# Simplification de modèles de mélange issus d'estimateur par noyau

Olivier SCHWANDER[1], Frank NIELSEN[1,2],

[1]École Polytechnique, Palaiseau, France

[2]Sony CSL, Tokyo, Japon

schwander@lix.polytechnique.fr, nielsen@lix.polytechnique.fr

**Résumé –** Les modèles de mélange gaussiens sont un outil universel pour modéliser des densités de probabilité complexes et variées. Ils peuvent être estimé grâce à l'Espérance-Maximisation ou par des estimateurs par noyaux. l'Espérance-Maximisation produit des modèles compacts mais parfois coûteux à produire alors que les estimateurs par noyaux produisent des modèles peu coûteux mais avec beaucoup de composantes dans le mélange. Dans cet article, nous présentons une nouvelle méthode pour obtenir des modèles de haute qualité qui sont à la fois compacts et rapides à produire. Nous utilisons des méthodes de partitionnement et des calculs de centroïdes. L'efficacité de notre approche est évaluée en terme de log-vraisemblance.

**Abstract –** Gaussian mixture models are a widespread tool for modeling various and complex probability density functions. They can be estimated using Expectation–Maximization or Kernel Density Estimation. Expectation–Maximization leads to compact models but may be expensive to compute whereas Kernel Density Estimation yields to large models which are cheap to build. In this paper we present new methods to get high-quality models that are both compact and fast to compute. This is accomplished with clustering methods and centroids computation. The quality of the resulting mixtures is evaluated in terms of log-likelihood.

## 1 Introduction

Les modèles statistiques sont très répandus dans le traitement du signal. On a principalement deux choix pour modéliser un jeu de données : une approche semi-paramétrique avec un mélange obtenu par l'algorithme Espérance-Maximisation (EM) ou une méthode non-paramétrique utilisant les fenêtres de Parzen (ou estimateurs par noyau).

Dans le premier cas, il faut choisir le nombre de composantes du mélange (soit manuellement, soit en l'apprenant) mais on obtient une représentation compacte et facile à utiliser. Dans l'autre cas, les fenêtres de Parzen décrivent de façon précise la distribution empirique mais cette précision s'obtient au prix d'un modèle très grand.

Nous proposons ici une nouvelle méthode (basée sur l'algorithme des $k$-moyennes) pour simplifier un modèle produit par un estimateur par noyau en utilisant une distance adaptée aux noyaux gaussiens. Nous utilisons ici la divergence de Kullback-Leibler, un outil classique pour comparer deux distributions, et la divergence de Fisher-Rao, justifiée par la géométrie hyperbolique des distributions gaussiennes. Comme nous ne connaissons pas de formule close pour le centroïde de Fisher-Rao, nous utiliserons plutôt une approximation appelée centroïde modèle. Enfin, nous introduisons une méthode non-itérative de simplification, qui consiste à n'effectuer qu'une seule étape de l'algorithme de simplification. Nos expériences montrent que notre approche produit des modèles d'aussi bonne qualité qu'EM tout en étant beaucoup plus rapide.

## 2 Modèles de distributions empiriques

### 2.1 Modèles de mélanges

Les modèles de mélange sont un outil très commun pour modéliser des données complexes dans de nombreux domaines, du traitement d'image au domaine biomédical en passant la reconnaissance de la parole. Ce succès vient de la capacité des modèles de mélanges (souvent gaussiens, mais pas toujours) à approximer des fonctions de densités associées à des variables aléatoires complexes. Pour un mélange $f$ à $n$ composantes, la fonction de densité s'écrit :

$$f(x) = \sum_{i=1}^{n} \omega_i \; g(x;\, \mu_i, \sigma_i^2)$$

où $\omega_i$ est le poids de la $i$-ième composante ($\sum \omega_i = 1$). Chaque composante $g(x;\, \mu_i, \sigma_i^2)$ est ici une distribution normale.

### 2.2 Estimateurs par noyaux

Un tel mélange peut être obtenu avec l'algorithme Espérance-Maximisation (EM) qui maximise de façon itérative la log-vraisemblance du mélange. Une autre possibilité est d'utiliser la méthode des fenêtres de Parzen (ou estimateur par noyaux). qui estime la fonction de densité connue par $N$ échantillons à l'aide d'une somme de $N$ noyaux (en général, gaussiens). Chaque gaussienne est centrée sur un échantillon et possède une matrice de variance-covariance fixée, contrôlée par un paramètre baptisé fenêtre. Ce paramètre, qui correspond au degré

de lissage du mélange, est déterminant pour la qualité de l'approximation.

