Partition latente et dégénérescence dans les mélanges gaussiens

Christophe Biernacki

Partition latente et dégénérescence dans les mélanges gaussiens

Christophe Biernacki

Université Lille 1 & CNRS, Villeneuve d'Ascq, France
biernack@math.univ-lille1.fr

Résumé

Dans le cas des mélanges gaussiens, il est notoire que la vraisemblance peut augmenter sans limite si, par exemple, une des gaussiennes est centrée en une observation et simultanément la matrice de variance correspondante tend vers la singularité. Les tentatives actuelles pour résoudre ce problème reposent sur la détention d'information hypothétique sur les matrices de variance elles-mêmes. Notre proposition consiste à introduire une information naturelle sur la partition latente impliquée dans le modèle génératif sous-jacent qui permet de borner la vraisemblance tout en préservant la convergence des estimateurs. Constatant que l'algorithme EM impliqué dans l'optimisation de cette vraisemblance particulière implique des difficultés combinatoires pour plus de deux composantes gaussiennes, l'information de partition peut être alternativement utilisée pour proposer une borne probabiliste inférieure non asymptotique sur les variances généralisées qui est simple à calculer dans la plupart des situations et qui conserve la convergence des estimateurs. Des expériences numériques illustrent les deux propositions.

Mots clés : Maximum de vraisemblance, algorithme EM, borne probabilisée, solutions pathologiques

Abstract

In the case of Gaussian mixtures, it is well-known that the likelihood function increases without bound if, for instance, one of the mixture means coincides with a sample observation and if the corresponding variance matrix tends to singularity. Previous attempts to avoid such situations systematically rely on some hypothetical additional information about the variance matrices themselves. Our proposal is to introduce some rational additional information on the latent partition involved in the generative scheme. It is shown that the corresponding likelihood is bounded and leads to consistent estimates. Since the associated EM algorithm is quite computational embarrassing for more that two components, the additional information alternatively allows to propose a quite simple and usually computationally tractable non-asymptotic stochastic lower bound on generalized variances, while still preserving consistency. Some numerical illustrations of both proposals are also given.

Keywords : Maximum likelihood, EM algorithm, stochastic lower bound, spurious solutions

1 Introduction

Reconnus comme méthode de modélisation particulièrement flexible, les mélanges gaussiens multivariés ont suscité un intérêt croissant au fil des années, tant du point de vue théorique que pratique. Cependant, il est bien connu que la fonction de vraisemblance associée n'admet pas de borne supérieure (Kiefer et Wolfowitz, 1956; Day, 1969), cette dérive arrivant par exemple quand une des gaussiennes est centrée en une observation et simultanément la matrice de variance correspondante tend vers la singularité (les situations conduisant à la dégénérescence sont plus genérales). Au delà des questions théoriques ainsi posées, le praticien est lui-même souvent confronté à des scénarios de dégénérescence lors de l'estimation par l'algorithme EM par exemple.

Les solutions classiques pour éviter la dégénérescence consistent à contraindre ou bien à transformer la fonction vraisemblance. Par exemple, Hathaway (1985), puis plus récemment Ingrassia et Rocci (2007), imposent des contraintes relatives entre matrices de variance. Alternativement, Ciuperca et al. (2003), dans le cas univarié, et Snoussi et Mahammad-Djafari (2001), dans le cas multivarié, adoptent un point de vue bayésien qui revient à pénaliser la vraisemblance par une loi *a priori* pénalisant lourdement les matrices de variances dégénérées. Contrairement aux autres tentatives, la convergence de l'estimateur est préservée. Cependant, chacune des propositions précédentes utilise une information sur l'espace des paramètres qui peut affecter la qualité de l'estimateur obtenu, au moins à taille finie, si elle est introduite de manière subjective.

L'originalité du présent travail est d'introduire une information sur la partition inconnue des données (intervenant dans l'interprétation générative du mélange) au lieu des matrices de variances comme cela est fait traditionnellement. L'hypothèse qui est faite est particulièrement faible et naturelle puisqu'elle revient à supposer tout simplement qu'au moins $d+1$ individus proviennent de chacune des composantes gaussiennes de dimension d.

