Evolution du critère « K-Produit » pour l'estimation des mélanges de lois

NICOLAS PAUL[1], ALEXANDRE GIRARD[1], MICHEL TERRE[2]

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

[1]nicolas.paul@edf.fr, alexandre.girard@edf.fr,
[2]terre@cnam.fr

Résumé – On s'intéresse ici à la conception d'un nouveau critère d'optimisation pour l'estimation des mélanges de lois. Une évolution du critère « K-Produit » initialement proposé dans [5] est présentée. Dans le cas monodimensionnel on démontre que le critère ne possède aucun minimiseur local non-global. Ce résultat est également observé en dimension supérieur. La pertinence du minimiseur unique pour l'estimation des mélanges de lois est étudiée dans le cas simplifié d'un mélange de lois uniformes séparées, pour lequel les écarts asymptotiques respectifs entre les trois éléments du minimiseur (unique) du critère et les trois espérances des composantes sont inférieurs à la moitié de l'écart-type des composantes. La pertinence du minimiseur est également observée par simulation dans des cas multidimensionnels.

Abstract – We focus on the definition of new optimization criteria for mixture estimation. An evolution of the K-Product criterion [5] is proposed. In the univariate case we show that the criterion does not have any local non-global minimizer. This property is also observed in the multivariate case. The relevance of the new K-Product criterion is studied with simulation (in some multivariate case) and theory : in particular we show that for a mixture of three separate uniform component the distance between the criterion unique minimizer and the true component expectations is less than half the components standard deviation.

1 Introduction

L'objet de cette communication est l'estimation non supervisée des paramètres d'un mélange de loi. On dispose d'un ensemble de N observations multidimensionnelles $\{\mathbf{x}_n\}$ d'un mélange de K distributions. La densité de probabilité des observations s'écrit :

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

avec h_k (resp. p_k) la distribution (resp. le poids) de la k-ième composante. Le nombre K de composantes est supposé connu. L'objectif est de chercher à estimer les K espérances $\{\mathbf{a}_k\}$ données par :

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

Dans le cas fréquent où les moyennes coïncident avec les K modes principaux de $\mathrm{h}(\mathbf{x})$, des approches non-paramétriques peuvent être utilisées pour estimer les K modes de la distribution des observations. Ces méthodes ne font pas d'hypothèses sur la forme des composantes du mélange. Elles consistent à partitionner l'espace d'observations IR^{D} en hyper-cubes pour estimer un échantillonnage de la densité de probabilité (ddp) $\mathrm{h}(\mathbf{x})$. Une fonction noyau (typiquement une Gaussienne) est associée à chaque observation. $\mathrm{h}(\mathbf{x})$ est ensuite estimée en ajoutant la contribution de chaque noyau sur chaque hyper-cube. Les K modes principaux de la ddp estimée sont ensuite calculés [1]. Cette méthode présente plusieurs inconvénients, en particulier le réglage délicat de la taille des hyper-cubes et de largeur des fonctions noyaux. Si le nombre d'observations n'est pas suffisant, la ddp estimée peut de plus contenir de nombreux modes qui ne correspondent pas aux modes de la distribution d'origine. La convergence des modes de la ddp estimée vers les modes de la vraie ddp est étudiée dans [2].

L'approche paramétrique, plus répandue, consiste à supposer que les différentes composantes du mélange sont modélisables par des fonctions paramétrées, typiquement des Gaussiennes, et à rechercher l'ensemble des paramètres qui maximisent la vraisemblance des observations. L'algorithme EM (« espérance – maximisation », en anglais : « expectation-maximisation ») [3], plus généralement dédié à l'estimation des problèmes à données manquantes, est souvent utilisé pour trouver ces paramètres. Dans le cas d'un mélange Gaussien équiprobable où les supports des composantes sont séparés et où les composantes ont la même variance, la fonction de log-vraisemblance est très proche du critère K-Means [4], très fréquemment utilisé dans les problèmes de classification non supervisée et de quantification vectorielle :

$$\mathrm{J}_{\mathrm{K-Means}}(\mathbf{u}_1, \ldots, \mathbf{u}_K) = \sum_{n=1}^{N} \min_k \|\mathbf{x}_n - \mathbf{u}_k\|^2 \quad . \quad (3)$$

