

Estimation de mélanges de lois : évolution du K-Produit vers le critère KP-log

ALEXANDRE GIRARD¹, NICOLAS PAUL¹, JEROME IDIER²

¹ EDF R&D département STEP
6, quai Watier, 78401 Chatou, France

² Institut de Recherche en Communication et Cybernétique de Nantes (IRCCyN, UMR CNRS 6597)
1 rue de la Noë, B.P. 92101, 44321 Nantes Cedex 3.

¹ alexandre.girard@edf.fr, nicolas.paul@edf.fr,
² jerome.idier@irccyn.ec-nantes.fr

Résumé – Cet article s’intéresse à proposer un nouveau critère d’optimisation qui permette de résoudre des problèmes d’estimation posés par l’application du K-produit décrit dans [6] sur certains mélanges en monodimensionnel ou en multidimensionnel. Ce nouveau critère est testé sur des cas illustrant ces problèmes et une application de classification d’images.

Abstract - We focus on the definition of a new optimization criterion for mixture estimation which allows solving some estimation difficulties seen with K-product algorithm presented in [6] in univariate or multivariate cases. This new criterion is tested on the problematic cases and an application in image classification is proposed.

1 Introduction

Cette communication traite de l’estimation non supervisée des espérances d’un mélange de K lois de distributions inconnues. On dispose pour cela d’un ensemble de N observations multidimensionnelles $\{\mathbf{x}_n\}$.

Dans [6] a été détaillée la liste des méthodes classiques d’estimation [1-4] telles que EM et K-Means avec leur principale limite qui est l’existence de minima locaux qui impose d’appliquer l’algorithme à partir d’un nombre suffisant d’initialisations.

Un nouveau critère d’optimisation K-Produit a été ensuite proposé, critère qui s’écrit avec la norme euclidienne :

$$J_{KP,N}(u_1, \dots, u_K) = \sum_{n=1}^N \prod_{k=1}^K \|x_n - u_k\|$$

L’unicité du minimiseur a été démontrée dans le cas monodimensionnel (cf. [6]). Ce minimiseur fournit des approximations des espérances des lois avec un objectif de « consistance » : dans notre contexte de classification, un ensemble de K vecteurs (u_1, \dots, u_K) est dit « consistant » si chaque support de classe contient un vecteur u_k . En asymptotique, ce critère s’écrit :

$$J_{KP}(u_1, \dots, u_K) = \int_{\mathbb{R}} \prod_{k=1}^K \|x - u_k\| h(x) dx$$

En 1-D, ce critère, pour lequel il a pu être démontré que la « consistance » est atteinte pour des mélanges de trois lois uniformes, n’est en revanche pas robuste pour certains mélanges de quatre lois, ce que nous illustrerons dans la suite de l’article.

Par ailleurs, dans le cas d -D, il peut présenter au moins un minimum local non global. La première partie de l’article détaillera ces points.

Pour pallier ces problèmes, une adaptation du critère a été réalisée en gardant le principe de conserver la forme suivante :

$$J_N(u_1, \dots, u_K) = \sum_{n=1}^N \prod_{k=1}^K \psi(x - u_k)$$

avec $\psi(y)$ présentant un unique minimum en 0 et tendant vers l’infini quand y tend vers l’infini.

C’est cette adaptation qui fera l’objet de la deuxième partie de l’article. Une méthode de résolution par point fixe sera présentée et un exemple numérique illustrera les gains acquis par rapport à K-Produit.

Les résultats obtenus permettent aussi de présenter une application en classification d’images. Ce sera l’objet de la troisième partie de l’article.

2 Limitations du K-Produit

Comme indiqué dans l’introduction, le K-Produit présente quelques limitations qui vont être décrites dans cette partie.

2.1 Mélange de quatre lois

Dans [6], il a été démontré que K-Produit peut traiter un mélange de trois lois uniformes de façon inconditionnelle. En revanche, la « consistance » n’est plus garantie pour l’application de K-Produit à un mélange de quatre lois dont le quatrième support est suffisamment éloigné des trois autres. Une hypothèse pouvant expliquer ce phénomène est que la prise en compte des éléments de l’échantillon qui sont loin du support concerné de façon linéaire peut conduire à créer des biais. Ceci sera illustré dans les applications numériques de KP-log.