2 Contrainte sur la partition latente

Le modèle de mélange gaussien Dans l'hypothèse de mélange gaussien multivarié, chaque individu $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de $\mathbb{R}^d$ ($n \geq g(d+1)$) provient de façon i.i.d. de la densité $f(\cdot; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ où π_k est la proportion de la k^{e} composante ($0 < \pi_k < 1$ pour $k = 1, \ldots, g$ et $\sum_k \pi_k = 1$) et où $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ correspond à la densité gaussienne de moyenne $\boldsymbol{\mu}$ and et de matrice de variance $\boldsymbol{\Sigma}$ ($|\boldsymbol{\Sigma}| > 0$ avec $|\boldsymbol{\Sigma}|$ la *variance généralisée* de $\boldsymbol{\Sigma}$). Le paramètre de mélange $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_g)$ évolue dans l'espace Θ et $\hat{\boldsymbol{\theta}}$ désigne l'estimateur de $\boldsymbol{\theta}$ qui maximise la vraisemblance $L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\theta})$ s'appuyant sur l'échantillon au complet $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Décomposition de la vraisemblance D'un point de vue génératif, le jeu de données $\mathbf{x}$ est construit en deux étapes successives : (i) tout d'abord une partition $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$

est obtenue par n réalisations i.d.d. $\mathbf{z}_i = (z_{i1}, \ldots, z_{ig})'$ de la loi multinomiale d'ordre un et de paramètre $(\pi_1, \ldots, \pi_g)$, $\mathbf{z}_i$ dénotant un vecteur binaire tel que $z_{ik} = 1$ si le i^{e} provient de la k^{e} composante et 0 sinon ; (ii) puis, conditionnellement à $\mathbf{z}$, chaque $\mathbf{x}_i$ est généré indépendamment par la composante gaussienne indiquée par z_{ik}, donc par la densité $\prod_{k=1}^{g} \phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}}$.

Notant $n_k = \sum_{i=1}^{n} z_{ik}$ l'effectif du k^{e} groupe de $\mathbf{z}$ et $\mathcal{Z}^* = \{\mathbf{z} : n_k \geq d+1\}$ l'ensemble des partitions contenant au moins $d+1$ individus dans chaque groupe, la vraisemblance $L(\boldsymbol{\theta}; \mathbf{x})$ se décompose en les deux autres vraisemblances suivantes :

$$L(\boldsymbol{\theta}; \mathbf{x}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathcal{Z}^*) + L(\boldsymbol{\theta}; \mathbf{x}, \bar{\mathcal{Z}}^*), \tag{1}$$

où $\bar{\mathcal{Z}}^*$ désigne le complémentaire de $\mathcal{Z}^*$. Il est facile de vérifier que maximiser la seconde vraisemblance $L(\boldsymbol{\theta}; \mathbf{x}, \bar{\mathcal{Z}}^*)$ conduit nécessairement soit à une dégénérescence (tous les optima locaux sont dégénérés), soit à une non identifiabilité des paramètres. Considérant donc uniquement les partitions $\mathcal{Z}^*$, la proposition suivante établit que la vraisemblance correspondante possède cette fois de bonnes propriétés : elle est bornée avec probabilité un dans ce cas, tout en préservant la convergence de l'estimateur associé.

Proposition 1 *Pour tout $\boldsymbol{\theta} \in \Theta$, alors $P(L(\boldsymbol{\theta}; \mathbf{x}, \mathcal{Z}^*) < \infty) = 1$. Par ailleurs l'estimateur de $\boldsymbol{\theta}$ obtenu en maximisant $L(\boldsymbol{\theta}; \mathbf{x}, \mathcal{Z}^*)$ est convergeant.*

Algorithme EM associé Afin de maximiser $L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z} \in \mathcal{Z}^*)$ seule l'étape E de l'algorithme EM diffère de la situation classique. Notant respectivement $\boldsymbol{\theta}$ et $\boldsymbol{\theta}^+$ les estimateurs associés à deux itérations successives de l'algorithme, on a alors :

- **Étape E** : calculer les probabilités conditionnelles $t^*_{ik} = E[z_{ik}|\mathbf{x}, \mathcal{Z}^*; \boldsymbol{\theta}]$.
- **Étape M** : utiliser les formules traditionnelles de calcul de $\boldsymbol{\theta}^+$ en remplaçant simplement les probabilités conditionnelles traditionnelles $t_{ik} = E[z_{ik}|\mathbf{x}; \boldsymbol{\theta}]$ par t^*_{ik}.

