An EM Stopping Rule for Avoiding Degeneracy in Gaussian-based Clustering with Missing Data

Christophe Biernacki, Vandewalle Vincent

An EM Stopping Rule for Avoiding Degeneracy in Gaussian-based Clustering with Missing Data

Christophe Biernacki[1] and Vincent Vandewalle[2]

[1] Inria, Univ. Lille, CNRS, UMR 8524 - Lab. P. Painlevé, F-59000 Lille,
christophe.biernacki@inria.fr,
[2] Univ. Côte d'Azur, Inria, CNRS, UMR 7351 - Lab. J.A. Dieudonné, F-06108 Nice,
vincent.vandewalle@inria.fr

Abstract. Missing data frequency increases with the growing size of multivariate modern datasets. In Gaussian model-based clustering, the EM algorithm easily takes into account such data but the degeneracy problem is dramatically aggravated during the EM runs: parameter degeneracy is quite slow and also more frequent than with complete data. Consequently, parameter degenerated solutions may be confused with valuable parameter solutions and, in addition, computing time may be wasted through wrong runs. In this work, a simple and low informational condition on the latent partition allows to propose a very simple partition-based stopping rule of EM which shows good behavior on numerical experiments.

Keywords: clustering, degeneracy, EM algorithm, Gaussian mixtures, latent partition, missing data.

1 Introduction

Gaussian mixtures models easily fit many practical settings, like the clustering paradigm. Various estimation approaches are available, but the maximum likelihood (ML) approach is usually much preferred. Nevertheless, it is well-known that the likelihood function of heteroscedastic 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] may converge, as observed by any practitioner, towards such degenerate solutions. When it encounters a degenerated solution the likelihood goes to infinity, which is a symptom that the parameters are close of the border. These degenerated solutions on the border of the parameters space are obviously out of interest.

The solutions to avoid the problem of degeneracy are either to modify the estimator of the parameters or to detect it through the dynamic of the EM algorithm. Avoiding degeneracy is usually handled by some *parameter-based constraints*, typically by constraining the singular values of the covariance matrices; see for instance [8, 3]. Alternatively, two very different approaches have been

proposed. A very recent one is proposed by [7] through the so-called (bounded) marginal likelihood consisting to estimating first the variances (and the mixture proportions) while integrating over all positioning information, and then to estimate the centers conditionally to the previous estimates. Despite its conceptual interest, this approach produces some estimates loosing some Fisherian information (is hardly implemented in the multivariate Gaussian case and for more than two components). A second alternative approach has been proposed three decades ago by [6] and could be qualified as a *partition-based constraint* approach. The author imposes a constraint on the latent partition underlying the data, that leads to maximize a (bounded) conditional 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, very few results are available for the degeneracy of Gaussian mixtures with some missing data while, with the increasing of the number of available variables, the risk that data contain missing values also increases. In this article we study the degeneracy properties in the missing data framework. Some unexpected results will be revealed, underlying the necessity to avoid degeneracy with most priority than in the complete data case itself. It will lead to design a interesting solution consisting of an absolutely standard EM algorithm combined with an easy-to-compute partition-based stopping rule.

The outline of this paper is the following. In Section 2, we characterize the EM algorithm dynamics in the missing data case within the Gaussian framework. In Section 3, we present the partition-based constraint solution for stopping the EM algorithm when engaged along a possible degeneracy run. Section 4 illustrates through empirical studies the efficiency of our proposal. The last section concludes this work.

2 Degeneracy of the EM algorithm with missing data

2.1 The EM algorithm for missing data

Missing data Let consider a sample $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ coming from a mixture of K Gaussian components in $\mathcal{R}^d$. Let denote by $O_i \subseteq \{1, \ldots, d\}$ the set of observed variables for the individual i ($i \in \{1, \ldots, n\}$) and by M_i the complementary set for the missing variables. Let z_{ik} denote the class of i: $z_{ik} = 1$ if $\mathbf{x}_i$ comes from class k ($k \in \{1, \ldots, K\}$) and 0 otherwise. The whole partition is denoted $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \mathcal{Z}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$. Let $\mathbf{x}_i^o$ denote the observed variables for unit i, and $\mathbf{x}^o$ the observed data set.