L'inconvénient principal des méthodes EM ou K-Means est l'éventuelle convergence vers des points stationnaires qui ne sont pas des optimiseurs globaux,

par exemple des optimiseurs locaux non-globaux. Un cas typique est représenté Fig. 1 : les observations sont issues d'un mélange de trois lois Gaussiennes bidimensionnelles dans une configuration relativement « simple » : les composantes sont séparées, équiprobables et de même covariance. Même dans ces conditions, les algorithmes EM ou K-Means ne convergent pas nécessairement vers la bonne solution. Les résultats obtenus avec EM et K-Means étant sur cet exemple relativement similaires, seul les résultats obtenus avec K-Means sont présentés pour plus de clarté.

2 Critères K-Produit

2.1 Définition

Notre approche est sensiblement différente des approches de l'état de l'art décrite section 1. Elle consiste à rechercher le minimum de critères « K-Produit » définis par :

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

où $\varepsilon > 0$ est un paramètre de régularisation permettant au critère d'avoir un gradient continu sur IR^D . Lorsque ε=0 le critère devient simplement :

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

Ce critère est une évolution du critère K-Produit proposé dans [5], dont quelques applications sont présentées dans [6] :

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

La différence tient au choix de la norme utilisée dans le produit : norme 2 dans la proposition initiale (6), norme 1 (resp. norme 1 « régularisée ») dans la définition (5) (resp. (4)) proposée ici. L'inconvénient principal du critère initial, « K-Produit norme 2 », est l'important biais entre son minimiseur et les moyennes des composantes : lorsque le nombre d'observations tend vers l'infini, le minimiseur de (6) ne tend pas vers les moyennes des composantes. Par exemple, dans le cas d'un mélange à deux composantes unidimensionnelles, équiprobables, de moyennes $-a$ et a et de même écart-type σ, le minimiseur de (6) tend vers le couple $\left\{a\sqrt{1+\sigma^2/a^2}, -a\sqrt{1+\sigma^2/a^2}\right\}$. La figure 1 illustre un autre exemple de biais dans le cas du mélange de $K = 3$ composantes évoqué section 1.

2.2 Minimisation

Un algorithme de relaxation de complexité O(*NKD*) peut être utilisé pour minimiser ce critère. En effet, au voisinage d'une estimation courante $\left\{\mathbf{u}_1^{\text{ite}}, \dots, \mathbf{u}_k^{\text{ite-1}}, \dots, \mathbf{u}_K^{\text{ite-1}}\right\}$ au cours de l'itération « ite » le gradient du critère (4) par rapport au *k*-ème vecteur $\mathbf{u}_k$ a pour équivalence :

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

avec $D_{n,k} = \frac{C_{n,k}}{\left(\varepsilon + C_{n,k}\left\|\mathbf{x}_n - \mathbf{u}_k^{\text{ite-1}}\right\|^2\right)^{0.5}}$ et $C_{\text{n,k}} = \prod_{l=1}^{l=k-1}\left\|\mathbf{x}_n - \mathbf{u}_l^{\text{ite}}\right\|^2 \prod_{l=k+1}^{l=K}\left\|\mathbf{x}_n - \mathbf{u}_l^{\text{ite}-1}\right\|^2$. $C_{\text{n,k}}$ pouvant se mettre à jour par récurrence sur *k*, la mise à jour du *k*-ème vecteur $\mathbf{u}_k^{\text{ite}}$ par annulation de (7), donnée par :

$$\mathbf{u}_k^{\text{ite}} = \sum_{n=1}^{N}\frac{D_{n,k}}{\sum_{m=1}^{N} D_{m,k}}\mathbf{x}_n \ , (8)$$

s'obtient avec une complexité O(*ND*).

2.3 Minimiseurs

La Fig. 2 montre l'évolution du critère obtenue avec différentes initialisations sur un même ensemble de 100 observations du mélange décrit Tab 1. Quelque-soit l'initialisation, l'algorithme a convergé vers le même minimiseur.