Des choix populaires pour la fenêtre $h$ sont le coefficient de Silverman ([8]) et la méthode de Sheater et Jones ([4]).

### 2.3 Avantages et inconvénients

Construire un estimateur par noyaux est très peu coûteux mais conduit à des modèles très grands, puisque l'on a une composante par point. Sur un jeu de données typique, le modèle peut devenir inutilisable (par exemple, estimer la courbe de niveaux de gris d'une image $120\times120$ donne un mélange à 14400 composantes) : simplement évaluer la fonction de densité, ou encore calculer la log-vraisemblance est trop coûteux dès que le temps est un paramètre critique. L'algorithme Espérance-Maximisation produit au contraire des mélanges compacts, au risque de converger vers un optimum local et au prix d'un nombre d'itérations parfois important (mais ce prix n'est payé que lors de la phase d'apprentissage, et non pas lors de l'utilisation du modèle). Puisque des modèles compacts issus d'EM ont prouvé leur efficacité, il serait intéressant de construire des mélanges avec peu de composantes, tout en évitant les coûteuses itérations d'EM.

## 3 Simplification et centroïdes

### 3.1 Simplification d'estimateurs par noyaux

Étant donné un estimateur par noyaux, la méthode la plus simple pour diminuer le nombre de composantes est d'effectuer un partitionnement des noyaux à l'aide d'un algorithme dérivé de $k$-moyennes. Goldberger et Roweis [3] ainsi que Garcia *et al.* [2] ont proposé de telles méthodes, utilisant la divergence de Kullback-Leibler pour comparer les composantes du mélange. De plus, à la fois la divergence et les centroïdes sont connus en formule close : cela permet un calcul efficace des étapes du $k$-moyennes.

Bien que ces méthodes donnent expérimentalement de bons résultats (Figure 1) le problème du nombre d'itérations demeure. La solution que nous proposons est de n'effectuer que l'étape d'initialisation suivie d'une seule étape de calcul. En effet, dans beaucoup d'applications, il n'est pas nécessaire de trouver le meilleur modèle, mais seulement un modèle suffisamment bon pour l'application considérée.

### 3.2 Divergence de Kullback-Leibler

La divergence de Kullback-Leibler (KLD) mesure l'entropie relative entre deux distributions. Pour les gaussiennes, on a

$$\begin{aligned}\mathrm{KLD}(f_p, f_q) &= \frac{1}{2}\log\left(\frac{\det\Sigma_p}{\det\Sigma_q}\right)\\ &\quad+\frac{1}{2}\mathrm{tr}(\Sigma_q^{-1}\Sigma_p)\\ &\quad+\frac{1}{2}(\mu_q-\mu_p)^T\Sigma_q^{-1}(\mu_q-\mu_p)-\frac{d}{2}\end{aligned}$$

Cette divergence n'est pas symétrique, on a donc plusieurs sortes de centroïdes :

– le centroïde droit :

$$\arg\min_c \sum_i \omega_i KLD(c, x_i)$$

– le centroïde gauche :

$$\arg\min_c \sum_i \omega_i KLD(x_i, c)$$

– le centroïde symmétrisé :

$$\arg\min_c \sum_i \omega_i \frac{1}{2}\mathrm{SKL}(x_i, c)$$

où $c$ parcourt l'ensemble des distributions gaussiennes et où $\omega_i$ est le poids associé à la gaussienne $p_i$ ($\sum\omega_i = 1$ et $\omega_i > 0$).

SKL est ici la divergence de Kullback-Leibler symmétrisée :

$$\mathrm{SKL}(p,q) = \frac{1}{2}\ (KLD(p,q) + KLD(q,p))$$

.

Des solutions à ces problèmes d'optimisation sont connues en forme close et sont données dans [5].

