PUB. IRMA, LILLE 2011
Vol. 71, N° IV

A Data-Driven Bound on Variances for Avoiding Degeneracy in Univariate Gaussian Mixtures

Christophe BIERNACKI[a], Gwénaëlle CASTELLAN[b]

Abstract

In the case of univariate Gaussian mixtures, unbounded likelihood 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 non-asymptotic stochastic lower bound on variances. We prove also that maximizing the likelihood under this data-driven constraint leads to consistent estimates.

Key words and phrases. Univariate Gaussian mixture, maximum likelihood, non-asymptotic bound, consistent estimate.

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 6, 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 [5, 1]. As a consequence, firstly some theoretical questions about the ML properties are raised and, secondly, optimization algorithms like EM [2, 8] may converge, as observed by any practicioner, towards such degenerate solutions.

Avoiding degeneracy is usually handled by constraining the variances. The main option consists to constraint variances 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 [9]. Another way is to impose relative constraints between variances [3, 4]. Alternatively, [7] imposed a constraint on the latent partition underlying the data (instead of a constraint on the variances), what leads to maximize a bounded likelihood and gives consistent estimates. The proposed assumption is weak and natural since it only

[a]University Lille 1 & CNRS & INRIA, Villeneuve d'Ascq, France
[b]University Lille 1 & CNRS , Villeneuve d'Ascq, France

requires that at least two data units arise from each univariate mixture Gaussian component. However, maximizing this likelihood is untractable because of combinatorial difficulties, in particular when more than two components are involved.

Using such a weak assumption on the latent partition, the present work establishes a non-asymptotic stochastic lower bound on the variance of each component. This data-driven lower bound is very simple to calculate from the sample and leads to consistent estimates of the mixture. It can be used by any practitioner without any modification of its preferred ML software (typically 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 variances is obtained and studied in Section 3. The last section (Section 4) concludes by discussing consequences and possible extensions of the present work.

2 Linking latent partition with degeneracy

2.1 Observed-data likelihood and degeneracy

In the univariate Gaussian mixture model assumption, each individual X_i of the data set $\mathbf{X} = (X_1, \ldots, X_n)$ i.i.d. arises from the density

$$f(\cdot; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \phi(\cdot; \mu_k, \sigma_k^2)$$

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 $\phi(.; \mu_k, \sigma_k^2)$ denotes the density of the Gaussian distribution of this kth component with mean μ_k and variance σ_k^2 ($\sigma_k^2 > 0$ for all $k = 1, \ldots, g$). These natural constraints on the mixture parameter $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_g, \mu_1, \ldots, \mu_g, \sigma_1^2, \ldots, \sigma_g^2)$ are summarized in the parameter space Θ.

It is well-known that the *observed-data* likelihood defined by

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

is unbounded from above [5, 1]. For say X_1, if $\mu_1 = X_1$ and σ_k^2 fixed for all $k \in \{2, \ldots, g\}$ then $L(\boldsymbol{\theta}; \mathbf{X}) \to \infty$ as soon as $\sigma_1^2 \to 0$. 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 $\mathbf{X}$ is built from the two following sequential steps:

1. First a partition $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is obtained by n i.d.d. realizations $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{ig})'$ of the multinomial distribution of order one and of parameter $(\pi_1, \ldots, \pi_g)$, $\mathbf{Z}_i$ 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 $\mathbf{Z}_i$, each X_i is independently generated from the Gaussian component indicated by Z_{ik}.

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

$$L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{g} \left[\pi_k \phi(X_i; \mu_k, \sigma_k^2)\right]^{Z_{ik}}$$

could be maximized on $\boldsymbol{\theta}$. However, even in this full data case, degeneracy arises as soon as a given $k \in \{1, \ldots, g\}$ is such that $N_k = 1$, 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 $\mathbf{Z}$ where

$$\mathcal{Z}^* = \{\mathbf{Z} : \ N_k \geq 2, \ k = 1, \ldots, g\}$$

is the set of all partitions containing at least two individuals from each component.

Starting from this remark, [7] proposed to maximize a likelihood taking into account the additional information $\mathcal{Z}^*$ on the latent partition $\mathbf{Z}$. He chooses to maximize the *conditional* likelihood $L(\boldsymbol{\theta}; \mathbf{X}|\mathcal{Z}^*)$ and establishes that it is now bounded and leads to consistent estimates. He gives the detail of a specific EM algorithm for maximizing $L(\boldsymbol{\theta}; \mathbf{X}|\mathcal{Z}^*)$ but it is computational untractable as soon as $g > 2$. Note that the *augmented-data* likelihood $L(\boldsymbol{\theta}; \mathbf{X}, \mathcal{Z}^*)$ would lead exactly to the same problem.

