Sur l'estimation des paramètres de mixture de gaussiennes généralisées

Mohamed Ould Mohamed Mahmoud, Mériem Jaïdan

Sur l'estimation des paramètres de mixture de gaussiennes généralisées

Mohamed Ould Mohamed Mahmoud, Mériem Jaïdane

Ecole Nationale d'Ingénieurs de Tunis, Unité de recherche Signaux et Systèmes (U2S), Tunis, BP 37, Le Belvédère, 1002, Tunisie.

Abstract

The parameters estimation of mixture distributions is an important task in statistical signal processing, Pattern recognition, blind equalization. And other modern statistical tasks often call for mixture estimation. This paper aims to provide a realistic distribution based on Mixture of Generalized Gaussian distribution (MGG), which has the advantage to characterize the variability of shape parameter in each component in the mixture. We propose a formulation of the Expectation Maximization (EM) under Generalized Gaussian distribution. For this two different methods are presented. In the first method a derivation of the Maximum-Likelihood estimation of the parameters is used to update the mixture parameters. In the second approach we propose an extension of the (EM) algorithm derived in the case of "standard" mixture Gaussian to include the shape parameter estimation. The Kullback-Leibler divergence (KLD)[5] is used to compare, and evaluate these algorithms of MGG parameters estimation. An application of this technique is considered for modeling load distribution which exhibits an heterogeneity with a high variability of shape parameters

Résumé

Dans cette étude on propose une méthode à faible complexité pour estimer le paramètre de forme dans le cas d'un mélange des lois gaussiennes généralisées qui grâce à ces propriétés prend de plus en plus d'importance dans des domaines très variés (classification audio, modélisation des temps d'inter-arrivée en transmission IP, modélisation des puissances de pointes journalières lectriques,...). Une reformulation de l'algorithme standard EM est effectuée pour inclure le paramètre de forme. La méthode proposée est basée sur l'utilisation de la relation établie entre ce paramètre et le kurtosis. Une deuxième méthode basée sur généralisation d'une méthode existante est aussi proposée. Les performances de ces approches sont comparées et vérifiées par mesure de divergence de Kullback Leibler sur des données simulées et des données réelles de courbes de charge[1]

Mots-clés : Modèle de mixture, gaussienne généralisée estimation des paramètres, algorithme EM.

[1]Etude menée en marge d'un projet VRR - Projet de Valorisation des Résultats de la Recherche (2004-2007) financé par le MESRST (Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de la Technologie) en Tunisie. Projet entrant dans le cadre des actions incitatives R&D Université/Industrie. Projet réalis entre la Direction d'Etudes et de Planification de la STEG et et l'Unité Signaux et Systèmes du sur la prévision à moyen terme de la pointe avec prise en compte des effets de climatisation.

1 Modèle de mixture de gaussiennes généralisées (MGG)

La modélisation d'une densité de probabilité par un mélange de gaussiennes généralisées consiste à décomposer cette ddp en une somme pondérée de K composantes en suposant que chaque composante est modélisée par une gaussienne généralisée. grâce á ces propriétès cette distribution de mélange prend de plus en plus d'importance dans des domaines trés variés [1, 2, 4]). Dans ce modle la densité de probabilité en un point x est ainsi donnée par :

$$p(x \mid \Theta) = \sum_{i=1}^{K} \omega_i p_i(x \mid m_i, \sigma_i, c_i) \tag{1}$$

où - K est le nombre de composantes dans le mélange.
- ω_i est la proportion de mélange. Ces proportions ω_i représentent les probabilités a priori des différentes classes.
- $p_i(x \mid m_i, \sigma_i, c_i)$ repésente la densité de probabilité d'une loi gaussienne généralisée représentant la i-ième composante [1, 2] :

$$p_i(x \mid m_i, \sigma_i, c_i) = \frac{c_i \gamma_i}{\Gamma(1/c_i)} e^{-\gamma_i^{c_i} [|(x-m_i)|]^{c_i}} \tag{2}$$