### 3.3 Distance de Fisher-Rao

Puisque la géométrie des distributions gaussiennes est hyperbolique, on peut exprimer la distance de Fisher-Rao entre deux gaussiennes en utilisant la distance hyperbolique dans le demi-plan de Poincaré.

$$\mathrm{FRD}(f_p, f_q) = \sqrt{2}\ln\frac{|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_q}{\sqrt{2}},\sigma_q)|+|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|}{|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|-|(\frac{\mu_p}{\sqrt{2}},\sigma_p)-(\frac{\mu_p}{\sqrt{2}},\sigma_p)|}$$

### 3.4 Centroïdes modèles

Pour appliquer l'algorithme $k$-moyennes avec la distance de Fisher-Rao, il faut définir des centroïdes dans l'espace hyperbolique. Puisque les centroïdes de Fisher-Rao ne sont pas connus en formule close, nous allons utiliser les centroïdes modèles proposés par Galpérin [1], qui sont une façon de définir des centroïdes dans les catégories d'espaces à courbure constante (positive, négative et nulle). Pour un espace à courbure constance de dimension $d$, on commence par trouver un modèle de dimension $d+1$ plongé dans un espace euclidien. Pour un espace hyperbolique de dimension 2, on utilisera le modèle de Minkowski, la partie supérieure de l'hyperboloïde dont l'équation est $-x^2-y^2+z^2=1$.

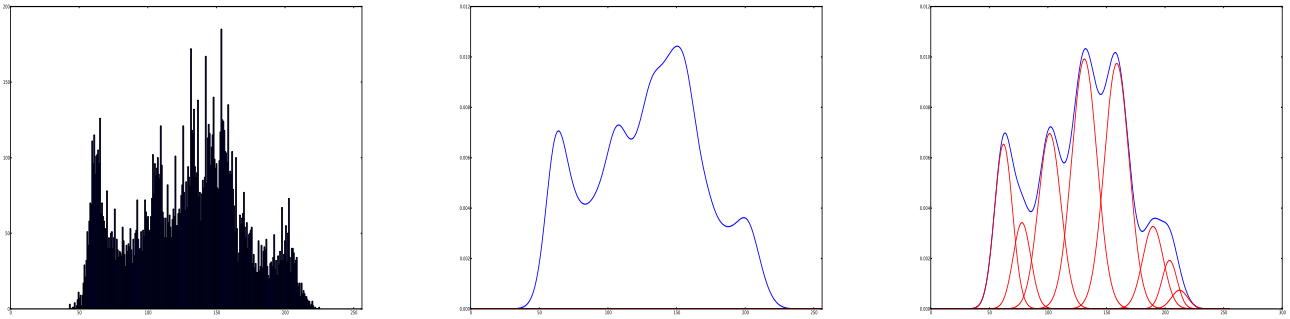

FIGURE 1 – Histogramme original, estimateur par noyau (14400 composantes) et mélange simplifié (8 composantes). Malgré le peu de composantes du mélange simplifié, l'approximation reste de très bonne qualité.

En premier lieu, chaque point $p$ (de coordonnées $(x_p, y_p)$) du disque de Klein est projeté vers un point $p'$ du modèle de Minkowski (voir par exemple [7] pour plus de détails sur ces formules) :

$$x_{p'} = \frac{x_p}{1 - x_p^2 + y_p^2} \qquad y_{p'} = \frac{y_p}{1 - x_p^2 + y_p^2}$$
$$z_{p'} = \frac{1}{1 - x_p^2 + y_p^2}$$

Ensuite, le centre de masses des points sur le modèle de Minkowski est calculé :

$$c'' = \sum \omega_i p'_i$$

Le point $c''$ n'est pas forcément sur le modèle de Minkowski, il faut donc le renormaliser en calculant l'intersection entre le vecteur $Oc''$ et l'hyperboloïde :

$$c' = c''/(-x_{c''}^2 - z_{c''}^2 + z_{c''}^2)$$

On effectue ensuite la transformation inverse, pour calculer la représentation du point $c'$ sur le disque de Klein :

$$x_c = \frac{x_{c'}}{z_{c'}} \qquad y_c = \frac{y_{c'}}{z_{c'}}$$

Cette méthode construit le centroïde de ponts situés sur le disque de Klein. Comme les paramètres de la distribution gaussienne sont sur le demi-plan supérieur de Poincaré, il faut convertir les coordonnées d'un modèle vers un autre, en utilisant le disque de Poincaré comme intermédiaire. Étant donné un point $(a, b)$ sur le demi-plan, on obtient le point en coordonnées complexe $z = a+ib$. La transformation de Möbius[6] vers le disque de Poincaré est la suivante :

$$z' = \frac{z - i}{z + i} \qquad z = \frac{i(z' + 1)}{1 - z'}$$

La transformation entre un point $k$ du disque de Poincaré et un point $k$ du disque de Klein s'effectue avec la bijection.

