Expression explicite de l'évidence pour l'estimation du nombre de composantes d'un mélange de gaussiennes : cas particulier d'un petit nombre d'observations

Jessica SODJO, Audrey GIREMUS, Jean-François GIOVANNELLI

IMS (Univ. Bordeaux, CNRS, Bordeaux-INP), F-33400 Talence, France
jessica.sodjo@ims-bordeaux.fr,audrey.giremus@ims-bordeaux.fr,giova@ims-bordeaux.fr

Résumé – Dans cette communication, nous présentons un calcul explicite de l'évidence pour sélectionner le nombre de composantes d'un mélange de lois normales. Plus précisément, nous nous plaçons dans une approche bayésienne puis écrivons la loi jointe des observations et des paramètres du mélange. Nous marginalisons ensuite cette quantité suivant les paramètres de mélange pour exprimer l'évidence. Le choix de lois *a priori* conjuguées permet de trouver l'expression explicite. Nous comparons notre approche au « Bayesian Information Criterion » (BIC) et à la moyenne harmonique et montrons son intérêt plus spécifiquement quand le nombre d'observations est réduit.

Abstract – In this paper, we present a direct approach to select the number of components of a Gaussian mixture by computing the evidence. The joint distribution of the observations and the mixture parameters is written considering a Bayesian framework. Then, this quantity is marginalized over the mixture parameters in order to express the evidence. The choice of prior conjugate distributions makes it possible to compute the closed-form. Our approach is hence compared to the Bayesian Information Criterion (BIC) and the harmonic mean and its interest is shown more specifically for small data sets.

1 Introduction

Les mélanges de lois normales sont principalement utilisés en classification comme dans [1] où ils sont appliqués dans le diagnostic de maladies par l'analyse de concentrations de protéines dans le sang. Ils peuvent également être mis à profit pour modéliser des bruits non gaussiens. Ils se présentent sous la forme suivante. Considérons un ensemble de N observations indépendantes $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ où $\boldsymbol{x}_n \in \mathbb{R}^M$, $n = 1, \ldots, N$. $\boldsymbol{X}$ est de plus supposé séparable en K sous-ensembles appelés classes au sein desquels les observations sont identiquement distribuées. La distribution d'une observation est donc [2, chapitre 9] :

$$p(\boldsymbol{x}_n \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{1}$$

avec $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ où $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$, $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ et $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_K\}$. π_k est la probabilité qu'une observation appartienne au k-ème sous-ensemble. Les coefficients de mélange π_k sont donc positifs et se somment à 1. $\mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ est une composante du mélange et est la loi normale de moyenne $\boldsymbol{\mu}_k \in \mathbb{R}^M$ et de matrice de précision $\boldsymbol{\Lambda}_k \in \mathbb{R}^{M \times M}$.

Cependant, l'utilisation de ces modèles requiert de résoudre deux problèmes majeurs : l'estimation des paramètres des différentes lois gaussiennes et l'évaluation du nombre de composantes dans le mélange.

Nous formulons l'estimation du nombre de composantes comme un problème de sélection de modèle. La sélection de l'ordre du modèle est liée au calcul de l'évidence. Cette dernière s'obtient en marginalisant la distribution jointe des observations et des paramètres du mélange par rapport à ces derniers. Grâce à un choix judicieux de lois *a priori* sur les paramètres, nous proposons une formule explicite de l'évidence qui est appliquée pour un nombre réduit d'individus en raison du coût calculatoire induit. Son utilisation dans ce contexte apporte une plus-value. En effet, la plupart des méthodes existantes présentent des limitations quand N est petit. Une première catégorie d'approches consiste à fixer un nombre maximal de classes et à réaliser l'estimation de leurs paramètres en utilisant par exemple l'algorithme « Expectation-Maximization » (EM) [3], les méthodes bayésiennes variationnelles ou le Ying-Yang [4]. Les classes de faible probabilité sont ensuite supprimées. Cependant l'imprécision sur l'estimation des paramètres impacte fortement la sélection de l'ordre du modèle. Une alternative est d'échantillonner la loi a posteriori jointe du nombre de composantes du mélange et de leurs paramètres en utilisant par exemple l'algorithme « Reversible Jump Markov Chain Monte Carlo ». L'ordre est ensuite estimé comme celui maximisant la distribution a posteriori estimée par échantillonnage. Une troisième catégorie de méthodes consiste à approcher l'évidence, soit analytiquement comme par exemple le « Bayesian Information Criterion » (BIC) [5], soit en utilisant des méthodes numériques d'intégration comme l'estimateur de la moyenne harmonique. Néanmoins, d'une part le BIC n'est valide qu'asymptotiquement. D'autre part, en ce qui concerne les méthodes de

simulation, la loi a posteriori peut être diffuse en présence d'un faible nombre d'observations et son exploration requiert un très grand nombre d'échantillons.