The mixture model From the model point of view, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the proportions of each component, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean and covariance matrix of the class k. Let $\boldsymbol{\theta} \in \Theta$ be the global parameter of the mixture, $\boldsymbol{\pi}$ included. Finally $\phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian density of expectation $\boldsymbol{\mu}_k$ and covariance

matrix $\mathbf{\Sigma}_k$. μ_{ik}^o is the sub-vector of $\boldsymbol{\mu}_k$ associated to the index O_i (*idem* for M_i). $\mathbf{\Sigma}_{ik}^{om}$ is the sub-matrix of $\mathbf{\Sigma}_k$ associated to the rows O_i and columns M_i (and the same for any combination of indexes o and m, like $\mathbf{\Sigma}_{ik}^{oo}$, *etc.*).

The related EM algorithm Starting from an initial parameter, the EM algorithm allowing to take into account missing data is described below, where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^+$ denote respectively parameters at two successive iterations (a similar convention is adopted for the missing data $\mathbf{x}^m$ and $\mathbf{z}$). Note also that this algorithm is stopped typically either after a predefined number of iterations, or when a predefined threshold in the increase of the log-likelihood is reached.

E step Missing data estimation through the computation of $t_{ik}^+ \propto \pi_k \phi(\mathbf{x}_i^o; \boldsymbol{\lambda}_k)$ and $\mathbf{x}_{ik}^{m^+} = \boldsymbol{\mu}_{ik}^m + \mathbf{\Sigma}_{ik}^{mo} \left(\mathbf{\Sigma}_{ik}^{oo}\right)^{-1} (\mathbf{x}_i^o - \boldsymbol{\mu}_{ik}^o)$. Here $\mathbf{x}_{ik}^{m^+}$ can be interpreted as an imputation of missing data given the class and the observed variables.

M step Parameter estimation through the computation of $\pi_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+$, $\boldsymbol{\mu}_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+ \mathbf{x}_{ik}^+$ and $\mathbf{\Sigma}_k^+ = \frac{1}{n_k^+} \sum_{i=1}^n t_{ik}^+ \left[(\mathbf{x}_{ik}^+ - \boldsymbol{\mu}_k^+)(\mathbf{x}_{ik}^+ - \boldsymbol{\mu}_k^+)' + \mathbf{\Sigma}_{ik}^+ \right]$ where $n_k^+ = \sum_{i=1}^n t_{ik}^+$, $\mathbf{x}_{ik}^+ = \begin{pmatrix} \mathbf{x}_i^o \\ \mathbf{x}_{ik}^{m^+} \end{pmatrix}$ and $\mathbf{\Sigma}_{ik}^+ = \begin{pmatrix} \mathbf{0}_i^o & \mathbf{0}_i^{om} \\ \mathbf{0}_i^{mo} & \mathbf{\Sigma}_{ik}^{m+} \end{pmatrix}$ with $\mathbf{0}$ being the null matrix $d \times d$ and $\mathbf{\Sigma}_{ik}^{m^+} = \mathbf{\Sigma}_{ik}^{mm} - \mathbf{\Sigma}_{ik}^{mo} \left(\mathbf{\Sigma}_{ik}^{oo}\right)^{-1} \mathbf{\Sigma}_{ik}^{om}$. $\mathbf{\Sigma}_{ik}^{m^+}$ can be interpreted as a variance fitting in order to take into account the under estimation of the variance caused by the missing data imputation.

Unbounded likelihood Let $\ell(\boldsymbol{\theta}; \mathbf{x}^o) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k \phi(\mathbf{x}_i^o; \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) \right)$ the log-likelihood. Classically, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is defined by $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{x}^o)$. However, if $\mathbf{\Sigma}_k$ is free (heteroscedastic case) then $\ell(\boldsymbol{\theta}; \mathbf{x}^o)$ is unbounded for instance by taking a center such that $\boldsymbol{\mu}_k = \boldsymbol{x}_i$ and the corresponding *generalized variance* such that $|\mathbf{\Sigma}_k| \to 0$ [5].

