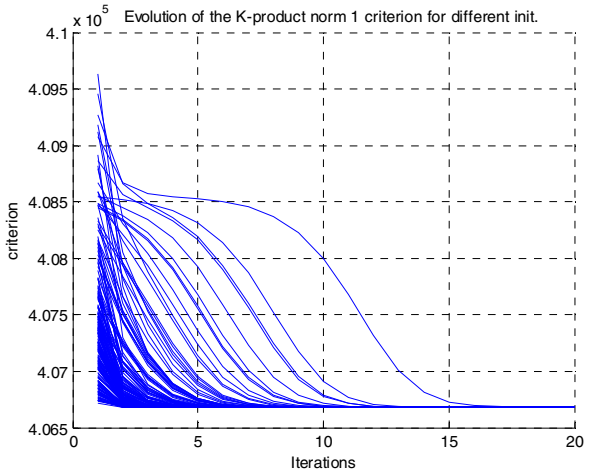

Figure 2: Evolution of the *K-product* criterion for different intializations. 100 observations of the gaussian mixture described in table 1 (K=3, D=4). Whatever the initialization is, the algorithm converges towards the same minimizer as indicated in the last column of table 1.

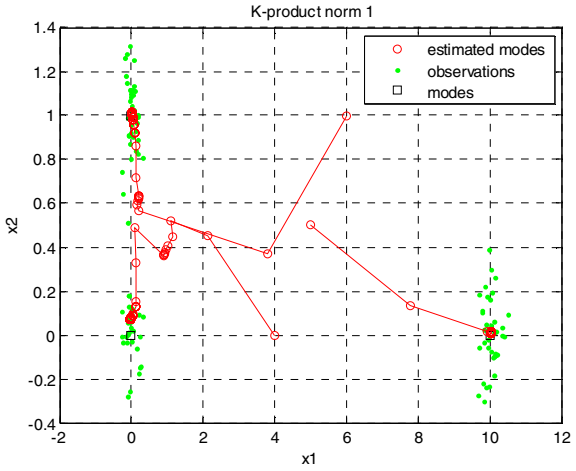

Figure 3: Evolution of the *K-product norm 1* estimated modes given by the relaxation algorithm. 100 observations of a multivariate Gaussian mixture (K=3, D=2). Estimated modes converge in less of 50 iterations towards the exact modes with a reduced bias. Covariance matrix being equal to $0{,}15^2\,\mathrm{I}$.

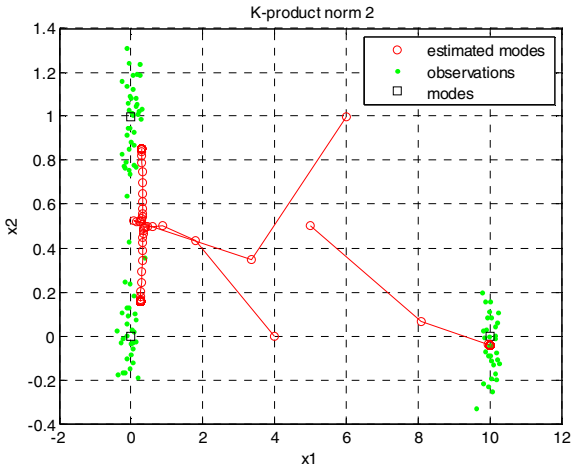

Figure 4: Evolution of the *K-product norm 2* estimated modes given by the relaxation algorithm. 100 observations of a multivariate Gaussian mixture (K=3, D=2). Estimated modes converge in less of 50 iterations towards the exact modes with an important bias. Covariance matrix being equal to $0{,}15^2$ I.

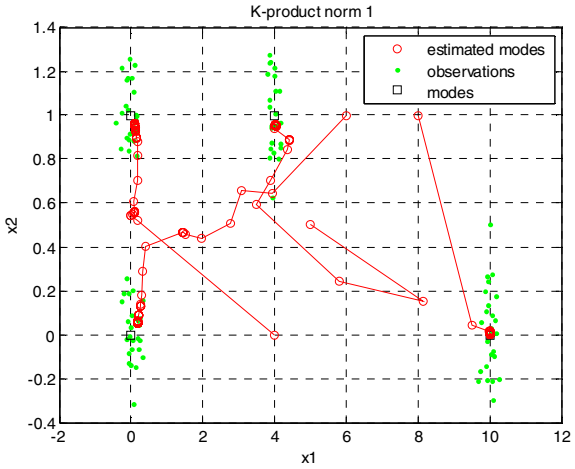

Figure 5: Evolution of the *K-product norm 1* estimated modes given by the relaxation algorithm. 100 observations of a multivariate Gaussian mixture (K=4, D=2). Estimated modes converge in less of 50 iterations towards the exact modes with a reduced bias. Covariance matrix being equal to $0{,}15^2$ I.

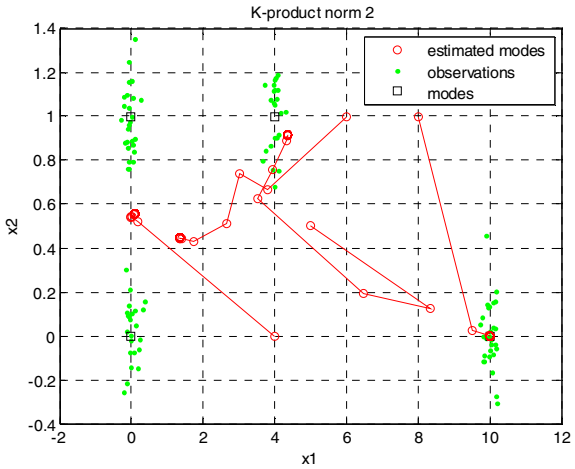

Figure 6: Evolution of the *K-product norm 2* estimated modes given by the relaxation algorithm. 100 observations of a multivariate Gaussian mixture (K=4, D=2). Estimated modes converge in less of 50 iterations towards the exact modes with an important bias in particular for modes on the left part of the figure. Covariance matrix being equal to $0{,}15^2$ I.