A Data-Driven Bound on Covariance Matrices for Avoiding Degeneracy in Multivariate Gaussian Mixtures

Christophe Biernacki, Gwénaelle Castellan

A Data-Driven Bound on Covariance Matrices for Avoiding Degeneracy in Multivariate Gaussian Mixtures

Christophe Biernacki [1] & Gwénaelle Castellan [2]

[1] *University Lille 1 & CNRS & INRIA, Villeneuve d'Ascq, France,*
Christophe.Biernacki@math.univ-lille1.fr
[2] *University Lille 1 & CNRS, Villeneuve d'Ascq, France,*
Gwenaelle.Castellan@math.univ-lille1.fr

Résumé. Le fait que la vraisemblance ne soit pas bornée dans les mélanges gaussiens est un handicap pratique et théorique. Utilisant la très faible hypothèse que chaque composante est d'effectif supérieur à la dimension de l'espace, nous proposons une borne aléatoire exacte très simple qui permet de contrôler les valeurs propres des matrices de covariances. Dans le cas univarié, l'estimateur du maximum de vraisemblance sous cette contrainte est convergeant, la preuve restant encore à établir dans le cas général. Cette stratégie est implémentée dans un algorithme EM et donne d'excellents résultats sur des données simulées.

Mots-clés. EM, dégénérescence, vraisemblance, borne non asymptotique

Abstract. Unbounded likelihood for multivariate Gaussian mixture is an important theoretical and practical problem. Using the weak information that the latent sample size of each component has to be greater than the space dimension, we derive a simple strategy relying on non-asymptotic stochastic lower bounds for monitoring singular values of the covariance matrix of each component. Maximizing the likelihood under this data-driven constraint is proved to give consistent estimates in the univariate situation, consistency for the multivariate case being still to establish. This strategy is implemented in an EM algorithm and its excellent performance is assessed through simulated data.

Keywords. EM, Gaussian mixture, degeneracy, likelihood, non-asymptotic bound

1 Introduction

Because Gaussian mixtures models are an extremely flexible method of modeling, they received increasing attention over the years, from both practical and theoretical points of view. Various approaches to estimate mixture distributions are available (see McLachlan and Peel, 2000, for a survey), including the method of moments, the Bayesian methodology or the maximum likelihood (ML) approach, the latter being usually much preferred. Nevertheless, it is well-known that the likelihood function of normal mixture models is not

bounded from above (Kiefer and Wolfowitz, 1956; Day, 1969). As a consequence, firstly some theoretical questions about the ML properties are raised and, secondly, optimization algorithms like EM (Dempster et al., 1977; Redner and Walker, 1984) may converge, as observed by any practicioner, towards such degenerate solutions.

Avoiding degeneracy is usually handled by constraining the singular values of the covariance matrices (so the variances in the univariate situation). The main option consists to constraint them to be greater than a given "small" value. Such a bound can be either arbitrarily chosen (typically the numerical tolerance of computer for many practitioners) or chosen in a smarter way for ensuring consistency of the constraint ML (Tanaka and Takemura, 2006). Another way is to impose relative constraints between singular values (Hathaway, 1985; Ingrassia and Rocci, 2007). Alternatively, Policello (1981) imposed a constraint on the latent partition underlying the data (instead of a constraint on the singular values), that leads to maximize a bounded likelihood and gives consistent estimates. The proposed assumption is weak and natural since it only requires that at least $d+1$ data units arise from each d-variate mixture Gaussian component. However, maximizing this likelihood is untractable in most situations because of combinatorial difficulties.

Using such a weak assumption on the latent partition, the present work establishes a non-asymptotic stochastic lower bound specific to each singular value and to each component. This data-driven lower bound is very simple to calculate from the sample and leads to consistent estimates of the mixture in the univariate case, consistency in the multivariate case being expected but the proof being still in progress. This strategy can be used by any practitioner with poor modification of its preferred ML software, like EM.

The outline of this paper is the following. In Section 2, we present the degeneracy problem and we introduce the constraint on the latent partition. The derived data-driven non-asymptotic stochastic lower bound on the singular values is obtained and studied in Section 3. The last section (Section 4) present numerical examples.