Cette communication est organisée de la façon suivante. Nous présentons la méthode bayésienne utilisée dans la partie 2. Dans la partie 3 est développé le calcul de l'évidence. L'algorithme de la moyenne harmonique et le BIC, auxquels sera comparée notre approche, sont introduits dans la partie 4. Des résultats de simulation sont présentés dans la partie 5 qui mettent en évidence l'intérêt de notre méthode pour un faible nombre d'observations.

2 Méthode bayésienne pour la sélection

L'estimation du nombre de composantes K du mélange est présentée comme un problème de sélection de modèle. L'estimation optimale au sens de la minimisation du risque bayésien associé au coût $0/1$ maximise la probabilité *a posteriori* :

$$\widehat{K} = \underset{m}{\operatorname{argmax}}\, Pr(K = m \mid \boldsymbol{X})$$

En utilisant le théorème de Bayes on a :

$$Pr(K = m \mid \boldsymbol{X}) \propto p(\boldsymbol{X} \mid K = m) Pr(K = m)$$

où $m \in [1, \ldots, K_{max}]$. La loi de probabilité suivie par le modèle dépend des applications. Une loi uniforme est ici choisie pour n'en favoriser aucun. Ainsi, maximiser $Pr(K = m \mid \boldsymbol{X})$ revient à maximiser l'évidence $p(\boldsymbol{X} \mid K = m)$. Cette évidence pour un modèle $K = m$ donné peut être calculée avec :

$$p(\boldsymbol{X} \mid K = m) = \int p(\boldsymbol{X}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \mid K = m) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\Lambda} \quad (2)$$

Dans la suite, $K = m$ sera omis par souci de concision. L'objectif de la plupart des algorithmes existants est d'approcher (2). Une contribution de ce travail est une forme explicite de l'évidence. Le calcul est rendu possible par un choix approprié de lois *a priori* pour les paramètres inconnus.

3 Calcul de l'évidence

Dans cette partie sera présenté le calcul de l'évidence. Il n'est pas possible d'obtenir une formule explicite avec la définition du modèle donnée en (1). Une solution classique pour s'affranchir de la somme sur les composantes dans (1) est d'introduire des variables latentes formalisant l'appartenance des individus à l'une des classes ; l'EM pour les mélanges de lois normales repose sur cette idée. Pour ce faire, un vecteur $\boldsymbol{z}_n$ associé à chaque observation $\boldsymbol{x}_n$ est introduit. L'ensemble des vecteurs latents discrets est : $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$. Ces vecteurs prennent des valeurs dans $\{\mathbb{1}_k\}_{k=1,\ldots,K}$ où $\mathbb{1}_k$ est un vecteur binaire dont le seul élément non nul est le k-ème. Conformément à la définition des coefficients de mélange, la distribution de $\boldsymbol{Z}$ est donc [2, chapitre 9] :

$$P(\boldsymbol{Z} \mid \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \quad (3)$$

où z_{nk} est le k-ème élément du vecteur $\boldsymbol{z}_n$.

Ainsi, la distribution des observations conditionnellement à l'ensemble des paramètres est :

$$p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)^{z_{nk}} \quad (4)$$

La formule de l'évidence pour un modèle $K = m$ devient :

$$p(\boldsymbol{X}) = \sum_{\boldsymbol{Z}} \int p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \mathrm{d}\boldsymbol{\pi} \mathrm{d}\boldsymbol{\mu} \mathrm{d}\boldsymbol{\Lambda} \quad (5)$$

Il est à présent nécessaire d'écrire la loi jointe $p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$. Des lois *a priori* conjuguées sont choisies pour rendre les calculs réalisables pour le cas de lois gaussiennes [2, chapitre 10]. Notons que ce genre de développement a été réalisé dans le cas de lois de Poisson [6]. Ces distributions assurent que les lois *a priori* et *a posteriori* appartiennent à la même famille. De plus, les K couples $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ sont supposés indépendants. Ainsi, la distribution du couple $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ peut être décomposée comme un produit de lois normale-Wishart :