We will overcome this difficulty in the next section by proposing an alternative solution through a cheaper computational lower bound on the variances.

3 A data-driven bound on variances

3.1 Establishing the lower bound

In order to prevent the (traditional) likelihood (1) to degenerate, we propose now alternatively a lower bound (noted $B_n^*(\alpha)$ below) on the variances σ_k^2 of each component $k = 1, \ldots, g$. 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 $\mathbf{Z}$ already discussed in the previous section.

Proposition 1. *For any $\alpha \in (0, 1)$, define the bound*

$$B_n^*(\alpha) = \frac{\min_{1 \leq i < j \leq n} (X_i - X_j)^2}{2\chi^2_{n-2g+1}((1-\alpha)^{1/g})}, \tag{2}$$

where $\chi^2_\lambda(\alpha)$ denotes the quantile of χ^2 with λ degrees of freedom and of order α. Then, we have

$$\mathbb{P}\left(\forall k \in \{1,\dots,g\}\ ,\ \sigma_k^2 > B_n^*(\alpha) \mid \mathcal{Z}^*\right) \geq 1-\alpha.$$

A proof is available in A.

Remarks

- The numerator involved in the bound $B_n^*(\alpha)$ relies on the calculus of the square of the minimum distance between two distinct observations. Note that this corresponds to the minimum of the unbiased empirical variance (up to a factor 2) of the pair (X_i, X_j) since

  $$(X_i - \bar{X}_{i,j})^2 + (X_j - \bar{X}_{i,j})^2 = (X_i - X_j)^2/2, \quad \text{with} \quad \bar{X}_{i,j} = (X_i + X_j)/2.$$

  Moreover, this bound is easy and fast to compute using the following equality:

  $$\min_{1\leq i<j\leq n} (X_i - X_j)^2 = \min_{1\leq i\leq n-1} (X_{(i+1)} - X_{(i)})^2,$$

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

- The lower 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 Using the bound for avoiding degeneracy

Since the bound $B_n^*(\alpha)$ is non asymptotic and data-driven, we can use it to compute the maximum likelihood estimator (MLE) over a (random) constrained subspace of the parameter space. Note that the constrained space allows to avoid the problem of the MLE existence. This strategy leads also to consistent estimates of the mixture parameter as it is claimed in the following proposition (a proof is given in A).

Proposition 2. *Let the following restricted mixture parameter data-driven set*

$$\Theta_n^*(\alpha) = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta, k \in \{1,\dots,g\}, \sigma_k^2 \geq B_n^*(\alpha)\}.$$

Noting

$$\hat{\boldsymbol{\theta}}_n^*(\alpha) = \underset{\boldsymbol{\theta}\in\Theta_n^*(\alpha)}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; \mathbf{X})$$

and $\boldsymbol{\theta}^0$ the true mixture parameter, then $\hat{\boldsymbol{\theta}}_n^*(\alpha) \xrightarrow{\text{pr}} \boldsymbol{\theta}^0$ as $n \to \infty$, up to a label permutation of components.

4 Discussion

In this work, we have used a quite natural and weak information on the latent partition in order to avoid degeneracy in univariate Gaussian mixtures. The proposed approach corresponds to a *non-asymptotic stochastic* lower bound on the variances which is very easy to calculate by any practitioner. In addition, it can be naturally combined with any standard ML optimization procedure like the EM algorithm by simply discarding any run crossing the variance bound $B_n^*(\alpha)$. Thus, this bound could provide an interesting response not only for theorists but also for practitioners.

Two main extensions of this work could be of interest. First, avoiding degeneracy in the *multivariate* Gaussian mixture is planned in our future works. We expect to address it in a proper manner by imposing again constraints on the latent partition. Second, it is also worth noting that our proposal does not solve the difficult problem of spurious solutions of the likelihood [see 6, Sections 3.10 and 3.11]. However, we think that introducing again some information on the latent partition could be an interesting way to explore for tackle this problem since spurious solutions can be seen as a Gaussian "captured" by a very small set of individuals.