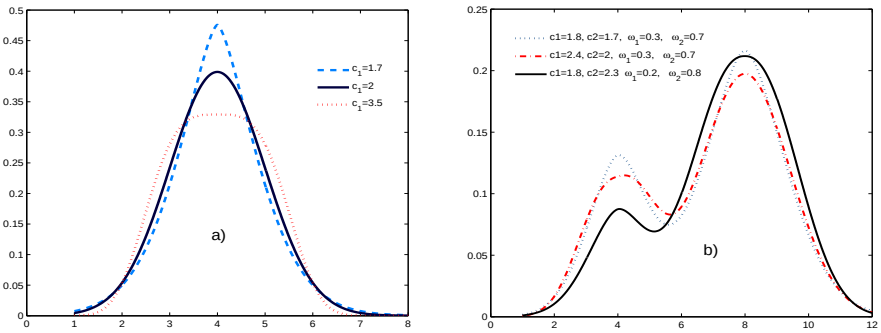

Figure 1: *La figure a) montre l'exemple de la densité de probabilité d'une loi de gausienne généralisée symétrique ($K = 1$) pour différentes valeurs de paramètre de forme de même valeurs $\sigma_1 = 1$, $m_1 = 4$. La figure b) représente un exemple de mixture MGG $K = 2$, avec $m_1 = 4$, $m_2 = 8$ et différentes valeurs de c_i et ω_i indiquées sur la figure.*

- γ_i est le paramètre d'échelle : $\gamma_i = \frac{1}{\sigma_i} \frac{\Gamma(3/c_i)}{\Gamma(1/c_i)}^{1/2}$, $\Gamma(x) = \int_0^\infty \tau^{x-1} e^{-\tau} d\tau$ étant la fonction Gamma
-$\Theta = (\omega_i, m_i, \sigma_i, c_i)$, $i = 1, 2 \cdots K$. m_i est la valeur moyenne, σ_i l'écart type, c_i le paramètre de forme.
Le paramètre de forme c_i caractérise l'applatissement de la ddp et contrôle la déviation de la noramlité (cf. figure 1). Par variation de ce paramètre il est possible de caractériser une large classe de distribution : gaussienne (c_i=2), sous-gaussienne, (c_i<2 : distribution plus pointue, que la gaussienne , faibles queues) et super-gaussienne (c_i>2 : distribution aplatie, plus uniform).

2 Reformulation de l'algorithme EM dans le cas de mélange MGG pour estimer les paramétres de forme c_i

L'estimation des paramètres de la mixture de gaussiennes généralisées est plus complexe que dans le cas de mixture de gaussiennes. La difficulté réside dans l'évaluation des paramètres de forme $c_i, i = 1...K$. Nous proposons une extension de l'algorithme EM (Expectation Maximisation) [2] dans l'estimation des paramètres du modèle MGG qui consiste à maximiser le log de la vraisemblance complète[2] donnée, dans le cas de gaussiennes généralisées, par :

$$L(X \mid \theta) = \sum_{i=1}^{K} \sum_{j=1}^{N} h_{i,j} \; ln \left(\omega_i p_i(x_j \mid m_i, \sigma_i, c_i)\right) \tag{3}$$

Oú $h_{i,j} = p(i \mid x_j)$ $(i = 1, \cdots, K$ et $j \in [0, N])$ représente l'espérance conditionnelle de p_i sachant l'observation x_j, càd la probabilité a posteriori pour que x_j se trouve dans la ième composante. Si on remplace dans l'équation 3 $p_i(x)$ par son expression (équation 2) on obtient l'expression suivante de $L(X \mid \theta)$:

$$L(X \mid \theta) = \sum_{i=1}^{K} \sum_{j=1}^{N} h_{i,j} ln(\omega_i) + \sum_{i=1}^{K} \sum_{j=1}^{N} h_{i,j} \left(lnc_i - ln2 - ln\gamma_i - ln\,\Gamma\left(\frac{1}{c_i}\right) - \gamma_i^{(c_i)} |x_j - m_i| \right) \tag{4}$$