Il est facile de voir que $t^*_{ik} \propto p_{ik} t_{ik}$ avec $p_{ik} = P(\mathcal{Z}^*|\mathbf{x}, \{z_{ik} = 1\}; \boldsymbol{\theta})$ donc la nouvelle étape E revient finalement à repondérer les valeurs habituelles t_{ik}. Par ailleurs, on peut prouver que pour tout $\boldsymbol{\theta} \in \Theta$ et pour tout $i \in \{1, \ldots, n\}$, $p_{ik} \to 1$ quand $n \to \infty$ et donc les anciennes et nouvelles versions de EM sont équivalentes pour les grandes valeurs de n.

L'implémentation du nouveau EM repose essentiellement sur le calcul des termes p_{ik} qui s'expriment comme de simples produits et sommes des probabilités t_{ik}. Malgré ce caractère explicite rassurant, le nombre d'opérations élémentaires augmente très fortement avec n et g. En particulier, dès $g > 2$ la combinatoire rend les calculs impossibles, ce qui justifie alors de l'approche alternative que nous décrivons maintenant.

3 Borne sur les variances généralisées

Obtention de la borne Dans le but de détecter toute dégénérescence dans la vraisemblance traditionnelle, nous proposons une borne inférieure facile à calculer sur les

variances généralisées $|\mathbf{\Sigma}_k|$ de chaque composante $k = 1, \ldots, g$. L'originalité repose sur le fait que cette borne est stochastique, non-asymptotique et adaptative. Comme l'établit la proposition suivante, son obtention repose uniquement sur l'hypothèse naturelle $\mathcal{Z}^*$ concernant les partitions et déjà faite dans la section précédente.

Proposition 2 *Soit $\mathbf{y}_1, \ldots, \mathbf{y}_c$ l'ensemble des $c = \binom{n}{d+1}$ combinaisons de $d+1$ individus parmi n où $\mathbf{y}_\ell = (y_{1\ell}, \ldots, y_{n\ell})'$ est la ℓ^e combinaison avec $y_{i\ell} = 1$ si $\mathbf{x}_i$ y est associé et $y_{i\ell} = 0$ sinon ($i = 1, \ldots, n$, $\ell = 1, \ldots, c$). Soit aussi $\bar{\mathbf{x}}_\ell$ la moyenne empirique de la ℓ^e combinaison et $\mathbf{V}_\ell$ la matrice de variance empirique correspondante, toutes deux obtenues à partir des formules standards*

$$\bar{\mathbf{x}}_\ell = \frac{1}{d+1} \sum_{i=1}^{n} y_{i\ell} \mathbf{x}_i \quad \text{et} \quad \mathbf{V}_\ell = \frac{1}{d+1} \sum_{i=1}^{n} y_{i\ell} (\mathbf{x}_i - \bar{\mathbf{x}}_\ell)(\mathbf{x}_i - \bar{\mathbf{x}}_\ell)', \tag{2}$$

et soit aussi $\mathbf{V}^*$ la matrice de variance empirique $\mathbf{V}_\ell$ ($\ell = 1, \ldots, c$) de déterminant minimum :

$$\mathbf{V}^* = \arg \min_{\mathbf{W} \in \{\mathbf{V}_\ell\}} |\mathbf{W}|. \tag{3}$$

Alors, sous l'hypothèse que $\mathcal{Z}^*$ est vrai, pour tout $k \in \{1, \ldots, g\}$ et $\alpha \in (0, 1)$, on a

$$P\left(|\Sigma_k| > \frac{(d+1)^d |\mathbf{V}^*|}{\left(\chi^2_{n-1-(g-1)(d+1),(1-\alpha)^{1/d}} \right)^d} \right) \geq 1 - \alpha, \tag{4}$$

où $\chi^2_{a,\alpha}$ est le quantile de χ^2 avec a degrés de liberté et d'ordre α.