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \underbrace{\mathcal{NW}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k; \boldsymbol{m}_k, \beta_k, \boldsymbol{W}_k, \nu_k)}_{p(\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k) p(\boldsymbol{\Lambda}_k)} \quad (6)$$

où $p(\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k)$ est une loi normale de moyenne $\boldsymbol{m}_k$ et de matrice de précision $\beta_k \boldsymbol{\Lambda}_k$ tandis que $p(\boldsymbol{\Lambda}_k)$ est une distribution de Wishart de paramètre d'échelle $\boldsymbol{W}_k$ et dont le nombre de degrés de liberté est ν_k. $\boldsymbol{m}_k \in \mathbb{R}^M$, $\boldsymbol{W}_k \in \mathbb{R}^{M \times M}$ et, β_k et ν_k sont des scalaires.

La distribution sur l'ensemble des coefficients de mélange est donnée par une loi de Dirichlet :

$$p(\boldsymbol{\pi}) = \mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}) \quad (7)$$

dont la constante de normalisation $C(\boldsymbol{\alpha})$ est définie par :

$$C(\boldsymbol{\alpha}) = \frac{\Gamma(\overline{\alpha})}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_K)} \quad \text{et} \quad \overline{\alpha} = \sum_{k=1}^{K} \alpha_k \quad (8)$$

où $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$ est un vecteur de valeurs positives. En considérant ensuite les indépendances conditionnelles, la loi jointe peut être factorisée en :

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad (9)$$

Pour calculer l'évidence, il est nécessaire de marginaliser cette expression suivant les paramètres : $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ et $\boldsymbol{\Lambda}$ et d'effectuer une somme sur $\boldsymbol{Z}$. Il est important de réarranger les différents facteurs de la loi jointe pour réaliser l'intégration.

Les coefficients de mélange apparaissent à la fois dans la distribution de $\boldsymbol{Z}$ et dans la loi *a priori* sur $\boldsymbol{\pi}$. L'expression obtenue en faisant le produit de ces facteurs est proportionnelle à la distribution *a posteriori* de $\boldsymbol{\pi}$:

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi}) = \frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\alpha}^p)} p(\boldsymbol{\pi} \mid \boldsymbol{Z}) \quad (10)$$

$\boldsymbol{\alpha}^p$ est le vecteur de paramètres de la densité *a posteriori* de $\boldsymbol{\pi}$. Celle-ci est obtenue en utilisant la propriété de conjugaison :

$$p(\boldsymbol{\pi} \mid \boldsymbol{Z}) = \mathcal{D}ir(\boldsymbol{\pi}; \boldsymbol{\alpha}^p)$$

avec

$$\eta_k^p = \sum_{n=1}^{N} z_{nk} \quad \text{et} \quad \alpha_k^p = \eta_k^p + \alpha_k \tag{11}$$

Notons que η_k^p est le nombre d'observations attribuées à la k-ème classe pour une configuration particulière de $\boldsymbol{Z}$.
En procédant de même pour les lois sur les couples $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ et la vraisemblance des paramètres attachée aux données, il s'ensuit que :

$$p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{X}, \boldsymbol{Z}) \times (2\pi)^{-NM/2} \times \prod_{k=1}^{K} \left[\left(\frac{\beta_k}{\beta_k^p} \right)^{M/2} \frac{B(\boldsymbol{W}_k, \nu_k)}{B(\boldsymbol{W}_k^p, \nu_k^p)} \right] \tag{12}$$

avec $B(\boldsymbol{W}_k, \nu_k)$ la constante de normalisation de la distribution Wishart définie par :

$$B(\boldsymbol{W}_k, \nu_k) = \left[|\boldsymbol{W}_k|^{\nu_k/2} \left(2^{\nu_k M/2} \Gamma_M(\nu_k/2) \right) \right]^{-1}$$

où $\Gamma_M(.)$ est la fonction Gamma multivariée. $\boldsymbol{W}_k^p$, ν_k^p et β_k^p sont les paramètres de la loi *a posteriori* de $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. Par conjugaison des lois, celle-ci s'exprime comme :