Ce résultat peut être démontré dans le cas monodimensionnel, en utilisant la fonction $\mathbf{w}(\mathbf{u})$ qui a tout vecteur $\mathbf{u}$ de IR^K associe un vecteur contenant les *K* Polynômes Symétriques Elémentaires (PSE) de $\mathbf{u} = (u_1, \dots, u_K)$:

$$w_k(\mathbf{u}) = \sum_{j_1<\dots<j_k\le K} u_{j_1} \times \dots \times u_{j_k} \ . (10)$$

Une propriété importante de $\mathbf{w}$ est que si deux vecteurs $\mathbf{u}_A$ et $\mathbf{u}_B$ sont tels que $\mathbf{w}(\mathbf{u}_A) = \mathbf{w}(\mathbf{u}_B)$ alors $\mathbf{u}_A$ et $\mathbf{u}_B$ sont égaux à une permutation près.

Soit maintenant H_ε la fonction strictement convexe définit par :

$$H_\varepsilon(\mathbf{v}) = \frac{1}{N}\sum_{n=1}^{N}\sqrt{\varepsilon + (x_n^K - (x_n^{K-1}, \dots, x_n^0)^t\mathbf{v})^2} \ , (11)$$

Pour tout $\mathbf{u}$ de IR^K, on a $J_\varepsilon(\mathbf{u}) = H_\varepsilon(\mathbf{w}(\mathbf{u}))$. On montre que si $\mathbf{u}_{\min}$ est un minimiseur de J_ε alors $\mathbf{w}(\mathbf{u}_{\min})$ est le minimiseur unique de H_ε. Tous les minimiseurs de J_ε sont donc égaux à une permutation près.

3 Estimation des mélanges de lois

La pertinence du minimiseur de K-Produit pour l'estimation des espérances du mélange peut être évaluée par simulation (cas multidimensionnel), et théoriquement (cas monodimensionnel). On montre en particulier au paragraphe 3.1 que pour un mélange de trois lois uniformes séparées la distance entre les trois éléments du minimiseur unique de K-Produit et les trois espérances est bornée par l'écart-type des composantes.

3.1 Théorie

Dans le cas unidimensionel l'optimisation du critère asymptotique $J_0^\infty(u_1, \dots, u_K) = \int_{-\infty}^{\infty}\prod_{k=1}^{K}|x - u_k|h(x)dx$ conduit à des relations intégrales valables pour tout mélange de lois. Pour $K = 3$ composantes, l'annulation

des combinaisons du type $\frac{\partial J_0^\infty}{\partial u_k} - \frac{\partial J_0^\infty}{\partial u_l}$ et $u_k \frac{\partial J_0^\infty}{\partial u_k} - u_l \frac{\partial J_0^\infty}{\partial u_l}$ permet d'éliminer les termes en u_k des intégrales puis et ainsi d'obtenir les conditions d'optimalité suivantes pour k=0, 1, 2 :

$$\int_{u_1}^{u_2} x^k h(x)dx + \int_{u_3}^{\infty} x^k h(x)dx = \frac{E_h\{X^k\}}{2}. \quad (12)$$

La dépendance des relations (12) en les u_i n'apparaît donc que sur les intervalles des intégrales. Ces relations permettent aussi d'établir directement des résultats sur des conditions d'unicité de minima locaux.

Ces relations se simplifient dans le cas de mélanges de lois uniformes à supports séparés $[a_k - b \;\; a_k + b]$ de largeur commune $2b$, d'écart-type $\sigma = b/\sqrt{3}$ et de moyenne a_k. En repartant de (12) la solution optimale $\{u_1, u_2, u_3\}$ vérifie alors:

$$\begin{gathered} -u_1 + u_2 - u_3 = -a_1 + a_2 - a_3 \\ -u_1^2 + u_2^2 - u_3^2 = -a_1^2 + a_2^2 - a_3^2 - b^2 \\ -u_1^3 + u_2^3 - u_3^3 = -a_1^3 + a_2^3 - a_3^3 + 3b^2(-a_1 + a_2 - a_3). \quad (13) \end{gathered}$$

Après développement, ce système conduit aux relations suivantes pour les erreurs d'estimation $v_k = u_k - a_k$:

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

$$v_2 = v_1 + v_3$$

$$v_1 v_3 + v_1(a_2 - a_1) - v_3(a_3 - a_2) + \frac{b^2}{2} = 0, \quad (14)$$

On peut borner les erreurs d'estimations à partir du système (14) : Pour des composantes séparées, on a $0 < \frac{1}{(a_k - a_l)} < 1/2b$ si $k > l$ donc $|v_2| < b/4$ par la première relation. D'après la deuxième relation si v_1 et v_3 sont de même signe ils sont en valeur absolue inférieurs à $b/4$. S'ils sont de signes opposés la troisième relation de (14) impose $v_1 < 0$ et $v_3 > 0$ puis $|v_1|(a_2 - a_1) + |v_3|(a_3 - a_2) < b^2/2$ donc $|v_{1,3}| < b^2/2 \times 1/2b = b/4$. Finalement, pour k = 1, 2 et 3 :

$$|u_k - a_k| < \frac{b}{4} < \frac{\sigma}{2}, \quad (15)$$

L'écart entre le minimiseur du critère K-Produit asymptotique et les espérances des composantes est inférieur à la moitié de l'écart-type des composantes. Cette propriété, démontrée ici pour le cas simplifié d'un mélange univarié de composantes uniformes, est également observé dans le cas de mélanges Gaussien multidimensionnels (paragraphe 3.2).

3.2 Simulations

Des simulations permettent également de vérifier la pertinence du critère dans des cas plus complexes, en particulier dans le cas déjà évoqué (Fig. 1) ou pour des dimensions supérieures (Tab. 1). Ces différents résultats montrent clairement l'intérêt de notre approche « K-Produit norme 1 » par rapport aux méthodes existantes.

Dans ces cas relativement « simples » (K = 3 composantes identiques, équiprobables, séparées) les algorithmes classiques tels que EM ou K-Means ne convergent pas nécessairement vers la bonne solution, mais peuvent converger vers un optimiseur non global, avec un vecteur qui « englobe » les deux composantes proches et les deux vecteurs restant qui sont répartis dans la classe restante. Ces algorithmes doivent être relancés plusieurs fois, avec différentes initialisations, afin d'être sûr d'obtenir une solution correcte.

4 Conclusions

Une évolution du critère « K-Produit » initialement proposé dans [5] est proposée pour estimer les espérances d'un mélange de loi lorsque le nombre de composante est connu. La pertinence du critère est étudié théoriquement. On montre en particulier, dans le cas unidimensionnel, que le critère n'admet aucun minimiseur local non-global et que, pour un mélange de trois lois uniformes séparées de largeur de support commune, l'écart entre le minimiseur unique et les espérances que l'on cherche à estimer est inférieur à la moitié de l'écart-type des composantes.

La validité de ces résultats dans le cas multidimensionnel, que suggèrent les simulations, est en cours d'étude, ainsi que l'extension au cas où le nombre de composantes n'est pas connu.

Références

[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 33:1065-1076 (1962)

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

[4] Hartigan J., Wong M., A K-Means Clustering Algorithm. Journal of Applied Statistics, 28:100-108 (1979)

[5] Paul N., Terre M., Fety M., A Global Algorithm to estimate the expectations of the components of an observed univariate mixture. Advances in Data Analysis and Classification 1(3):201-219 (2007)