Cette borne repose sur le calcul de la variance généralisée $|\mathbf{V}^*|$ qui est indépendant de g mais demande $c = \binom{n}{d+1}$ évaluations de la variance généralisée associée aux échantillons de taille $d+1$. La dégénérescence arrivant essentiellement dans le cas de petites tailles d'échantillon (Biernacki, 2007), le temps d'évaluation reste donc modéré dans la plupart des cas d'intérêt. On peut s'attendre en outre à ce que la borne ne soit pas très précise puisqu'elle est peut être vérifiée avec une probabilité bien supérieure à $1 - \alpha$ dans de nombreux cas mais cependant elle est suffisante pour la stratégie que nous décrivons maintenant.

Stratégie d'utilisation En pratique, la stratégie que nous proposons pour empêcher la dégénérescence de la vraisemblance classique $L(\boldsymbol{\theta}; \mathbf{x})$ est très simple, facilement implémentable et conduit à des estimateurs convergeants. Elle consiste à écarter tous chemin de EM (version classique de EM) dès que la borne probabiliste précédente est atteinte par au moins une variance généralisée des composantes gaussiennes.

Cette stratégie est beaucoup plus économique que de maximiser la vraisemblance $L(\boldsymbol{\theta}; \mathbf{x}, \mathcal{Z}^*)$ avec EM comme décrit à la section précédente. En effet, comme déjà mentionné, le coût de calcul ne dépend plus de g. Par ailleurs, la borne ne doit être calculée qu'une fois pour toutes et non à chaque itération de EM.

4 Illustration numérique

Nous nous limitons ici à la présentation des expériences menées dans le cas univarié afin d'illustrer très simplement les outils théoriques précedents. Cependant, des résultats obtenus dans des situations multivariées et sur données réelles (non reportés ici) amènent aux mêmes conclusions.

Un échantillon de taille $n = 20$ est généré à partir du mélange univarié bimodal suivant : $\pi_1 = \pi_2 = 0.5, \quad \mu_1 = 0, \quad \mu_2 = 4, \quad \sigma_1^2 = \sigma_2^2 = 1$, où σ_k^2 est la variance de la k^{e} composante. La borne probabiliste sur σ_k^2 est $\hat{\sigma}_{\text{inf}}^2 = 1.35\ 10^{-5}$ et est obtenue en un centième de seconde sur un ordinateur "standard".

Tout d'abord, seulement la valeur de σ_1^2 est inconnue et doit être estimée, la valeur μ_1 étant fixée à la valeur d'un des individus et tous les autres paramètres prenant leur vraie valeur. La figure 1 représente les log-vraisemblances $\ln L(\sigma_1^2; \mathbf{x}, \mathcal{Z}^*)$ et $\ln L(\sigma_1^2; \mathbf{x})$ pour différentes valeurs de σ_1^2. Nous notons, comme attendu, que (i) $\ln L(\sigma_1^2; \mathbf{x}, \mathcal{Z}^*)$ ne croît pas indéfiniment quand $\sigma_1^2 \to 0$, mais, au contraire décroît ; (ii) $\ln L(\sigma_1^2; \mathbf{x})$ croît sans limite quand $\sigma_1^2 \to 0$, mais la borne $\hat{\sigma}_{\text{inf}}^2$ détecte clairement cette dynamique dégénérescente ; (iii) les deux vraisemblances sont indissociables à grande distance de la situation dégénérée $\sigma_1^2 = 0$.