2 Linking latent partition with degeneracy

2.1 Observed-data likelihood and degeneracy

In the Gaussian mixture model assumption, each individual $\boldsymbol{X}_i \in \mathbb{R}^d$ of the data set $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ i.i.d. arises from the density

$$f(\cdot; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi(\cdot; \boldsymbol{\mu}_k, \Sigma_k)$$

where π_k is the mixing proportion of the kth component ($0 < \pi_k < 1$ for all $k = 1, \ldots, g$ and $\sum_k \pi_k = 1$) and where $\phi(.; \boldsymbol{\mu}_k, \Sigma_k)$ denotes the density of the Gaussian distribution of this kth component with mean $\boldsymbol{\mu}_k$ and covariance matrix Σ_k ($|\Sigma_k| > 0$ for all $k = 1, \ldots, g$). These natural constraints on the mixture parameter $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_g, \Sigma_1, \ldots, \Sigma_g)$

are summarized in the parameter space Θ. It is well-known that the *observed-data* likelihood defined by

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = \prod_{i=1}^{n} f(\boldsymbol{X}_i; \boldsymbol{\theta}) \tag{1}$$

is unbounded from above (Kiefer and Wolfowitz, 1956; Day, 1969). Indeed, in some situations where $|\Sigma_k| \to 0$ for a given $k \in \{2, \ldots, g\}$ it can arise that$L(\boldsymbol{\theta}; \boldsymbol{X}) \to \infty$. It corresponds to the so-called *degeneracy*.

2.2 Constraining the likelihood with the latent partition

From a generative point of view, the data set $\boldsymbol{X}$ is built from the two following sequential steps:

1. First a partition $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ is obtained by n i.d.d. realizations $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{ig})'$ of the multinomial distribution of order one and of parameter $(\pi_1, \ldots, \pi_g)$, $\boldsymbol{Z}_i$ denoting a binary vector where $Z_{ik} = 1$ if the ith data unit arises from the kth component and 0 otherwise.

2. Then, conditionally to $\boldsymbol{Z}_i$, each $\boldsymbol{X}_i$ is independently generated from the Gaussian component indicated by Z_{ik}.

In mixture models, $\boldsymbol{Z}$ is *latent*, but if it were known the *complete-data* likelihood

$$L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{g} [\pi_k \phi(X_i; \boldsymbol{\mu}_k, \Sigma_k)]^{Z_{ik}}$$

could be maximized on $\boldsymbol{\theta}$. However, even in this full data case, degeneracy can arise as soon as a given $k \in \{1, \ldots, g\}$ is such that $N_k \leq d$, where $N_k = \sum_{i=1}^{n} Z_{ik}$ denotes the number of individuals arising from the kth component. Consequenly, the unique solution for avoiding degeneracy to occur (with probability one) with complete-data likelihood in the general Gaussian case is to impose the constraint $\mathcal{Z}$ on $\boldsymbol{Z}$ where

$$\mathcal{Z} = \{\boldsymbol{Z} : \ N_k \geq d + 1, \ k = 1, \ldots, g\}$$

is the set of all partitions containing at least $d + 1$ individuals from each component.

Starting from this remark, Policello (1981) proposed to maximize a likelihood taking into account the additional information $\mathcal{Z}$ on the latent partition $\boldsymbol{Z}$. He chooses to maximize the *conditional* likelihood $L(\boldsymbol{\theta}; \boldsymbol{X}|\mathcal{Z})$ and establishes that it is now bounded and leads to consistent estimates. He gives the detail of a specific EM algorithm for maximizing $L(\boldsymbol{\theta}; \boldsymbol{X}|\mathcal{Z})$ but it is computational untractable as soon as $g > 2$. We will overcome this difficulty in the next section by proposing an alternative solution through a cheaper computational lower bound on the singular values.