2.2 Unexpected EM algorithm behaviour in case of missing data

Illustration of a specific EM dynamics on real data EM dynamics is well-identified in the *complete individual data* case [4]. Typically, when an EM iteration is close enough to a degenerate situation for a given component, then the corresponding generalized variance evolves exponentially fast towards zero along the subsequent iterations of the algorithm. However, a very different behavior has been identified in the case of missing data as we illustrate now on the real data set breast cancer tissue of the UCI database repository[3]. It is composed with $n = 106$ statistical units and $d = 9$ features. We have artificially hidden 10% of the data completely at random. Then we have adjusted a mixture model with $K = 4$ clusters. Then, the evolution of the log-likelihood for a degenerated solution is given in Figure 1 (left panel). We can see that the growth of

[3] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

the log-likelihood seems to be linear. If we look now at the evolution of the logarithm of the generalized variance of the degenerating component Figure 1 (right panel) we also see that it seems to decrease linearly as the number of iteration increases. Such a log-likelihood and a log-generalized variance behavior indicate clearly that the EM dynamics in the missing individual data case is radically different from this one observed in the complete individual data case.

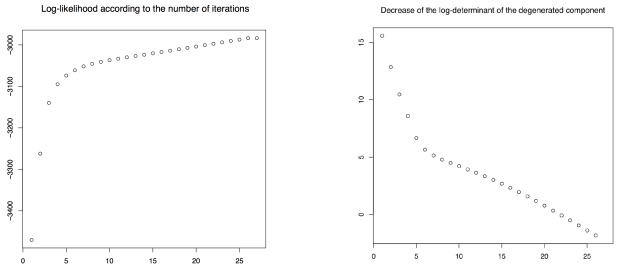

Fig. 1: Illustration of the EM degeneracy dynamics with missing data, according to the number of iterations: (left) Evolution of the likelihood; (right) Evolution of the logarithm of the generalized variance for the degenerating component.

Explaining the observed degeneracy speed through a toy example In order to guess the dynamic of the EM algorithm, let start with a Gaussian univariate framework ($d = 1$) without mixture ($K = 1$) and with only one datum that we will denote simply by x. In this framework the estimator of the mean μ is equal to x and the estimator of the variance Σ is equal to 0 which leads to an infinite likelihood. Let imagine that in addition to this unique observed datum, $n-1$ data have not been observed (they are missing). Then, it is possible to perform an EM algorithm (useless here) which will converge towards the expected degenerated solution. However, in this oversimplified instance, it is possible to explicitly express the algorithm pathway, since updating formulas at the M step become $\mu^+ = n^{-1}((n-1)\mu + x)$ and $\Sigma^+ = n^{-1}((n-1)\Sigma + (x-\mu^+)^2)$. After simple algebra, we can then express at iteration $q \geq 0$ both the convergence speed of the log-likelihood ℓ and of the variance Σ:

$$\ell(\boldsymbol{\theta}^{(q)}; x) \sim -0.5q \ln \frac{n-1}{n} \quad \text{and} \quad \Sigma^{(q)} \sim \Sigma^{(0)} \ln \left(\frac{n-1}{n}\right)^q,$$

where $\boldsymbol{\theta}^{(q)}$ denotes the parameter $\boldsymbol{\theta}$ at iteration q. We thus retrieve in this simple example, the previous experimental behavior of both the log-likelihood and of the variance along a degeneracy run. Let also remark that the degeneracy speed decreases as the rate of missing data increases. It can thus be expected that the EM algorithm can be trapped by degenerated solutions with a still slower dynamics, as we illustrate now.

Additional behaviour of EM: Influence of the missing data rate Let take again the instance on the breast cancer tissue of the UCI database repository, and now we vary the rate of missing data. From results presented Table 1 we can see that as the rate of missing data increases, the rate of degeneracy increases, and the number of iterations before degeneracy decreases.