Nous proposons d'optimiser $L(X \mid \theta)$ itérativement avec le même algorithme EM, en introduisant cette fois-ci l'estimateur relatif au paramètre de forme c_i. Ainsi, les étapes de l'algorithme de EM dans ce cas peuvent être résumé comme suite:

E-step (Expectation step) l'étape Espérance est représentée par le calcul de la probabilité conditionnelle $h_{i,j}$

$$h_{i,j}^{(n+1)} = \frac{\omega_i p(x_j | m_i^{(n)}, \sigma_i^{(n)}, c_i^{(n)})}{\sum_{r=1}^{K} \omega_r^{(n)} p(x_j | m_r^{(n)}, \sigma_r^{(n)}, c_r^{(n)})} \tag{5}$$

dans cet étape le calcul de la fonction $L(X \mid \theta)$ basée sur l'estimation de $\theta^{(n)}$ est effectuée.

M-Step (Maximization step) permet la maximisation numérique de la fonction de log vraisemblance. Dans le cas de mélange de simple gaussien les paramètres $(\omega_i, m_i, \sigma_i)$ sont estimés avec un ensemble des équations itératives [4] :

$$\hat{\omega_i}^{(n+1)} = \frac{1}{N} \sum_{j=1}^{N} h_{i,j}^{(n)} \quad \hat{m_i}^{(n+1)} = \frac{\sum_{j=1}^{N} h_{i,j}^{(n)} x_j}{\sum_{j=1}^{N} h_{i,j}^{(n)}} \quad \hat{\sigma_i^2}^{(n+1)} = \frac{\sum_{j=1}^{N} h_{i,j}^{(n)} (x_j - \hat{m}_i^{(n)})^2}{\sum_{j=1}^{N} h_{i,j}^{(n)}} \tag{6}$$

[2]Les observations x_j sont considérées comme des données incomplètes, auxquelles on rajoute les données manquantes représentées par le vecteur aléatoire d'appartenance d'un point x_j à la classe i

Cependant, l'algorithme tel qu'il est ne permet pas d'inclure les paramètres de forme. Ainsi, un développement supplémentaire doit être effectué pour trouver l'équation itérative correspondante aux paramètrex de forme. Dans les section suivantes nous proposons deux méthodes différentes pour résoudre ce problème.

2.1 Estimation des paramètres de forme par optimisation numérique (OPN)

Nous proposons dans cette première méthode une généralisation de l'approche proposée par Yakoub [2] dans le cas d'un mélange de deux gaussiennes génralisées. Cette méthode consiste à maximiser $L(X \mid \theta)$ en annulant les dérivées de la fonction de log de vraisemblance (équation 4) par rapport m_i,γ_i, c_i respectivement.

$$\frac{dL(X \mid \theta)}{dm_i} = 0 \, , \, \frac{dL(X \mid \theta)}{dc_i} = 0 \, , \frac{dL(X \mid \theta)}{d\gamma_i} = 0 \tag{7}$$

Dans cette méthode, le calcul de c_i revient á résoudre l'équation non linéaire (les équations relatives à l'estimation de la moyenne et paramètre d'échelle m_i et γ_i peuvent être obtenues dans [2]):

$$\frac{dL(X \mid \theta)}{dc_i} = \sum_{j=1}^{N-1} h_{i,j} \left[\frac{1}{c_i} + \frac{1}{c_i}^2 \Psi\left(\frac{1}{c_i}\right) - \left(\frac{|x_j - m_i^{(n)}|}{\gamma_i^{(n)}} \right)^{c_i} ln\left(\frac{|x_j - m_i^{(n)}|}{\gamma_i^{(n)}} \right) \right] = 0 \tag{8}$$

où $\Psi(\cdot)$ est la fonction digamma $\Psi(x) = \Gamma'(x)/\Gamma(x)$.