2.2 Application du K-Produit en d -D

En n dimensions, n supérieur à 2, des simulations ont fait apparaître pour certains cas de mélanges un point singulier où tous les u_k sont égaux. Selon les mélanges de lois à estimer, ce point singulier peut s’avérer être stable et être ainsi un minimum local non global. Selon

l'initialisation de l'algorithme, les itérations successives peuvent passer par ce bassin d'attraction et converger vers ce point singulier. Pour contourner le problème, un terme de régularisation peut être ajouté pour forcer les itérations à ne pas s'approcher de points à composantes égales, mais cela demande l'introduction d'un hyperparamètre de régularisation correspondant au diamètre de ce bassin d'attraction, et son réglage n'est pas facilement déterminable. La Figure 1 montre un exemple d'un tel point singulier qui se trouve en dehors des échantillons des lois et sur lequel l'algorithme peut converger selon l'initialisation. Sur 1000 initialisations, ce minimum local est apparu 180 fois.

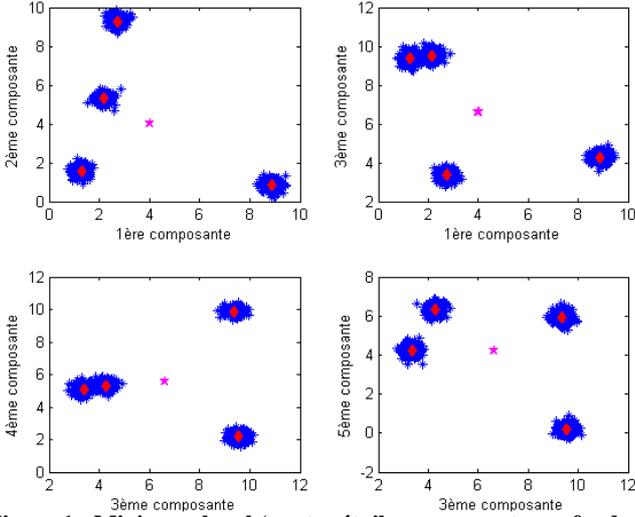


Figure 1 : Minimum local (quatre étoiles en magenta confondues en dehors des échantillons) avec K-Produit sur un mélange de quatre lois en dimension 5 et minimum global (losanges rouges)

3 Le nouveau critère KP-log

Dans cette partie, nous allons présenter le nouveau critère proposé. Mais, de façon préliminaire, nous allons décrire quelles sont les propriétés que devrait avoir *a priori* l'estimateur idéal, pour ensuite le traduire le mieux possible en un critère d'optimisation.

3.1 Propriétés d'un estimateur idéal

Revenons d'abord aux estimateurs existant. K-means est en un certain sens un quasi-oracle : s'il est possible de dire à quelle loi appartient tel élément de l'échantillon à classifier, alors la moyenne des éléments affectés par loi donne la meilleure estimation possible de l'espérance de la loi. La fonction-coût exploitant le minimum de distance pour l'affectation d'un élément au centre de classe ne permet pas d'effectuer une descente de gradient sur tout l'espace pour trouver des estimations des espérances de lois. Il s'en suit une caractéristique connue de K-means : le minimum obtenu est local.

L'apport essentiel de K-produit est de fournir une réponse à la question de l'estimation d'approximations des espérances avec une méthode rapide qui repose sur une descente de gradient, moyennant les limitations présentées dans la partie précédente.

En résumé, il faut un critère qui soit convexe pour faire une descente de gradient avec un seul minimum et qui assure une « consistance » inconditionnelle.

3.2 Définition d'un nouveau critère

La moyenne d'une loi donnée h peut être vue classiquement comme le minimum de la fonction-coût J suivante en norme euclidienne :

$$J(u) = \frac{1}{2} \int_{\mathbb{R}^d} \|x - u\|^2 h(x) dx$$

Par analogie, dans le cas d'un mélange de lois disjointes où l'on veut déterminer chaque espérance, cette intégrale devrait n'être réalisée que sur le support de la loi élémentaire considérée. Un moyen pour cela est d'introduire une fonction-seuil pour imposer que l'impact de la distance de x à un point u_i , $\|x - u_i\|$, soit négligeable au-delà d'une certaine valeur. Il faut maintenant définir une fonction-seuil qui permette de revenir ensuite à notre objectif d'optimisation convexe.

Une solution est d'introduire l'équivalent d'une fonction-seuil κ sur la distance :

$$\kappa(y) = \frac{2}{1 + \exp(\gamma^{-1} \phi(y))}$$

avec ϕ fonction d'énergie paire, donc $\phi'(0) = 0$, avec de plus $\phi(0) = 0$, $\phi(y) \xrightarrow{|y| \rightarrow +\infty} +\infty$ et ϕ dérivable continûment deux fois, et γ paramètre à ajuster.

