Estimation de densité via un algorithme EM-Kernel

Catherine Aaron

Clermont Université, Université Blaise Pascal, Laboratoire de Mathématiques, BP 10448, F-63000 CLERMONT-FERRAND CNRS, UMR 6620, Laboratoire de Mathématiques, F-63177 AUBIERE

Résumé

On s'intéresse, dans ce papier à la construction d'un algorithme automatique d'estimation de densité. Cet algorithme repose sur un modèle de mélange. Chaque composante de ce mélange peut etre estimée via une méthode à densité dépendant d'un paramètre λ . La valeur de λ pour chaque composante du mélange, ainsi que les probabilités d'appartenance seront estimées par une méthode type EM par abus de langage (on maximisera une pseudo vraisemblance en alternant les étapes optimisation sur λ, optimisation sur les probabilités). Le choix d'une telle méthode sera justifié, pour la dimension 1, en partie 1. Des résultats asymptotiques seront explicités en partie 2. Enfin on présentera l'algorithme proprement dit et quelques résultats, en parties 3 et 4.

mots clés : Modèles semi et non paramétriques, Apprentissage

Abstract

This paper presents an algorithm for density estimation from samples in $\Re^d$ that leads on a mixture model ; each component of the mixture is estimated with a kernel method which depends on a parameter λ. The parameter λ and the probabilities for each individual of the sampel to belong to a component are computed via an EM-type algorithm (i.e. we alternete a pseudo-likelyhood maximization on λ and on the probabilities). In a first part we justify the method for the case $d = 1$. Part 2 is dedicated to asymptotics results. Part 3 and 4 are dedicated to the practical algorithm and numerical results.

Introduction : justification de la méthode

Soit $X_1, ..., X_N$ un $N-$échantillon de variables qui suivent une loi de densité inconnue f. Il est bien connu [6] qu'on peut estimer f, via un noyaux K par :

$$\hat{f}_{N,h}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right)$$

Tout le problème d'une telle estimation repose sur le choix d'une valeur correcte pour la "taille de fenêtre" h (voir [7] pour un review des principales méthodes associées à ce

choix).

Il est bien connu qu'il peut être largement profitable d'introduire une taille de fenêtre locale afin d'améliorer la qualité de l'estimation de la densité, le modèle deviens alors :

$$\hat{f}_{N,h}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h(X_i)}K\left(\frac{x-X_i}{h(X_i)}\right)$$

La bibliographie sur le choix de la fonction $h(X_i)$ est immense et nous nous concentrerons ici sur le cas le plus connu, celui proposé par Abramson [1] qui propose de choisir une taille de fenêtre proportionnelle à la racine de la densité : $h(X_i) = \lambda f(X_i)^{1/2}$ (bien sur dans la pratique on considérera la racine de la densité estimée).

Si une telle approche va, sous les hypothèses idoines, améliorer l'estimation de la densité par rapport à une estimation à noyau global, elle pourra, elle même, être améliorée par le choix d'un λ local.

Pour voir cela considérons $\lambda^{opt}(\overrightarrow{X})$ le facteur multiplicatif optimal (au sens du MISE -Mean Integrated Squared Error-) pour un $N-$échantillon $X_1, ..., X_N$ de loi de densité inconnue $f(x)$ et $\lambda^{opt}(\overrightarrow{Y})$ le facteur multiplicatif optimal pour un $N-$échantillon $Y_1, ..., Y_N$ de loi de densité $f(\frac{x}{\mu})\frac{1}{\mu}$

On peut facilement montrer qu'on a la relation :

$$\lambda^{opt}(\overrightarrow{Y}) = \sqrt{\mu}\lambda^{opt}(\overrightarrow{X})$$

Ceci met en évidence le fait que, si l'on souhaite estimer la densité d'un mélange de deux lois identiques à un facteur d'échelle prêt on aurait intérêt à rechercher un coefficient λ qui dépend de la composante du mélange sur laquelle on se trouve (voir figure 1).

Pour tenir compte de cet effet on va proposer un de nouveau modèle d'estimation de densité de type $EM - Kernel$:

$$\hat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\frac{p_j(X_i)}{\lambda_j h(X_i)}K\left(\frac{x-X_i}{\lambda_j h(X_i)}\right)$$

avec :

- $p_j(X_i)$ probabilité pour l'individu i d'appartenir à la composante j du modèle ces probabilités vérifieront donc évidemment: $p_j(X_i) \in [0,1]$ et $\forall i, \sum_j p_j(X_i) = 1$
- λ_j coefficient multiplicatif pour la composante j
- h fonction noyau normée par $\int h^2 = 1$ soit une fonction constante sur son support dans le cas d'un noyau global, la racine de la densité dans le cas d'une estimation type abramson ou tout autre type de fonction pertinente (i.e. qui améliore le noyau global)