A Proofs of propositions

Proof of Proposition 1. For $k = 1, \ldots, g$, we note $V_k(\mathbf{Z})$ the unbiased estimate of σ_k^2 obtained by maximizing the complete-data likelihood $L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})$:

$$V_k(\mathbf{Z}) = \frac{1}{N_k - 1} \sum_{i:\ Z_{ik}=1} (X_i - \bar{X}_k(\mathbf{Z}))^2 \quad \text{where} \quad \bar{X}_k(\mathbf{Z}) = \frac{1}{N_k} \sum_{i:\ Z_{ik}=1} X_i.$$

Conditionally to the random variables $\mathbf{Z}$ and to the event $\mathcal{Z}^*$, the g variables $(N_k - 1)V_k(\mathbf{Z})/\sigma_k^2$ have a Chi-square distribution with $N_k - 1$ degrees of freedom and are independent. Thus we deduce that, for any $\alpha \in (0, 1)$,

$$\mathbb{P}\left(\forall k \in \{1, \ldots, g\} \ , \ \sigma_k^2 \geq \frac{(N_k - 1)V_k(\mathbf{Z})}{\chi^2_{N_k - 1}((1-\alpha)^{1/g})} \Big| \mathbf{Z}, \mathcal{Z}^* \right) = 1 - \alpha.$$

From Lemma 1 (see B) and noticing that $\{i : Z_{ik} = 1\} \subset \{1, \ldots, n\}$, we have $V_k(\mathbf{Z}) \geq \min_{1 \leq i < j \leq n} V_{i,j}$, where $V_{i,j}$ denotes the unbiased empirical variance of (X_i, X_j). Since $V_{i,j} = \frac{1}{2}(X_i - X_j)^2$ implies that

$$V_k(\mathbf{Z}) \geq \frac{1}{2} \min_{1 \leq i < j \leq n} (X_i - X_j)^2,$$

we obtain that (note that now the condition on $\mathbf{Z}$ can be weakened into a condition on $\mathbf{N} = (N_1 \ldots, N_g)$),

$$\mathbb{P}\left(\forall k \in \{1, \ldots, g\} \ , \ \sigma_k^2 \geq \frac{(N_k - 1) \min_{1 \leq i < j \leq n} (X_i - X_j)^2}{2\chi^2_{N_k - 1}((1-\alpha)^{1/g})} \Bigg| \mathbf{N}, \mathcal{Z}^* \right) \geq 1 - \alpha.$$

Moreover, using the fact that

- $\forall \lambda_1, \lambda_2 \in \mathbb{N}^*, \lambda_1 \leq \lambda_2 \Rightarrow \chi^2_{\lambda_1}(\alpha) \leq \chi^2_{\lambda_2}(\alpha)$,
- $N_k \geq 2$ for all $k \in \{1, \ldots, g\}$ (equivalent to the constraint $\mathcal{Z}^*$ on $\mathbf{Z}$) together with $\sum_{k=1}^{g} N_k = n$ implies that $N_k \leq n - 2(g-1)$ for all $k \in \{1, \ldots, g\}$,

the condition on $\mathbf{N}$ vanishes and we deduce that

$$\mathbb{P}\left(\forall k \in \{1, \ldots, g\} \ , \ \sigma_k^2 \geq \frac{\min_{1 \leq i < j \leq n} (X_i - X_j)^2}{2\chi^2_{n-2g+1}((1-\alpha)^{1/g})} \middle| \mathcal{Z}^* \right) \geq 1 - \alpha,$$

thus the conclusion follows.

Proof of Proposition 1. Convergence of estimates of $\boldsymbol{\theta}^0$ are understood below up to a label permutation of components. We will use a result of [9] on the consistency of the maximum likelihood estimator when the variances are constrainted to be lower bounded by some sequence $c_n = \exp(-n^\delta)$ for some $\delta \in (0, 1)$. More precisely, denoting by Θ_n the restricted mixture parameter set,

$$\Theta_n = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta, k \in \{1, \ldots, g\}, \sigma_k^2 > c_n\},$$

and $\hat{\boldsymbol{\theta}}_n$ a MLE restricted to Θ_n,

$$\hat{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta_n} L(\boldsymbol{\theta}; \mathbf{X}),$$

[9] demonstrate the strong consistency of the sequence $\hat{\boldsymbol{\theta}}_n$. Therefore if we prove that

$$\lim_{n \to +\infty} \mathbb{P}(B_n^*(\alpha) \geq c_n) = 1 \quad \text{and} \quad \lim_{n \to +\infty} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_n \in \Theta_n^*(\alpha)\right) = 1,$$

then we will deduce that

$$\lim_{n \to +\infty} \mathbb{P}\left(\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n^*(\alpha)\right) = 1,$$

and this leads to the weak consistency of the sequence of estimators $\hat{\boldsymbol{\theta}}_n^*(\alpha)$, up to a label permutation of components.