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \boldsymbol{X}, \boldsymbol{Z}) = \prod_{k=1}^{K} \{ \mathcal{NW}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k; \boldsymbol{m}_k^p, \beta_k^p, \boldsymbol{W}_k^p, \nu_k^p) \}$$

où

$$\beta_k^p = \beta_k + \eta_k^p \tag{13}$$

$$\nu_k^p = \nu_k + \eta_k^p \tag{14}$$

$$\overline{\boldsymbol{x}}_k^p = \frac{1}{\eta_k^p} \sum_{n=1}^{N} z_{nk} \boldsymbol{x}_n \tag{15}$$

$$\boldsymbol{m}_k^p = \frac{1}{\beta_k^p} (\beta_k \boldsymbol{m}_k + \eta_k^p \overline{\boldsymbol{x}}_k^p) \tag{16}$$

$$\boldsymbol{S}_k^p = \frac{1}{\eta_k^p} \sum_{n=1}^{N} z_{nk} (\boldsymbol{x}_n - \overline{\boldsymbol{x}}_k^p)(\boldsymbol{x}_n - \overline{\boldsymbol{x}}_k^p)^T \tag{17}$$

$$(\boldsymbol{W}_k^p)^{-1} = \boldsymbol{W}_k{}^{-1} + \eta_k^p \boldsymbol{S}_k^p + \frac{\beta_k \eta_k^p}{\beta_k + \eta_k^p} (\overline{\boldsymbol{x}}_k^p - \boldsymbol{m}_k)(\overline{\boldsymbol{x}}_k^p - \boldsymbol{m}_k)^T \tag{18}$$

$\overline{\boldsymbol{x}}_k^p$ et $\boldsymbol{S}_k^p$ sont respectivement les moyenne et matrice de covariance empiriques des observations attribuées à la k-ème classe pour une configuration $\boldsymbol{Z}$ donnée. De plus, il est possible de marginaliser séparément suivant chaque couple du fait de l'hypothèse d'indépendance.

Après intégration sur $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ et $\boldsymbol{\Lambda}$, l'expression de l'évidence pour un choix de modèle $K = m$ devient alors :

$$p(\boldsymbol{X}) = (2\pi)^{-NM/2} \sum_{\boldsymbol{Z}} \frac{C(\boldsymbol{\alpha})}{C(\boldsymbol{\alpha}^p)} \prod_{k=1}^{K} \left[\left(\frac{\beta_k}{\beta_k^p} \right)^{M/2} \frac{B(\boldsymbol{W}_k, \nu_k)}{B(\boldsymbol{W}_k^p, \nu_k^p)} \right] \tag{19}$$

Pour notre étude, nous avons considéré le même ensemble d'hyperparamètres noté $\{\boldsymbol{m}_0, \beta_0, \boldsymbol{W}_0, \nu_0\}$ pour les K couples et un vecteur $\boldsymbol{\alpha}$ composé des mêmes valeurs α_0.

Pour un nombre d'observations N donné et pour un modèle $K = m$ fixé, le nombre de configurations en $\boldsymbol{Z}$ à considérer est égal à $K!\, S(N, K)$. $S(N, K)$ est le nombre de Stirling de seconde espèce qui donne le nombre de partitions uniques d'un ensemble à N éléments en K classes distinctes et non vides. Après avoir construit ces partitions uniques, il suit que le numéro de la classe attribuée à une observation importe. Pour une partition donnée, il existe $K!$ configurations identiques à une permutation d'étiquettes près, les termes de la somme (19) correspondant à ces derniers sont identiques. Nous calculons donc, pour chaque partition unique le terme de la somme (19) correspondant et le multiplions par $K!$.

4 Approches alternatives

À titre de comparaison, nous considérons un algorithme de moyenne harmonique et le BIC dont nous rappelons brièvement le principe.

Algorithme de la moyenne harmonique

Cet algorithme [7] consiste à approcher la valeur de $p(\boldsymbol{X})$ pour un modèle $K = m$ considéré via un algorithme Markov Chain Monte Carlo (MCMC).
Le principe est le suivant. La loi jointe peut être réécrite suivant le théorème de Bayes :

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) = p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}) p(\boldsymbol{Z}, \boldsymbol{\theta}) = p(\boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X}) p(\boldsymbol{X})$$

$$\Rightarrow \frac{p(\boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X})}{p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})} = \frac{p(\boldsymbol{Z}, \boldsymbol{\theta})}{p(\boldsymbol{X})} \tag{20}$$