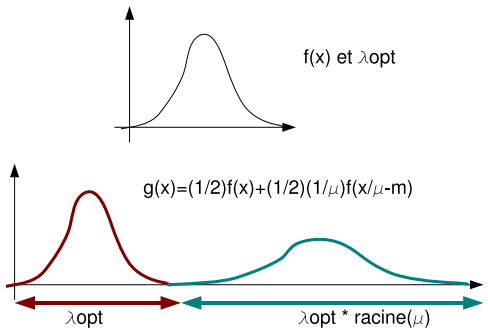

Figure 1: Lorsqu'on veut estimer la densité sur un mélange equiprobable de deux lois identiques a un facteur d'échelle μ prêt la valeur des coefficient multiplicatifs pour chacune des composante est différente il peut ainsi être interressant de considérer un coefficient "local"

L'introduction de modèles type $EM - Kernel$ a déjà été exposée par ([2] et [3]) mais l'intérêt en ce limitant à des mélanges de densités estimées avec des noyaux globaux (alors qu'on montre ici que cette approche garde de l'intérêt avec des noyaux locaux) et les algorithmes de calculs diffèrent du notre.

Quelques résultats asymptotiques

Pour l'instant seuls les calculs dans le cas d'une estimation par un mélange de densités dont chacune des composantes est estimée par une méthode à noyaux globale, i.e. si on se place en dimension 1, dans le cas où l'on estime la densité par une fonction du type :

$$\hat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{K} \frac{p_j(X_i)}{h_j} K\left(\frac{x - X_i}{h_j}\right)$$

ont été effectués. On peut démontrer que :

- Si les fonctions fp_j sont $\mathcal{C}^2$ alors :
- $MISE = CN^{-4/5}$

Ce qui signifie qu'une fois les fonctions poids optimales obtenues, la vitesse de convergence notre estimateur de densité sera la même que celle à noyaux gobal mais on abaissera la constante. Si ce résultat ne semble pas très puissant, il permet néanmoins d'améliorer, parfois grandement, la qualité de l'estimation de densité. On peut noter que des méthodes à noyaux locaux peuvent avoir des vitesses de convergence moindre que celles à noyaux globaux mais améliorer les résultats pour des échantillons de taille "raisonnable" (celles

dont on dispose en pratique) [8]

Le résultat exposé en dimension 1 se généralise relativement aisément aux autres dimensions si l'on continue à considérer que chaque composante est estimée par une méthode à noyaux global. En revanche les calculs pour des estimations à noyaux locaux restent à effectuer (on s'attend néanmoins au même type de résultat).

Algorithme

Les quelque résultats que nous allons présenter dans la section suivante ont été réalisés en utilisant l'algorithme suivant :

- On choisit la méthode d'estimation de densité pour chacune des composantes du mélange (en accord avec les remarques sités dans [1] nous choisiront une estimation par des noyaux globaux en dimension 1 et un noyau type Abramson en dimension supérieure)
- On choisit K le nombre de composante du mélange
- **Initialisation :** On initialise les $p_j(X_i)$ en réalisant un algorithme EM classique de mélange de gaussienne sur les données
- **Itération de :**
  - Actualisation des λ_i à poids fixés par maximisation de la pseudo-vraisemblance ([4] [5])
  - Actualisation des $p_j(X_i)$ à λ_i fixés par maximisation de la vraisemblance

Quelques résultats

On a appliqué l'algorithme présenté à l'estimation de densité pour l'hidalgo stamps data, ces données correspondent à des épaisseurs de papiers de timbre et le problème fréquement posé sur cette base est l'estimation du nombre de modes, à chaque modes étant associé un type de papier.

L'algorithme a aussi été appliqué, en dimension 1 aux données "Old Faithful" (temps entre deux irruption d'un geyser aux Etats-Unis), On retrouve le résultat connu d'un mélange de deux densités de variance différentes. On retrouve des résultats assez similaires a ceux de [8] qui eux aussi mettent en évidence la présence d'un coude dans la première composante du mélange (coude non obtenu par d'autres méthodes d'estimation de densité)

Enfin, pour avoir une idée du comportement de l'algorithme en dimension supérieure, nous avons appliqué l'algorithme au cas des iris de Fisher (les graphiques correspondront aux résultats projetés sur les deux premiers axes d'une ACP) pour un mélange de 3 densités.