Ces équations peuvent être résolues numériquement, comme suggéré dans [2], par la procédure itérative de Newton-Raphson. C'est ce que nous avons mis en oeuvre pour le calcul de $c_i^{(n+1)}$ par cette méthode. Cette approche d'estimation est très complexe, et caractérisée par un temps de calcul très important. Elle révèle également une importante sensibilité aux conditions initiales.

2.2 Estimation des paramètres de forme par utilisation de statistiques d'ordre supérieur (SOS)

Nous proposons, dans cette nouvelle méthode, d'intégrer un estimateur de c proposé par Regazoni [3] dans le cas d'une gaussienne généralisée à partir du lien entre le kurtosis κ et le paramètre de forme c selon l'expression suivante.

$$\widehat{c} \approx \sqrt{\frac{5}{\widehat{\kappa} - 1.865}} - 0.12 \tag{9}$$

Ainsi, nous estimons en premier temps le Kurtosis κ_i à l'itération $(n+1)$ de la même façon (avec mêmes poids) que m_i et σ_i dans le cas de mélange de simples gaussiennes (équation 6):

$$\hat{\kappa_i}^{(n+1)} = \frac{\sum_{j=1}^{N} h_{i,j}^{(n)} (x_j - m_i^{(n)})^4}{(\sigma_i^{(n)})^4 \sum_{j=1}^{N} h_{i,j}^{(n)}} \tag{10}$$

puis nous utlisons l'estimateur précédent pour déduire $c_i^{(n+1)}$ sachant l'estimation de $\kappa_i^{(n+1)}$ selon :

$$\hat{c_i}^{(n+1)} \approx \sqrt{\frac{5}{\hat{\kappa_i}^{(n+1)} - 1.865}} - 0.12 \tag{11}$$

Cette méthode simple est particulièrement appropriée pour les grandes valeurs de κ_i ($\kappa_i > 1.865$).

2.3 Premiers résultats

Les méthodes d'estimation proposées ont d'abord été testées et validées sur des exemples théoriques. On présente ici l'exemple d'un mélange de 3 gaussiennes qui doit avoir des paramètres de forme $c_i = 2$. On dénote par M, C, σ respectivement les vecteurs de m_i, c_i, et σ_i, $i = 1...K$. Nous avons généré un vecteur de nombres aléatoires issu d'un mélange de 3 gaussiennes avec les moyennes $M = [-4, 0, 5]$ et les variances $\sigma = [1, 2, 1.5]$. Les poids du mélange sont fixés égales à $1/3$. L'estimation est effectuée avec les conditions initiales suivantes $M_0 = [-1, 5, 1]$ $\sigma_0 = [1, 1, 2]$, $C_0 = [3, 3, 3]$.

Les résultats de l'estimation des paramètres sont conformes avec l'exemple généré. Ainsi, avec la deuxième méthode **SOS**, les paramètres de formes obtenus sont $c_1 = 2.08$, $c_2 = 2.21$, $c_3 = 2.29$ proches de 2, valeur critique de c dans le cas gaussien, les valeurs des moyens et variance estimées $\widehat{M} = [-3.9, 0.1, 4.9]$ $\widehat{\sigma} = [1, 1.8, 1.4]$, sont également proches des valeurs réelles.

Pour la comparaison entre les deux méthodes d'estimation nous avons mesuré la divergence de Kullback Leibller. Les résultats obtenus, dans l'exemple de la figure 2, sont $KLD = 9.35.10^{-4}$ pour la méthode SOS, $KLD = 1.2.10^{-3}$ pour la méthode numérique OPN.