3.2.1 Démarche de définition du critère

Nous allons présenter la démarche conduisant au critère final en 1D pour alléger les notations.

La fonction-seuil κ permet de transformer $(x - u)$ en $(x - u)\kappa(x - u)$ et $\frac{1}{2}(x - u)^2$ en $\int_u^x \kappa(\xi - u)(\xi - u) d\xi$.

La fonction-coût complète pour le calcul des différentes espérances s'écrit alors :

$$J(u_1, \dots, u_K) = \frac{1}{2} \int_{-\infty}^{+\infty} \prod_{k=1}^K \int_{u_k}^x \frac{2(\xi - u_k)h(x)}{1 + \exp\left(\frac{\phi(\xi - u_k)}{\gamma}\right)} d\xi_k dx$$

Dans cette fonction-coût, seuls les x proches des u_k induisent une variation de la fonction-coût pour une variation de u_k . La présence des exponentielles et des intégrales limite son utilisation en optimisation.

L'idée est alors de réaliser un développement limité de la fonction-seuil au premier ordre. D'où :

$$1 + \exp\left(\frac{\phi(\xi_k - u_k)}{\gamma}\right) \approx 1 + 1 + \frac{\phi''(0)}{2} (\xi_k - u_k)^2$$

Cette expression permet de calculer explicitement l'intégrale entre u_k et x indépendamment de l'expression de ϕ dès lors que celle-ci satisfait aux conditions posées :

$$\int_{u_k}^x \frac{2(\xi_k - u_k) d\xi_k}{1 + \exp(\gamma^{-1} \phi(\xi_k - u_k))} \approx \frac{2\gamma}{\phi''(0)} \ln \left(1 + \frac{\phi''(0)(x - u_k)^2}{4\gamma} \right)$$

3.2.2 Nouveau critère de classification

Le nouveau critère résultant, baptisé KP-log, s'écrit :

$$J_{KP-\log}(u_1, \dots, u_K) = \int_{\mathbb{R}^d} \prod_{k=1}^K \ln \left(1 + \frac{\|x - u_k\|^2}{\alpha^2} \right) h(x) dx$$

en utilisant la norme euclidienne et, en discret, :

$$J_{KP-\log, N}(u_1, \dots, u_K) = \frac{1}{N} \sum_{n=1}^N \prod_{k=1}^K \ln \left(1 + \frac{\|x_n - u_k\|^2}{\alpha^2} \right)$$

avec

$$\alpha^2 = \frac{1}{K^2} \int_{\mathbb{R}^d} \|x - E(h)\|^2 h(x) dx \quad (2)$$

Ce critère s'inscrit dans la suite des précédents en ce que, d'une part, la notion de produit est conservée, ce qui permet de faire du calcul différentiel, et, d'autre part, comme $\ln \left(1 + \alpha^{-2} \|x - u_k\|^2 \right) \approx \alpha^{-2} \|x - u_k\|^2$ pour x voisin de u_k , le critère se rapproche du K-means dans ce voisinage, les autres termes variant assez peu.

La différence essentielle tient au fait que, par construction, l'évolution à l'infini du critère est beaucoup plus lente que celle du K-Produit.

3.3 Méthodes numériques

L'annulation du gradient de $J_{KP-\log}$ pour tout l de $\{1, \dots, K\}$ donne :

$$\int \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{\|x - u_k\|^2}{\alpha^2} \right) \frac{(x - u_l) h(x)}{\alpha^2 + \|x - u_k\|^2} dx = 0 \quad (3)$$

Une procédure par descente de gradient pourrait être utilisée. Il est proposé ici d'exploiter un principe de point fixe. Pour cela, nous réécrivons (3) comme suit :

$$u_l = \frac{\int \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x - u_k)^2}{\alpha^2} \right) \frac{x}{\alpha^2 + (x - u_l)^2} h(x) dx}{\int \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x - u_k)^2}{\alpha^2} \right) \frac{1}{\alpha^2 + (x - u_l)^2} h(x) dx} \quad (4)$$

Ceci crée une relation itérative. Il est possible de démontrer que pour un choix de α suffisamment grand, cette application est contractante en considérant des itérations sur chaque u_l . En parcourant successivement les u_l , la convergence s'effectue vers un point fixe global. En pratique, les u_l sont modifiés simultanément.