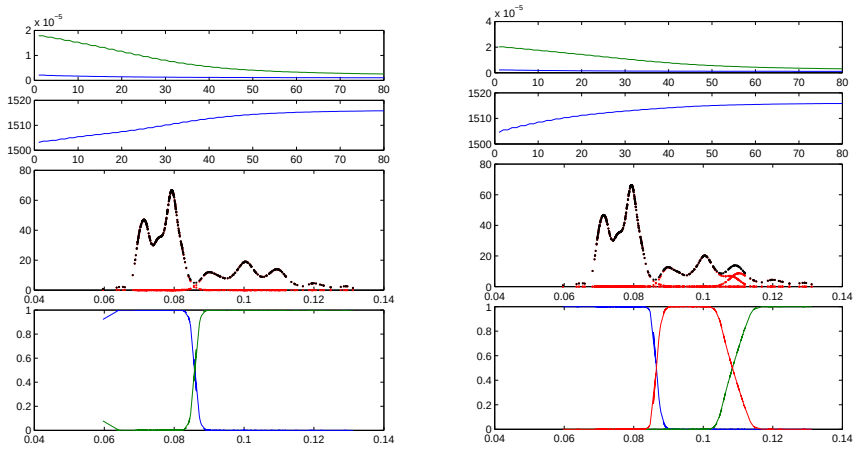

Figure 2: Résultats pour l'estimation de densité pour l'hidalgo stamps data : la première figure correspond à une estimation sur un mélange de deux densités, la seconde pour un mélange a trois densités, pour chaque figure on observe : l'évolution des tailles de fenêtres, l'évolution de la pseudo-vraisemblance, la densité estimée (noir) et chaque composante (rouge), et, pour finir les poids

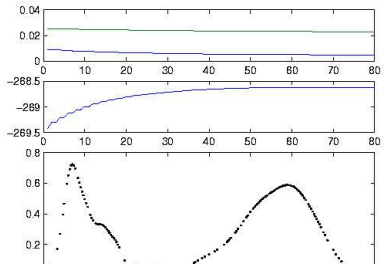

Figure 3: Résultats pour l'estimation de densité pour Old Faithful data

Conclusion - Perspectives

Les résultats présentés sont relativement bons mais ont été obtenus éssentiellement en "trichant" c'est-à-dire en connnaissant a priori le bon nombre de composantes du modèle. Il apparait néanmoins que dans le cadre de l'estimation de densité seule le résultat semble relativement peu sensible au nombre de composantes (par exempe dans le cas de la base "hidalgo stamps" où le nombre de composantes est inconnu la densité varie très peu qu'on considère un mélange de 2 ou 3 densités). En revanche ce choix s'avère primordial dans le cadre d'une utilisation en classification (type iris) et il est alors indispensable de trouver un moyen de choisir un bon nombre de composantes (a priori ou a posteriori).

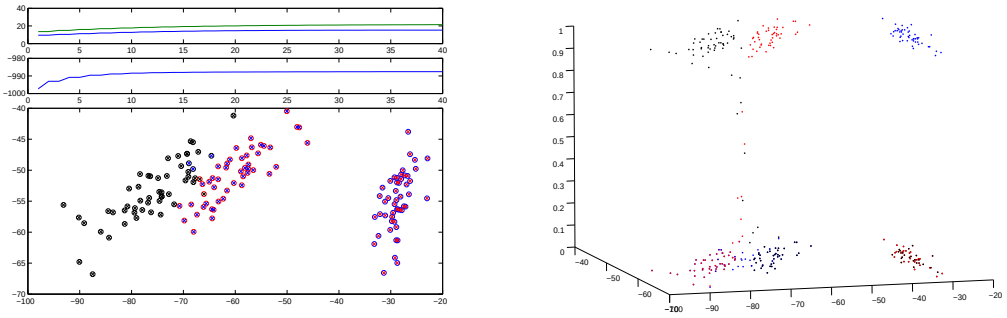

Figure 4: Résultats pour les iris de Fisher, utilisé en classification non supervisée on retrouve, avec un faible taux de mal classés les 3 “vraies classes” (le graphique en bas à gauche met en relation les classes obtenues -cercles- et les vraies classes -croix-) de plus les individus mal classés sont essenctiellement ceux pour lesquels la probabilité d'appartenance aux classes est la moins discriminante (graphique de droite illustrant la probabilité d'appartenance aux classes)

Bibliographie

[1] Abramson S. (1982) *The Annals of Statistics*, Vol. 10, No. 4, 1217-12223

[2] Bhattia V. Mulgrew B. (2004), A EM-kernel density method for channel estimation in non-gaussian noise *Proceedings IEEE VTS Fall 60th Vehicular Technology Conference*

[3] Bugeau A., Perez P. (2007) Estimation à noyau adaptatif dans des espaces multidimensionnels hétérogènes *Technical report, IRISA*

[4] Duin R.P. (1976). On the choice of smoothing parameter of parzen estimator of probability density functions, *EEE transaction on Computers*, I 25, 1175-1179.

[5] Hermans J. Van den Broek K. Habbema J.D.F. (1974). A stepwise discrimination program using density estimation, *COMPSTAT'74 Proceedings in Computational Statistics.*

[6] Parzen (1962). On estimation of a probability density function and mode, *Ann. Math. Stat*, 33, 10651076.

[7] Simon J. Sheather (2004) Density Estimation *Statistical Science*, Vol. 19, No. 4, 588-597.

[8] Sain, S.R., Scott, D.W., (1996). On locally adaptive density estimation *Journal of the American Statistics Association* ,91, 1525-1534.