Table 1: Frequency and speed of degeneracy (deg.) according to the rate of missing data on the breast cancer data set.

% missing data	0	5	10	15	20	25	30
% deg.	16	4	12	11	46	51	100
Average nb of iterations before deg.	2	13	13	82	304	138	215

3 Minimal partition information for avoiding degeneracy

Usual solutions to avoid degeneracy have been already briefly presented in Section 1. Here we focus on the solution consisting of introducing a constraint on the latent partition within the estimation process, arguing that it represents the method with the weakest information required within a model-based clustering context. We present this method below, which has also the advantage to be easily applicable to the missing data case, but by introducing as a novelty a specific adaptation which avoids its associated combinatorics limitations.

3.1 The partition information and the associated EM* algorithm

The core idea is here to transpose some natural constraints from the supervised setting to the current unsupervised setting, *i.e.* enough data should be available in each class to perform the parameter estimation (an obvious requirement to obtain non-singular covariance matrices). It is equivalent to constrain the partition space $\mathcal{Z}$. This idea was first introduced by [6] and we detail it below.

This strategy being originally introduced for the complete data case, let $n_k = \sum_{i=1}^{n} z_{ik}$ denoting the number of individuals in component k and let define $\mathcal{Z}^* = \{\mathbf{z} : \forall k, n_k \geq d+1\}$ the set of partitions with at least $d+1$ elements by class. In that case, [6] proposed to maximize the *conditional* log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{x}|\mathcal{Z}^*)$, which can be optimized with a specific EM algorithm and which leads also to consistent mixture estimates. At this step, the key point is to notice that $L(\boldsymbol{\theta}; \mathbf{x}|\mathcal{Z}^*) < \infty$ with probability one. In the following, we however pursue with the *joint* likelihood $L(\boldsymbol{\theta}; \mathbf{x}, \mathcal{Z}^*) = L(\boldsymbol{\theta}; \mathbf{x}|\mathcal{Z}^*)\mathrm{p}(\mathcal{Z}^*)$ since it will be more tractable from the computational point of view in the M step of EM (see further) than its conditional version and both joint and conditional likelihoods are extremely

similar ($L(\boldsymbol{\theta};\mathbf{x},\mathcal{Z}^*) \simeq L(\boldsymbol{\theta};\mathbf{x}|\mathcal{Z}^*)$) because $\mathrm{p}(\mathcal{Z}^*) \simeq 1$ even for moderate values of n.

In the missing data case, we have obviously to slightly adapt the definition of the so-called partition information $\mathcal{Z}^*$. Now $\mathcal{Z}^*$ corresponds to a minimum of $d+1$ *complete* individual data in each component, thus namely we define now $\mathcal{Z}^* = \{\mathbf{z} : \forall k, n_k^o \geq d+1\}$, with $n_k^o = \sum_{i:O_i=\{1,\ldots,d\}} z_{ik}$ designating the number of complete individuals in component k. This adaptation is motivated by the fact that the observed likelihood including the partition information $\mathcal{Z}^*$, namely $L(\boldsymbol{\theta};\mathbf{x}^o,\mathcal{Z}^*)$, is still bounded above with probability one (thus as in the non-missing data case where such a property was straightforward). We make this property clear through the following proposition[4].

Proposition *For* $\mathcal{Z}^* = \{\mathbf{z} : \forall k, n_k^o \geq d+1\}$ *then* $p(L(\boldsymbol{\theta};\mathbf{x}^o,\mathcal{Z}^*) < \infty) = 1$.

We adapt now EM for optimizing this particular likelihood, leading to a specific EM algorithm, called here EM*. This latter is expressed as follows:

E step $t_{ik}^{*+} \propto \tau_{ik} t_{ik}^+$ with $\tau_{ik} = \mathrm{p}(\mathcal{Z}^*|\mathbf{x}^o, z_{ik}=1;\boldsymbol{\theta})$ and $t_{ik}^+ = \mathrm{p}(z_{ik}=1|\mathbf{x}^o;\boldsymbol{\theta})$.
M step Standard formulas of EM (see Section 2.1) where t_{ik}^+ is replaced by t_{ik}^{*+}.