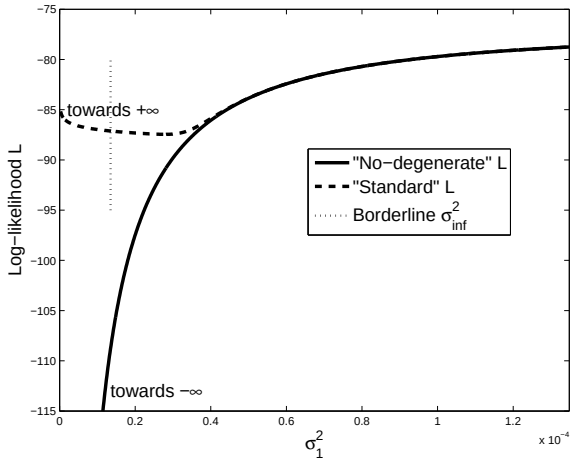
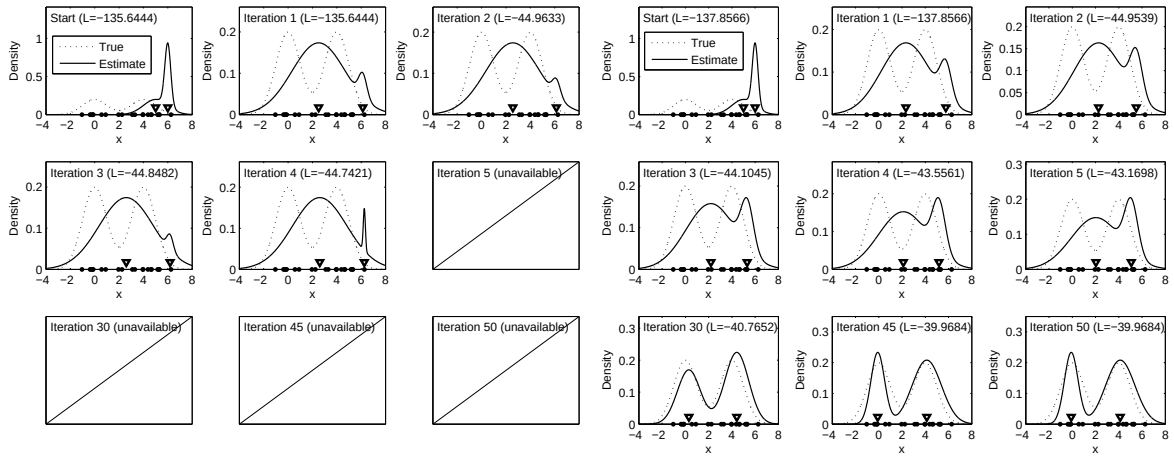


FIG. 1 – Comparaison de $\ln L(\sigma_1^2; \mathbf{x}, \mathcal{Z}^*)$ et $\ln L(\sigma_1^2; \mathbf{x})$ au voisinage de la dégénérescence.

Dans un second temps, tous les paramètres sont de $\boldsymbol{\theta}$ sont inconnus et doivent être estimés. Les figures 2 (a) et (b) proposent une situation où le EM standard conduit à la dégénérescence tandis que (i) la stratégie de borne inférieure détecte cette dynamique et l'arrête et (ii) aucune dégénérescence n'apparaît avec la version modifiée de EM. Notons d'ailleurs que l'estimation ainsi fournie est très proche des vrais paramètres du mélange.



(a) (b)

FIG. 2 – (a) EM classique avec borne inférieure sur la variance et (b) EM ayant la propriété de non dégénérescence. Les triangles indiquent la positions des moyennes des composantes.

Références

[1] Biernacki, C. (2007). Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures for grouped data and behaviour of the EM algorithm. *Scandinavian Journal of Statistics*, 34, 569–586.

[2] Ciuperca, G., Ridolfi, A., and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30(1), 45–59.

[3] Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika*, 56, 463–474.

[4] Hathaway, R. (1985). A constrained formulation of maximum-likelihood estimation for normal distributions. *Annals of Statistics*, 13, 795–800.

[5] Ingrassia, S. and Rocci, R. (2007). Constrained monotone em algorithms for finite mixture of multivariate gaussians. *Computational Statistics & Data Analysis*, 51, 5339–5351.

[6] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, 127, 887–906.

[7] Snoussi, H. and Mahammad-Djafari, A. (2001). Penalized maximum likelihood for multivariate gaussian mixture. In *Bayesian Inference and Maximum Entropy Methods (MaxEnt)*, pages 36–46. American Institute of Physics.