Par intégration des deux rapports, il vient que :

$$\sum_{\boldsymbol{Z}} \int \frac{p(\boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X})}{p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} = \sum_{\boldsymbol{Z}} \int \frac{p(\boldsymbol{Z}, \boldsymbol{\theta})}{p(\boldsymbol{X})} \mathrm{d}\boldsymbol{\theta}$$

$p(\boldsymbol{X})$ n'est pas fonction des paramètres $\{\boldsymbol{Z}, \boldsymbol{\theta}\}$ et $\sum_{\boldsymbol{Z}} \int p(\boldsymbol{Z}, \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 1$. Par conséquent, l'égalité devient :

$$[p(\boldsymbol{X})]^{-1} = \mathbb{E}_{\boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X}} \left[[p(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})]^{-1} \right] \tag{21}$$

L'évidence peut donc être approchée de la manière suivante :

- Générer L échantillons $\{[\boldsymbol{Z}^{(1)}, \boldsymbol{\theta}^{(1)}], \ldots, [\boldsymbol{Z}^{(L)}, \boldsymbol{\theta}^{(L)}]\}$ suivant la distribution $p(\boldsymbol{Z}, \boldsymbol{\theta} \mid \boldsymbol{X})$ à l'aide d'un algorithme MCMC, dans notre cas un algorithme de Gibbs.
- La loi des grands nombres permet d'en déduire que :

$$p(\boldsymbol{X}) \simeq \left[\frac{1}{L} \sum_{i=1}^{L} [p(\boldsymbol{X} \mid \boldsymbol{Z}^{(i)}, \boldsymbol{\theta}^{(i)})]^{-1} \right]^{-1} \tag{22}$$

Formule de l'algorithme BIC

L'algorithme BIC [5], fondé sur l'approximation de Laplace, consiste à approcher asymptotiquement l'évidence par :

$$\log(p(\boldsymbol{X} \mid K = m)) \simeq \log(p(\boldsymbol{X} \mid \widehat{\boldsymbol{\theta}}_m)) - \frac{\lambda_m}{2} \log(N) \tag{23}$$

où $\widehat{\boldsymbol{\theta}}_m$ est l'estimation du maximum de vraisemblance pour un modèle $K = m$ donné et λ_m est la dimension de $\boldsymbol{\theta}_m$. La sélection du meilleur modèle revient donc à choisir $\widehat{K} = m_{BIC}$ tel que $m_{BIC} = \underset{m}{\operatorname{argmin}}[-2 \log(p(\boldsymbol{X} \mid \widehat{\boldsymbol{\theta}}_m)) + \lambda_m \log(N)]$.

5 Résultats

Nous avons appliqué les trois méthodes pour $N = 12$ et un mélange de deux lois gaussiennes scalaires de paramètres : $\boldsymbol{\mu} = [-1 \ \ 1]$, $\boldsymbol{\Lambda} = [5 \ \ 5]$ et $\boldsymbol{\pi} = [0.5 \ \ 0.5]$. Pour les lois *a priori*, nous avons choisi des valeurs pour les hyperparamètres conduisant à des lois peu informatives : $\boldsymbol{m}_0 = 0$, $\beta_0 = 1$, $\boldsymbol{W}_0 = 10^4$, $\nu_0 = 1$ et $\alpha_0 = 1$. Nous avons simulé 100 réalisations du mélange pour évaluer les performances des trois approches en nous limitant à un nombre maximal de classes égal à 4.