$$p = \frac{1 - \sqrt{1 - \langle k, k\rangle}}{\langle k, k\rangle} \qquad k = \frac{2}{1 + \langle p, p\rangle} p$$

La figure 2 décrit ces différentes étapes.

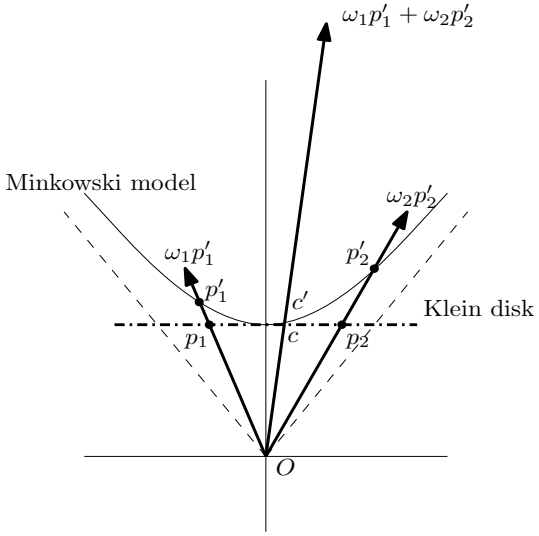


FIGURE 2 – Calcul du centroïde $c$ associé au système $(\omega_1, p_1), (\omega_2, p_2)$

La combinaison de ces différentes transformations permet de calculer le modèle centroïde et donc, d'obtenir, en formule close, le représentant d'un ensemble de gaussiennes en dimension 1. Ce qui, avec la distance de Fisher-Rao, nous permet d'appliquer efficacement la simplification basée sur $k$-moyennes.

# 4 Expériences

On compare ici la qualité (en terme de log-vraisemblance) des modèles obtenus par différentes méthodes : le classique EM, la simplification avec KL, la simplification avec les centroïdes modèles et à titre de comparaison, la simplification avec de vrais centroïdes de Fisher-Rao obtenus par recherche exhaustive. Le jeu de données utilisé à titre d'exemple est l'histogramme de niveau de gris de l'image Lena. Pour la divergence de Kullback-Leibler, nous présentons seulement les résultats obtenus avec le centroïde droit qui se comporte mieux que les deux autres ([2]) pour un coût identique.