3 A data-driven bound on singular values

3.1 Establishing the bound

In order to prevent the (traditional) likelihood (1) to degenerate, we propose now alternatively a lower bound on the jth singular value λ_{jk} of the kth component, for all $j = 1, \ldots, d$ and all $k = 1, \ldots, g$. Originality relies on the fact that this bound is stochastic, non asymptotic and data-driven. The next proposition establishes it by using the weak assumption $\mathcal{Z}$ on $\boldsymbol{Z}$ discussed above. In the following, X_{ijk} will denote the projection of $\boldsymbol{X}_i$ on the axis associated to the eigenvector of the singular value λ_{jk} and

$$S_{\mathcal{I}jk} = \sum_{i \in \mathcal{I}} (X_{ijk} - \bar{X}_{\mathcal{I}jk})^2$$

the non-normalized empirical variance of the subsample $\{X_{ijk}\}_{i \in \mathcal{I}}$ ($\mathcal{I} \subset \{1, \ldots, n\}$) with

$$\bar{X}_{\mathcal{I}jk} = \frac{1}{\#\mathcal{I}} \sum_{i \in \mathcal{I}} X_{ijk}$$

the corresponding empirical mean.

Proposition 1 *For any $\alpha \in (0, 1)$, we have,*

$$\mathbb{P}\left(\forall k \in \{1, \ldots, g\} \ , \ \lambda_{jk} \geq B^d_{jk}(\alpha) \mid \mathcal{Z}\right) \geq 1 - \alpha,$$

where $\chi^2_\nu(\alpha)$ denotes the quantile of χ^2 with ν degrees of freedom and of order α and where

$$B^d_{jk}(\alpha) = \frac{S^d_{jk}}{\chi^2_d(1-\alpha)}$$

with S^d_{jk} the minimum non-normalized variance among all subsamples of size $d+1$ in the whole sample $\{X_{ijk}\}_{i \in \{1,\ldots,n\}}$:

$$S^d_{jk} = \min_{\{\mathcal{I}:\#\mathcal{I}=d+1\}} S_{\mathcal{I}jk}.$$

The proof is straightforward. Notice first that the whole sample $\{X_{ijk}\}_{i \in \{1,\ldots,n\}}$ i.i.d. arises from a univariate Gaussian mixture with g components since it is the projection of a multivariate Gaussian mixture of g components. We note $\sigma^2_{kj\{1\}}, \ldots, \sigma^2_{kj\{g\}}$ the corresponding variances. Let $k_0 \in \{1, \ldots, g\}$ be the class number with the smallest variance, so such that $\sigma^2_{jk\{k_0\}} = \min_{1 \leq k' \leq g} \sigma^2_{jk\{k'\}}$. Conditionally to the event $\mathcal{Z}$, there exists $d+1$ distinct random variables $\{X_{ijk}\}_{i \in \{\mathcal{I}\}}$ which belong to the class k_0 (so such that $Z_{ik_0} = 1$ for all $i \in \mathcal{I}$). Since $\{X_{ijk}\}_{i \in \{\mathcal{I}\}}$ is a i.i.d. sample from a univariate Gaussian, then $S_{\mathcal{I}jk}$ have a Chi-square distribution with d degrees of freedom. Thus we deduce that, for any $\alpha \in (0, 1)$,

$$\mathbb{P}\left(\sigma^2_{jk\{k_0\}} \geq \frac{S_{\mathcal{I}jk}}{\chi^2_d((1-\alpha))} \Big| \{i \in \mathcal{I} : Z_{i,k_0} = 1\}, \mathcal{Z}^*\right) = 1 - \alpha.$$

We conclude by noticing two points: (1) $\lambda_{jk} \geq \sigma^2_{jk\{k_0\}}$ since $\lambda_{jk} = \sigma^2_{jk\{k\}}$; (2) $S^d_{jk} \leq S_{\mathcal{I}jk}$.

Remarks

- Note that the bound is easy and fast to compute using the following equality:

$$S_{jk}^d = \min_{1 \leq i \leq n-d} \sum_{i' \in \{i,\ldots,i+d\}} \left(X_{(i')jk} - \frac{1}{d+1} \sum_{i'' \in \{i,\ldots,i+d\}} X_{(i'')jk} \right)^2$$

  where $X_{(1)jk}, \ldots, X_{(n)jk}$ are the order statistics. It is also independent on g.

- The bound may be not very sharp since it is likely verified with far higher probability than $1-\alpha$ in most cases. However it will be convenient for the strategy of use that we describe now.