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

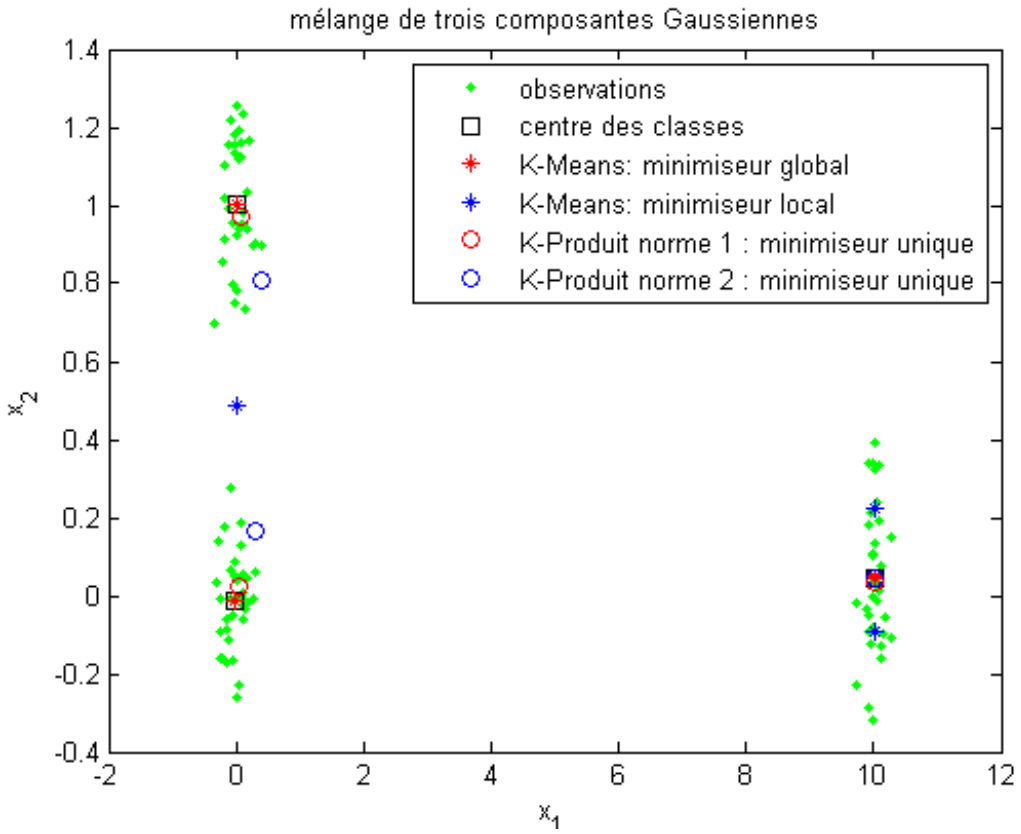

Figure 1 : 100 observations d'un mélange de *K*=3 composantes Gaussiennes et localisation des optimiseurs de différents critères : sur la classe de droite toutes les solutions sont superposées à la solution exactes, excepté l'optimiseur non global de K-Means ; sur les deux classes de gauche seuls le minimiseur global de K-Means et le minimiseur unique de « K-Produit norme 1 » donne une solution pertinente. Attention à la différence d'échelle entre l'axe des abscisses et l'axe des ordonnées.

Tab 1 : minimiseurs de K-Means et « K-Produit norme 1 » sur 100 observations d'un mélange de 3 composantes Gaussiennes, équiprobables et de matrice de variance-covariance $0.2^2\mathbf{I}$

Moyennes réelles	K-Means minimiseur global	K-Means minimiseur non-global	K-Produit norme 1 minimiseur unique
{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}

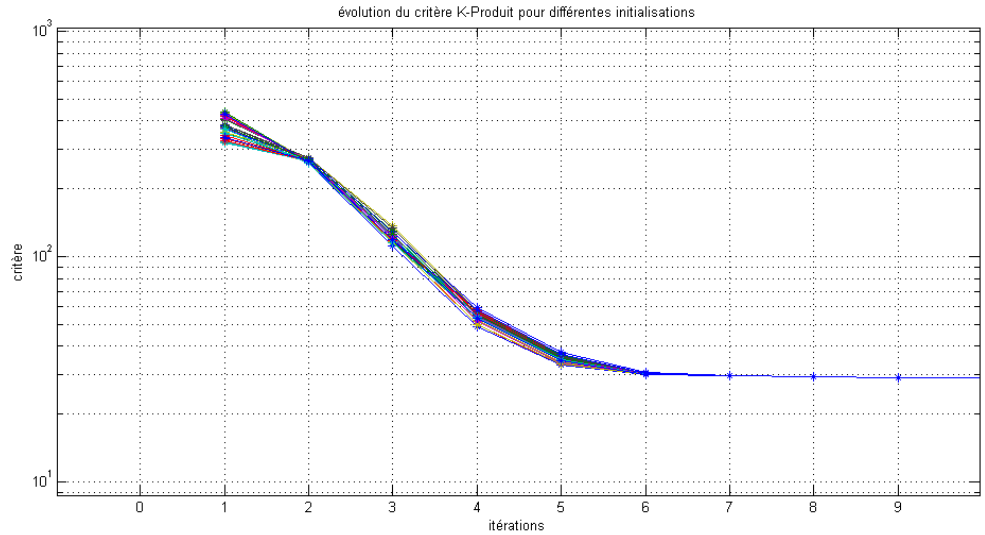

Figure 2 : évolution du critère K-Produit (4), pour différentes initialisations, sur 100 observations du mélange Gaussien décrit Tab. 1 (D=4, K=3). Quelque-soit l'initialisation, l'algorithme converge vers le même minimiseur indiqué Tab. 1 (dernière colonne)