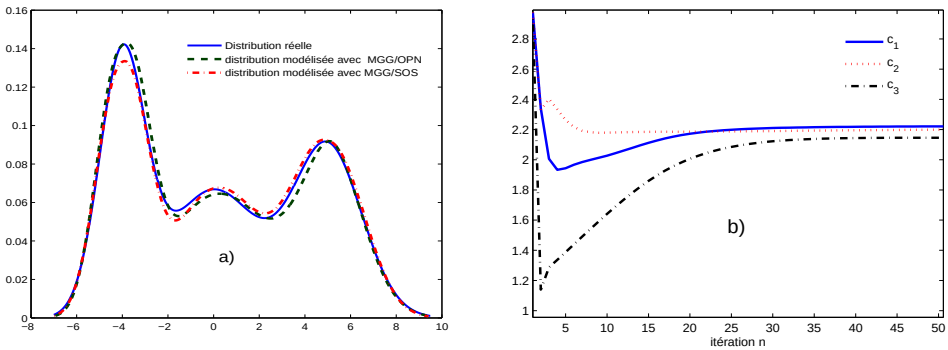

Figure 2: *Figure a) montre la densités de probabilité réelle (trait continu) et estimées selon les deux approches proposées d'un mélange de 3 gaussiennes, pour N=5000. Le résultat d'estimation des paramètres de forme avec la méthode SOS sont* $c_1 = 2.08$, $c_2 = 2.21$, $c_3 = 2.29$. Figure b) montre la convergence des paramètres obtenue avec la méthode SOS.

Application dans le cas de la modélisation de la distribution de courbes de

charge électrique Nous présentons ici les premiers résultats d'application du modèle MGG pour cerner la distribution annuelle des courbes de charge électrique réellede la Société Tunisienne d'Électricité et du gaz (STEG) (cf. figure 3) Ces premiers résultats

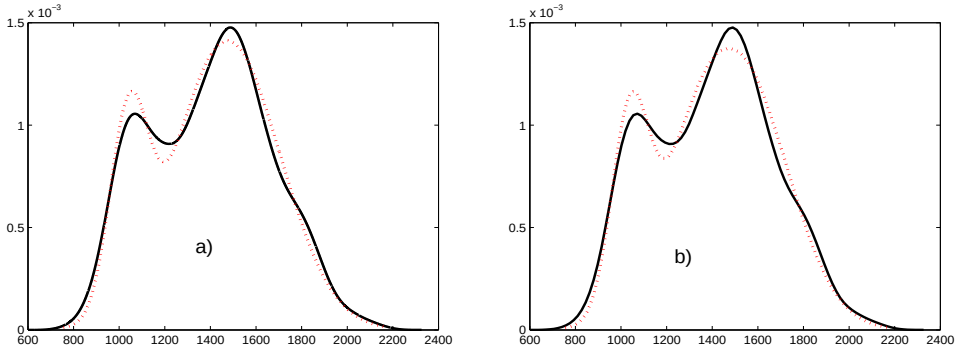

Figure 3: *Figure a) densités de probabilité réelle (trait continu) et estimée (trait pointillé) avec les deux méthode d'estimation par optimisation numérique (figure a) et SOS (figure b) sur l'exemple de l' année 2005. les paramètres de forme estimés avec la méthode SOS sont* $c_1 = 1.92$*,* $c_2 = 2.14$.

montrent que la deuxième méthode proposée a des performances proches voir meilleures que la première méthode malgré sa simplicité. une bonne adéquation du modèle MGG dans le cas des courbes de charge.

References

[1] C. Tzagkarakis, A. Mouchtaris, P. Tsakalides; *"Musical Genre Classification via Generalized Gaussian and Alpha-Stable Modeling,"* Acoustics, Speech and Signal Processing, ICASSP 2006.

[2] Y. Bazia, L.Bruzzone, and F. Melgania *Image thresholding based on the EM algorithm and the generalized Gaussian distribution*, Pattern Recognition vol.40 pp. 619–634 (2007).

[3] A. Tesei, and C.S. Regazzoni,*HOS-based generalized noise pdfs models for signal detection optimisation* IEEE Signal processing, vol. 65, No.2 pp. 267–281, March. 1998.

[4] Mohand Saïd Allili, Nizar Bouguila. *Finite Generalized Gaussian Mixture Modeling and Applications to Image and Video Foreground Segmentation.* Fourth Canadian Conference on Computer and Robot Vision(CRV'07) IEEE.

[5] David W. Scott, *Mulivariate Density Estimation*, Wiley-Interscience. 1992.