3.2 Strategy for using the bound

The previous bound can be used to compute the maximum likelihood estimator (MLE) over the following (random) constrained subspace of the parameter space:

$$\Theta(\alpha) = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta, j \in \{1,\ldots,d\}, k \in \{1,\ldots,g\}, \lambda_{jk} \geq B_{jk}^d(\alpha)\}.$$

We have already proved that $\hat{\boldsymbol{\theta}}(\alpha) = \text{argmax}_{\boldsymbol{\theta} \in \Theta(\alpha)} L(\boldsymbol{\theta}; \boldsymbol{X})$ is a consistent estimate of $\boldsymbol{\theta}$ in the univariate case $d = 1$ (Biernacki and Castellan, 2011). Consistency in the general multivariate is expected to hold also but the proof is still in progress.

In practice, this strategy is very simple to implement in an EM algorithm. First, at each iteration $[r]$, compute at the step M the current parameter $\boldsymbol{\theta}^{[r]}$ as usual and compute also all corresponding $\lambda_{jk}^{[r]}$ and $B_{jk}^{d[r]}(\alpha)$ for all j and k. Then, use the associated constraint $\Theta^{[r]}(\alpha)$ to stop or not the EM process at this iteration $[r]$. Stopping an EM run in this manner means that the corresponding EM trajectory would lead to degeneracy and have to be run again from other starting points.

4 Numerical experiments

To access practical efficiency of the previous strategy for avoiding degeneracy, we perform now numerical experiments on simulated data. We consider $g = 2$ Gaussians of dimension $d \in \{1, 2, 4, 8\}$ with same proportions ($\pi_1 = \pi_2 = 1/2$), with centers $\mu_1 = \mathbf{0}$ and $\mu_2 = \mathbf{1}$ and covariance matrices $\Sigma_1 = \Sigma_2 = \mathbf{I}_d$. For each value of d, 1000 samples of size $n = 10d$ are drawn and two versions of the EM algorithm are run with identical starting parameters $\boldsymbol{\theta}^{[0]}$ choosen at random but differing on their stopping rule. The first version corresponds to a classical EM (denoted EM_0 below): It stops either when relative increase of the log-likelihood is smaller than a standard threshold $\varepsilon = 10^{-6}$ ("normal stop") or if the numerical tolerance of the computer is reached when estimating covariance matrices

	EM_0 stop:	crash		normal	
d	EM_α stop:	degeneracy	crash or normal	normal	degeneracy or crash
1		189/189	0/189	811/811	0/811
2		57/57	0/57	943/943	0/943
4		34/34	0/34	966/966	0/966
8		37/37	0/37	963/963	0/963

Table 1: Numerical comparison of EM_0 and EM_α: Counting runs.

("crash stop"; indicating probably degeneracy). The second version of the EM algorithm corresponds to our new strategy (denoted EM_α below): It stops either with a "normal stop" or a "crash stop" (the same "normal stop" and "crash stop" as EM_0), or when our bound on singular matrices is reached with $\alpha = 0.01$ (our so-called "degeneracy stop"). Results displayed on Table 1 show that EM_α prevent *all* crash situations which occur in EM_0 (only true positive cases) while it *never* considers normal runs as degenerate runs (none false positive cases). The table has to be read in the following way: For $d = 1$, EM_α has 189 degeneracy stops among the 189 crash stops of EM_0, *etc.*

Bibliographie

[1] Biernacki, C. and Castellan, G. (2011), A Data-Driven Bound on Variances for Avoiding Degeneracy in Univariate Gaussian Mixtures, *Pub. IRMA Lille*, Vol. 71-IV.

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

[3] Dempster, A. P. and Laird, N. M., Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc. B*, 39, 1–38.

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

[5] Ingrassia, S. and Rocci, R. (2007), Constrained monotone EM algorithms for finite mixture of multivariate Gaussians, *Comput. Stat. Data Anal.*, 51, 5339–5351.

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

[7] McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, Wiley, New York.

[8] Policello, G. E. (1981), Conditional maximum likelihood estimation in Gaussian mixtures, *Stat. Dist. Sci. Work*, 5, 111–125.

[9] Redner, R. and Walker, H. (1984), Mixture densities, maximum likelihood and the EM algorithm, *SIAM Rev.*, 26 (2), 195–239.

[10] Tanaka, K. and Takemura, A. (2006), Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when the scale parameters are exponentially small, *Bernoulli*, 12 (6), 1003–1017.