En faisant une analogie statistique, cela veut dire que chaque u_l est égal à l'espérance conditionnelle sachant les u_k de la loi de densité de probabilité :

$$g_l(x|u_1, u_{k \neq l}) = \frac{\prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x - u_k)^2}{\alpha^2} \right) \frac{h(x)}{\alpha^2 + (x - u_l)^2}}{\int \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x - u_k)^2}{\alpha^2} \right) \frac{h(x)}{\alpha^2 + (x - u_l)^2} dx}$$

Par ailleurs, les applications vont s'appuyer sur un échantillon fini noté (x_n) . L'équation (3) s'écrit alors :

$$u_l = \frac{\sum_{n=1}^N \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x_n - u_k)^2}{\alpha^2} \right) \frac{x_n}{\alpha^2 + (x_n - u_l)^2}}{\sum_{n=1}^N \prod_{\substack{k=1 \\ k \neq l}}^K \ln \left(1 + \frac{(x_n - u_k)^2}{\alpha^2} \right) \frac{1}{\alpha^2 + (x_n - u_l)^2}}$$

3.4 Classification des éléments de l'échantillon

L'application pratique de ces développements se faisant sur un échantillon, un point intéressant est de pouvoir affecter à chaque loi les éléments correspondant pour pouvoir faire une estimation de l'espérance de type K-means. Pour cela, le fait que chaque g_l soit une densité de probabilité permet de comparer les valeurs de g_l pour un point x_n donné et de l'affecter à la loi l pour laquelle le maximum est atteint.

3.5 Résultats

Les résultats vont être focalisés sur les points limitant de K-Produit.

3.5.1 Mélange de quatre lois en 1D

La Figure 2 montre l'application des algorithmes K-Produit et KP-log sur 4000 observations issues d'un mélange de quatre lois où la dernière est éloignée des trois autres.

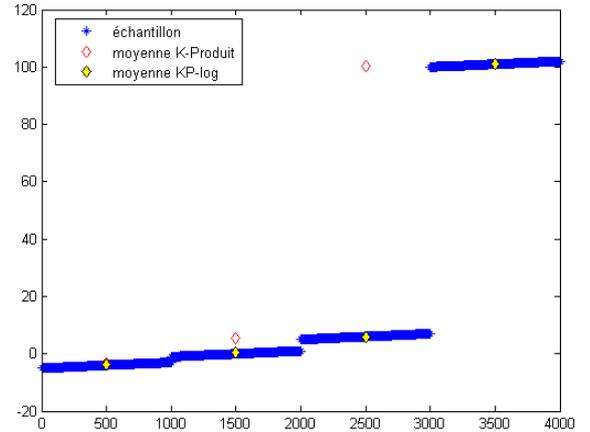


Figure 2 : Comparaison des performances de K-Produit et de KP-log sur un mélange de quatre lois

1000 initialisations différentes ont été testées, confirmant l'unicité des minima des critères. Les estimations des moyennes issues de KP-log se confondent avec les moyennes réelles des lois. La moyenne de la troisième loi estimée avec K-Produit est très biaisée et l'estimation de la deuxième loi est un peu biaisée et hors support.

3.5.2 Cas en d -D

KP-log est appliqué ici sur l'échantillon qui avait mis en défaut, selon l'initialisation, K-Produit. Les moyennes estimées correspondent aux espérances réelles. Il est aussi montré sur cet échantillon qu'il existe une faible sensibilité sur le choix de α autour de la définition choisie (cf. équ. 2). La Figure 3 présente aussi le résultat de la séparation du mélange de lois (une couleur par loi) en fonction du critère de classification présenté.

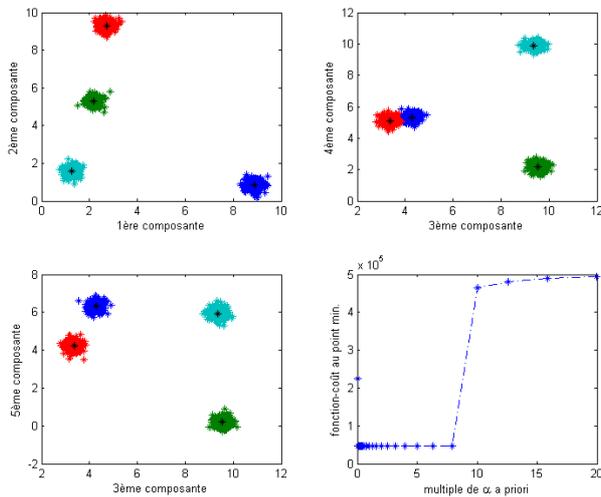


Figure 3 : Résultat sur un échantillon à quatre classes en dimension cinq – sensibilité au choix de α