First, we prove that $\lim_{n \to +\infty} \mathbb{P}(B_n^*(\alpha) \geq c_n) = 1$. It is the consequence of Lemmas 2 and 3 (see B). Lemma 2 allows to control the random term $\min_{1 \leq i < j \leq n} (X_i - X_j)^2$ and Lemma 3 establishes the equivalence $\chi^2_n(\alpha) \sim n$. Thus, by definition of $B_n^*(\alpha)$ (given in Equation (2)),

$$\mathbb{P}(B_n^*(\alpha) \geq c_n) = \mathbb{P}\left(\min_{1 \leq i < j \leq n} (X_i - X_j)^2 \geq 2c_n \chi^2_{n-2g+1}((1-\alpha)^{1/g}) \right).$$

We then apply Lemma 2 with $c = 2c_n \chi^2_{n-2g+1}((1-\alpha)^{1/g})$ and Lemma 3 allows to conclude that $\lim_{n \to +\infty} n(n-1) \left(2c_n \chi^2_{n-2g+1}((1-\alpha)^{1/g})\right)^{1/2} = 0$.

Secondly, we prove that $\lim_{n\to+\infty}\mathbb{P}\left(\hat{\boldsymbol{\theta}}_n \in \Theta_n^*(\alpha)\right) = 1$. It follows from the two following convergences. On the one hand $\hat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}^0$ almost surely [see 9]. On the other hand, $\lim_{n\to+\infty}\mathbb{P}\left(\boldsymbol{\theta}^0 \in \boldsymbol{\Theta}_n^*(\alpha)\right) = 1$ because the bound $B_n^*(\alpha)$ tends almost surely to 0 when n tends to $+\infty$ (it is clear since $\min_{1\leq i<j\leq n}(X_i - X_j)^2$ is a decreasing positive sequence and $\chi^2_{n-2g+1,(1-\alpha)^{1/g}}$ tends to $+\infty$).

B Lemmas

Lemma 1. *For $I = \{1,\dots,m\}$, we note V_I the unbiased empirical variance of $(X_i)_{i\in I}$ given by*

$$V_I = \frac{1}{m-1}\sum_{i\in I}(X_i - \bar{X}_I)^2 \quad \textit{where} \quad \bar{X}_I = \frac{1}{m}\sum_{i\in I} X_i.$$

Then, for any $k \in \{2,\dots,m\}$ the following equality holds

$$V_I = \frac{1}{\binom{m}{k}}\sum_{J\subset I,\#J=k} V_J,$$

where $\#J$ denotes the cardinal of the set J. Consequently,

$$V_I \geq \min_{J\subset I,\#J=k} V_J.$$

Proof. We develop the mean of the sum of $\{V_J\}_{J\subset I,\#J=k}$:

$$\begin{aligned}
\sum_{J\subset I,\#J=k} V_J &= \sum_{J\subset I,\#J=k}\Big(\frac{1}{k-1}\sum_{j\in J}X_j^2 - \frac{1}{k(k-1)}\sum_{j,j'\in J}X_jX_{j'}\Big)\\
&= \frac{1}{(k-1)}\Big(\sum_{J\subset I,\#J=k}\sum_{j\in J}X_j^2 - \frac{1}{k}\sum_{J\subset I,\#J=k}\sum_{j,j'\in J}X_jX_{j'}\Big)\\
&= \frac{1}{(k-1)}\Big(\sum_{i\in I}X_i^2\#\{J\subset I,\#J=k,i\in J\}\\
&\quad -\frac{1}{k}\sum_{i\in I}X_i^2\#\{J\subset I,\#J=k,i\in J\}\\
&\quad -\frac{1}{k}\sum_{i\neq i'\in I}X_iX_{i'}\#\{J\subset I,\#J=k,i,i'\in J\}\Big)\\
&= \frac{1}{(k-1)}\Big(\frac{k-1}{k}\binom{m-1}{k-1}\sum_{i\in I}X_i^2 - \frac{1}{k}\binom{m-2}{k-2}\sum_{i\neq i'\in I}X_iX_{i'}\Big)\\
&= \binom{m}{k}\Big(\frac{1}{m}\sum_{i\in I}X_i^2 - \frac{1}{m(m-1)}\sum_{i\neq i'\in I}X_iX_{i'}\Big)\\
&= \binom{m}{k}\Big(\frac{1}{m-1}\sum_{i\in I}X_i^2 - \frac{1}{m(m-1)}\sum_{i,i'\in I}X_iX_{i'}\Big) = \binom{m}{k}V_I.
\end{aligned}$$