La courbe de gauche de la figure 3 présente l'évolution de la

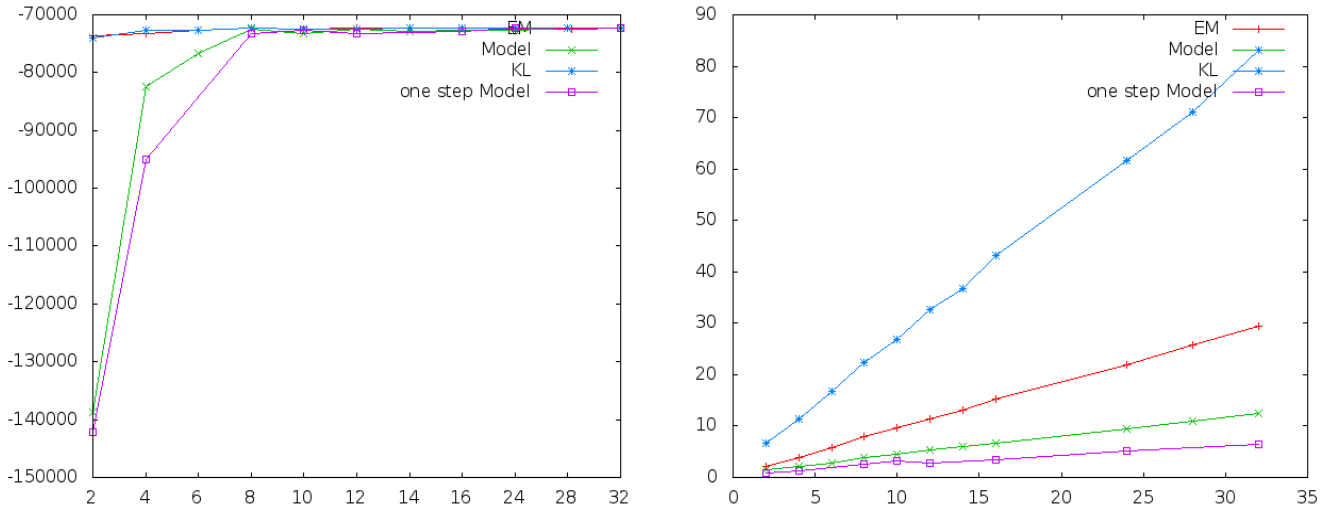


FIGURE 3 – log-vraisemblance des mélanges simplifiés et temps de calcul, en fonction du nombre de composantes.

log-vraisemblance en fonction du nombre $k$ de composants du mélange. On remarque immédiatement que toutes les méthodes ont des performances similaires et convergent rapidement vers une valeur maximale (la courbe pour la simplification avec KL est confondue avec celle obtenue par EM).

La divergence de Kullback-Leibler et la distance de Fisher-Rao se comportent de façon similaire mais sont assez différents d'un point de vue théorique : KL suppose une géométrie sous-jacente plate alors que Fisher-Rao est liée au caractère hyperbolique des distributions gaussiennes. Cependant, à une échelle infinitésimale, leur comportement est identique.

La partie droite de la figure 3 décrit le temps de calcul (en secondes) en fonction de $k$. Alors que les log-vraisemblances sont presque les mêmes d'un algorithme à l'autre, les coûts sont très différents. La simplification avec KL est la plus lente (en dépit de leur forme close, les formules sont complexes). La simplification avec les centroïdes modèles est très rapide, bien plus qu'EM, tout en donnant des modèles de même qualité.

Bien que plus lent à converger quand $k$ augmente, la simplification en une étape a de bonnes performances tout en étant bien plus rapide qu'une version avec toutes les itérations. Ici, l'initialisation est aléatoire, nous n'utilisons pas $k$-means++ car son coût à l'initialisation annule le bénéfice de ne faire qu'une seule itération.

## 5 Conclusion

Nous avons introduit ici une nouvelle méthode de construction de modèles de mélanges qui est à la fois rapide et précise. À partir d'un estimateur par noyaux, nous sommes capables de produire de nouveau modèles possédant la même qualité d'approximation mais plus rapides à construire que des modèles issus de l'Espérance-Maximisation. De plus, nous montrons expérimentalement que les centroïdes modèles sont de bonnes approximations des centroïdes de Fisher-Rao.

Les résultats présentés ici le sont en dimension 1, mais peuvent s'étendre en dimension arbitraire. De plus, si les résultats ici sont prometteurs, il reste à valider notre approche sur une application réelle.

Les codes sources et détails des expériences sont disponibles en ligne [1].

## Références

[1] G.A. Galperin. A concept of the mass center of a system of material points in the constant curvature spaces. *Communications in Mathematical Physics*, 154(1) :63–84, 1993.

[2] V. Garcia, F. Nielsen, and R. Nock. Levels of details for gaussian mixture models. *Computer Vision–ACCV 2009*, pages 514–525, 2010.

[3] J. Goldberger and S. Roweis. Hierarchical clustering of a mixture model. *Advances in Neural Information Processing Systems*, 17 :505–512, 2005.

[4] M.C. Jones, J.S. Marron, and S.J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433) :401–407, 1996.

[5] Frank Nielsen and Vincent Garcia. Statistical exponential families : A digest with flash cards. *CoRR*, abs/0911.4863, 2009.

[6] Frank Nielsen and Richard Nock. Hyperbolic voronoi diagrams made easy. *CoRR*, abs/0903.3287, 2009.

[7] J.G. Ratcliffe. *Foundations of hyperbolic manifolds*, volume 149. Springer Verlag, 2006.

[8] B.A. Turlach. Bandwidth selection in kernel density estimation : A review. *CORE and Institut de Statistique*, pages 23–493, 1993.

1. `http://www.lix.polytechnique.fr/~schwander/resources/gretsi2011`