Thus, comparing to the E step of the classical EM, we can observe that a kind of regularization of the conditional probabilities z_{ik}^+ is performed. However, the E step of EM* implies combinatorics within the computation of τ_{ik} as soon as $K > 2$ which greatly limits its practical use (both for the complete and the missing data cases). The following proposed strategy overcomes this problem.

3.2 A simpler solution: an early EM stopping rule relying on $\mathcal{Z}^*$

The term τ_{ik} is the pivotal quantity involved in our strategy. We start by observing two key points in EM*. Firstly, the combinatorial problem only comes from the computation of the term τ_{ik}. Secondly, its value is essentially such that $\tau_{ik} \simeq 1$ (for all i and k values) along most of the iterations of any EM* run, except in a limited number of iterations where the run approaches a degenerate solution. Consequently, most of EM* iterations are greatly equivalent to classical EM iterations (remind that EM and EM* are identical when $\tau_{ik} = 1$), meaning that computing the τ_{ik} values at each iteration of EM* is mostly usefulness. On the contrary, when approaching closely a degenerate solution, we must observe $\tau_{ik} \simeq 0$ for at least one value of the couple (i,k), meaning the corresponding iteration of EM and EM* diverge, except if simultaneously the value of t_{ik}^+ is such that $t_{ik}^+ \simeq 0$.

From this core remark, we formulate a first strategy consisting to replace the calculus of each τ_{ik}, a generally unfeasible task, by a sampling of a partition value $\mathbf{z}_{|ik}$ drawn from $\mathrm{p}(\mathbf{z}|\mathbf{x}^o, z_{ik}=1;\boldsymbol{\theta})$, which is on the contrary an excessively simple task. Then we simply check if the generated partition value $\mathbf{z}_{|ik}$ belongs or not to $\mathcal{Z}^*$. Indeed, if $\tau_{ik} \simeq 0$, then it is expected to draw $\mathbf{z}_{|ik} \notin \mathcal{Z}^*$, whereas if $\tau_{ik} \simeq 1$, then it is expected to draw $\mathbf{z}_{|ik} \in \mathcal{Z}^*$. Then, in practice, a run of the EM

[4]The proof is not given in this too short paper version.

algorithm is stopped as soon as $\mathbf{z}_{|ik} \notin \mathcal{Z}^*$ for at least one couple (i,k), where $\mathbf{z}_{|ik}$ is drawn from $\mathrm{p}(\mathbf{z}|\mathbf{x}^o, z_{ik} = 1; \boldsymbol{\theta})$. Thus, this strategy acts as a partition stopping rule and produces a complete restart of EM from a new run.

However, this first strategy is expected to be too stringent since it could stop some runs where $z_{ik} \simeq 0$ for some (i,k) values, situation where $\tau_{ik} t_{ik}^+ \simeq t_{ik}^+$ even if $\tau_{ik} \simeq 0$. In other words, we could stop some EM runs which are finally equivalent to EM* runs, whereas we target to stop EM runs only in situations where there are different from the EM* ones. It thus leads to the second (and finally retained) strategy that we describe now and which should be preferred to avoid stopping valid EM runs. The principle is to consider the product $\tau_{ik} t_{ik}^+$ as a whole, instead of τ_{ik} individually. In that way, the strategy consists to firstly draw a value z_{ik} from t_{ik}^+, for each couple (i,k) (thus a full partition $\mathbf{z}$ from $\mathrm{p}(\mathbf{z}|\mathbf{x}^o; \boldsymbol{\theta})$). Then, only for each couple (i,k) s.t. $z_{ik} = 1$, perform the 1st strategy previously described.