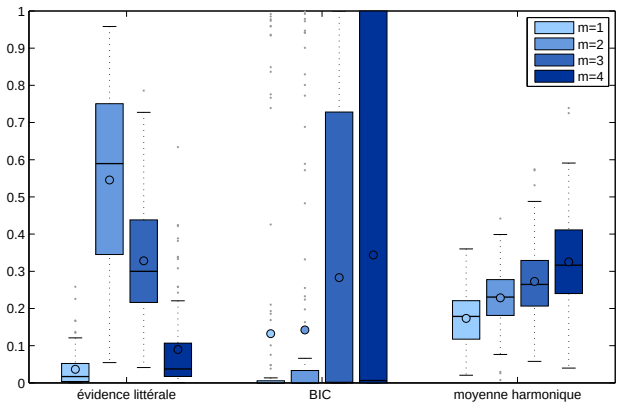

FIGURE 1 – Comparaison de la probabilité a posteriori des modèles pour $N = 12$

Sur la figure 1, sont représentées sous forme de boîte à moustaches les probabilités *a posteriori* des modèles calculées pour chacune des méthodes. La valeur moyenne est représentée par un rond. Nous remarquons que dans les conditions de simulation considérées, seule notre approche conduit à une bonne estimation du nombre de classes. En effet, comme attendu, le BIC qui est une approche asymptotique ne permet pas de prendre la bonne décision si le nombre d'observations est trop réduit : le modèle d'ordre le plus grand a toujours la probabilité la plus élevée. En ce qui concerne l'algorithme de la moyenne harmonique, il fournit aussi une mauvaise estimation de la probabilité *a posteriori* des différents modèles. Ce résultat met en évidence les problèmes liés à cet estimateur. Comme souligné dans [8], il n'est approprié que si la loi *a priori* a une forte influence sur la loi *a posteriori* comparativement aux données. Il est donc très sensible au réglage des hyperparamètres. Pour illustrer cette remarque, nous avons représenté en figure 2 les résultats obtenus avec le réglage suivant : $\boldsymbol{W}_0 = 1$, qui conduit à une loi plus informative. Dans ce cas, nous observons en effet que la sélection de modèle est correctement réalisée.

6 Conclusion

Nous considérons dans cette communication la question difficile de l'estimation du nombre de composantes d'un mélange de lois gaussiennes. Dans un cadre bayésien, notre approche est fondée sur un calcul explicite de l'évidence, rendu possible par le choix de lois *a priori* conjuguées. Notre méthode ne peut s'appliquer à un nombre trop grand d'observations en raison

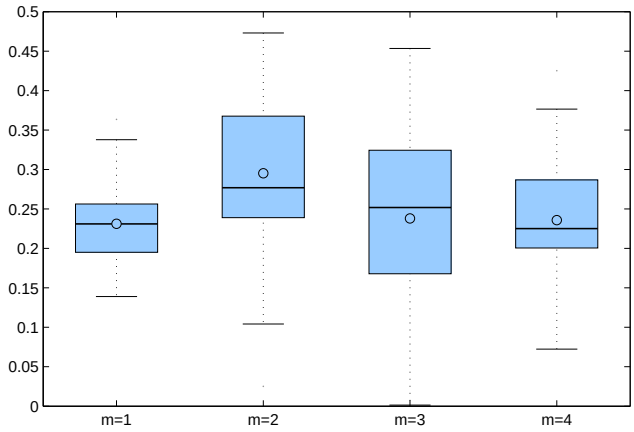

FIGURE 2 – Probabilité *a posteriori* des modèles dans le cas informatif donnée par l'algorithme de la moyenne harmonique

de la limitation de la puissance de calcul. Cependant, dans ce contexte délicat, nous montrons son intérêt par rapport à des méthodes classiques de sélection d'ordre. Nous étudions actuellement l'influence des différents facteurs intervenant dans notre formule en fonction des valeurs des hyperparamètres des lois *a priori* afin d'en proposer une approximation applicable à un plus grand nombre d'individus.

Références

[1] K.-A. Do, P. Muller et M. Vannucci, *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, 2006.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] G. Celeux et G. Goavert, *A classification EM algorithm for clustering and two stochastic versions*. Compuational Statistics Quaterly, vol 2, num 1, pp 73-82, 1991.

[4] L. Shi, S. Tu et L. Xu, *Learning gaussian mixture with automatic model selection : a comparative study on three Bayesian related approaches*. Front. Electr. Electron. Eng. China, pp 215-244, 2011.

[5] Y.-X. Geng et W. Wu, *A bayesian information criterion based approach for model complexity selection in speaker identification*. IEEE Computer Society, pp 264-268, 2008.

[6] P. Fearnhead, *Direct simulation for discrete mixture distributions*. Statistics and Computing, vol 15, Issue 2, pp 125-133, 2005.

[7] A. Raftery, M. Newton, J. Satagopan et P. Krivitsky, *Estimating the integrated likelihood via posterior simulation using the harmonic mean identity*. Bayesian Statistics, pp 1-45, 2007.

[8] N. Chopin et C. Robert, *Comments on « Estimating the integrated likelihood via posterior simulation using the harmonic mean identity (with discussion) »*. Bayesian Statistics 8, edited by O. U. P. Bernardo, J. M. et al. (eds), pp 371-416, 2007.