The second part of this lemma is immediate. □

Lemma 2. *For any $c > 0$,*

$$\mathbb{P}\left(\min_{1\le i<j\le n}(X_i - X_j)^2 \ge c\right) \ge 1 - \frac{n(n-1)}{\sqrt{\sigma_*^2\pi}}\sqrt{c},$$

where $\sigma_*^2 = \min_{1\le k\le g}\sigma_k^2$.

Proof.

$$\begin{aligned}
\mathbb{P}\left(\min_{1\le i<j\le n}(X_i - X_j)^2 \ge c\right) &= \mathbb{P}\left(\bigcap_{1\le i<j\le n}\left\{(X_i - X_j)^2 \ge c\right\}\right)\\
&= 1 - \mathbb{P}\left(\bigcup_{1\le i<j\le n}\left\{(X_i - X_j)^2 \le c\right\}\right)\\
&\ge 1 - \sum_{1\le i<j\le n}\mathbb{P}\left((X_i - X_j)^2 \le c\right).
\end{aligned}$$

And, conditioning with respect to $(\mathbf{Z}_i, \mathbf{Z}_j)$, the law of $X_i - X_j$ is a normal random variable and more precisely, if $Z_{ik} = 1$ and $Z_{jl} = 1$ then $X_i - X_j$ is a normal random variable with mean $\mu_k - \mu_l$ and variance $\sigma_k^2 + \sigma_l^2$. Thus,

$$\begin{aligned}
\mathbb{P}\left((X_i - X_j)^2 \le c\right) &= \mathbb{E}\left[\mathbb{P}\left((X_i - X_j)^2 \le c\right)|\mathbf{Z}_i, \mathbf{Z}_j\right]\\
&= \mathbb{E}\left[\mathbb{P}\left(|X_i - X_j| < \sqrt{c}\right)|\mathbf{Z}_i, \mathbf{Z}_j\right]\\
&= \sum_{1\le k,l\le g}\pi_k\pi_l\nu\left(\left[\frac{-\sqrt{c} - (\mu_k - \mu_l)}{\sqrt{\sigma_k^2 + \sigma_l^2}}, \frac{\sqrt{c} - (\mu_k - \mu_l)}{\sqrt{\sigma_k^2 + \sigma_l^2}}\right]\right),
\end{aligned}$$

where ν denotes the standard normal distribution. And the Mean Value Theorem implies for any real numbers $x \le y$,

$$\nu([x, y]) = \Phi(y) - \Phi(x) \le \frac{1}{\sqrt{2\pi}}(y - x).$$

From which we deduce that

$$\mathbb{P}\left((X_i - X_j)^2 \le c\right) \le \frac{2\sqrt{c}}{\sqrt{4\sigma_*^2\pi}} = \sqrt{\frac{c}{\sigma_*^2\pi}}.$$

□

Lemma 3. *For any $\alpha \in (0, 1)$, the quantile $\chi_n^2(\alpha)$ of χ^2 with n degrees of freedom and of order α is equivalent to n when n tends to $+\infty$.*

Proof. It follows directly from the fact that if X_n is a χ^2 random variable with n degrees of freedom then $X_n/n \xrightarrow{\text{pr}} 1$. Consequently, for all $\varepsilon > 0$ there exists $N \in \mathbb{N}^*$ such that for all $n \ge N$

$$\mathbb{P}(X_n \le (1 - \varepsilon)n) \le \alpha \quad \text{and} \quad \mathbb{P}(X_n \ge (1 + \varepsilon)n) \le 1 - \alpha.$$

The last inequality can be written as $\mathbb{P}(X_n \leq (1+\varepsilon)n) \geq \alpha$, and this leads to

$$(1-\varepsilon)n \leq \chi^2_n(\alpha) \leq (1+\varepsilon)n.$$

□

References

References

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

[2] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

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

[4] Ingrassia, S., Rocci, R., 2007. Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. Computational Statistics & Data Analysis 51, 5339–5351.

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

[6] McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

[7] Policello, G. E., 1981. Conditional maximum likelihood estimation in Gaussian mixtures. Statistical Distributions in Scientific Work 5, 111–125.

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

[9] Tanaka, K., 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.