4 Application sur une image

Une application classique de la classification est la séparation de zones sur une image. Les exemples pris ici proviennent d’une bibliothèque Internet d’images de ciel (<http://www.museumtextures.com>). L’image de la Figure 4 est un exemple de classification binaire couleur. L’image analysée Figure 5 présente à la fois une rupture franche entre le disque solaire et le ciel, et des nuances plus continues. La Figure 6 montre plus précisément la classification effectuée et en particulier les séparations non franches.

5 Conclusions

Le nouveau critère proposé fournit des réponses à des problèmes identifiés de K-Produit sur la consistance du minimiseur et, en dimension supérieure à 2, son unicité. Son utilisation permet de plus d’affecter les éléments de l’échantillon aux espérances de lois correspondantes.

L’illustration sur des images de ciel permet de montrer son utilisation sur une application concrète.

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., Terré 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] Paul N., Girard A., Terré M., Evolution du critère « K-Produit » pour l’estimation des mélanges de lois, GRETSI 2013, Brest, 2013

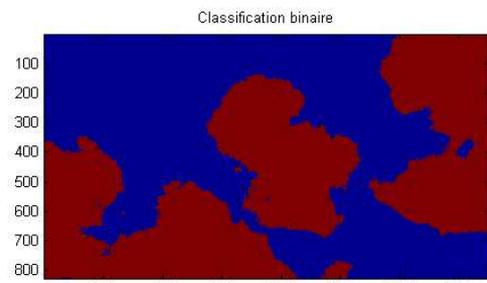
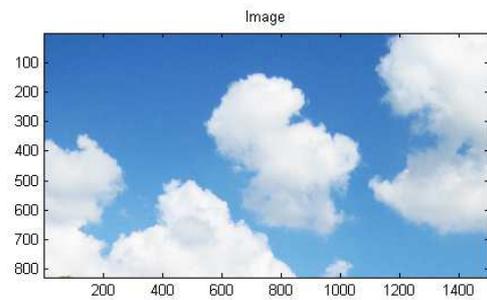


Figure 4 : Classification binaire avec KP-log

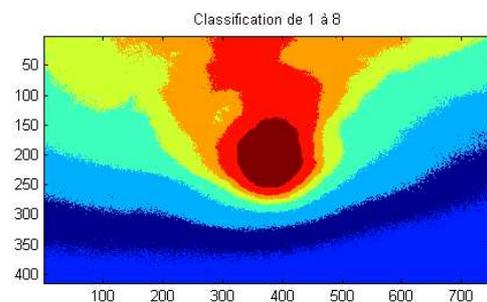
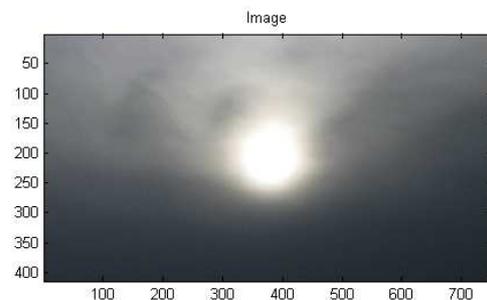


Figure 5 : Classification multiple avec KP-log

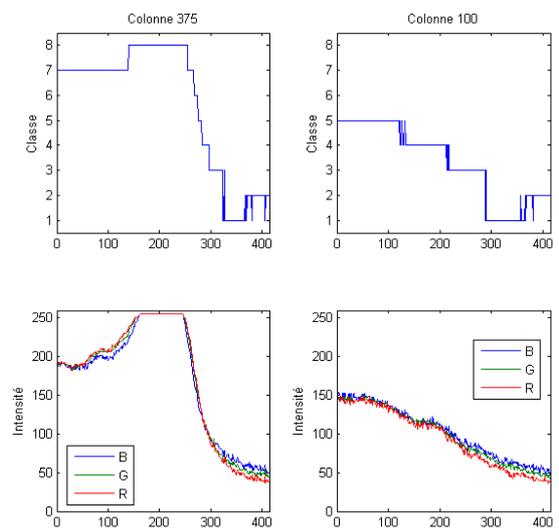


Figure 6 : Comparaison classification – image sur deux colonnes