Notice that, taken in its globality, this 2nd strategy is equivalent to the very simple procedure consisting finally to draw a $\mathbf{z}$ value from $\mathrm{p}(\mathbf{z}|\mathbf{x}^o; \boldsymbol{\theta})$ and then to check if $\mathbf{z} \in \mathcal{Z}^*$ (in that way EM is not stopped at this current iteration) or $\mathbf{z} \notin \mathcal{Z}^*$ (in that way EM is immediately stopped at this iteration and re-run from another starting value).

4 Numerical experiments

The proposed strategy is below compared to EM* (used as the reference algorithm and tractable for $K = 2$) with respect to the adjusted rand index (ARI) values between the obtained partition and the simulated partition.

Let consider 100 data sets generated from a 9-variate ($d = 9$) Gaussian mixture of two components. Classes have the same proportions ($\pi_1 = \pi_2 = 0.5$), the covariance matrices are the identity and the class centers are $\mu_1 = (0, 0, ..., 0)'$ and $\mu_2 = (6/\sqrt{d}, 6/\sqrt{d}, ..., 6/\sqrt{d})'$. The number of data n is equal to $n = 150$, and the probability for a data-cell to be missing equals to 0.2. This setting allows having the same Mahalanobis distance between cluster whatever the value. Moreover, the high separation between clusters implies that poor ARI solutions are mainly caused by a poor optimum of the likelihood.

Then, for each given dataset each algorithm is initialized ??? times, with the same initialization for all the algorithms. préciser nb it We then compare the ARI resulting from EM and EM*. For each data set, results are split considering three possible cases: first the stopping rule is not activated (thus in this case degeneracy cannot occur), second the stopping rule has been activated, but no degeneracy has been observed, third the stopping rule has been activated and a degeneracy due to numerical crash has been reported. Results are presented in Table 2. We observe that when the stopping rule is not activated, both EM and EM* give very close ARI values. In addition, when the stopping rule is activated, it appears to be justified either due to a degenerate run, or due to a poor partitioning output value run. Notice these latter runs should possibly correspond to a degenerate run, but it is difficult to access to this information since we know that degeneracy could be quite very slow in the missing data case.

Table 2: Average ARI for EM and EM* (the frequency of each case is also given).

Stopping rule (frequency)	EM	EM*
Not activated (94)	0.909	0.915
Activated without EM degeneracy (5)	0.074	0.340
Activated with EM degeneracy (1)	-	0.604

5 Concluding remarks

We have identified unexpected, and unpleasant, behaviour of the EM algorithm in the case of missing data, potentially leading to poor clustering results and a waste of computing time. We have proposed an easy-to-implement early stopping rule directly on a *classical* EM algorithm, and relying on a minimal partitioning information. The first promising numerical experiments we obtained have now to be confirmed by future numerical experiments.

References

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

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data (with discussion)". In: *Journal of the Royal Statistical Society, Series B* 39 (1977), pp. 1–38.

[3] Salvatore Ingrassia and Roberto Rocci. "Constrained monotone EM algorithms for finite mixture of multivariate Gaussians". In: *Computational Statistics & Data Analysis* 51.11 (2007), pp. 5339–5351.

[4] Salvatore Ingrassia and Roberto Rocci. "Degeneracy of the EM algorithm for the MLE of multivariate Gaussian mixtures and dynamic constraints". In: *Computational statistics & data analysis* 55.4 (2011), pp. 1715–1725.

[5] Jack Kiefer and Jacob Wolfowitz. "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters". In: *The Annals of Mathematical Statistics* (1956), pp. 887–906.

[6] George E Policello II. "Conditional maximum likelihood estimation in Gaussian mixtures". In: *Statistical Distributions in Scientific Work*. Springer, 1981, pp. 111–125.

[7] Monia Ranalli, Bruce Lindsay, and David Hunter. "A classical invariance approach to the normal mixture problem". In: *Statistica Sinica* (Jan. 2020). DOI: 10.5705/ss.202016.0483.

[8] Kentaro Tanaka and Akimichi Takemura. "Strong Consistency of the Maximum Likelihood Estimator for Finite Mixtures of Location: Scale Distributions When the Scale Parameters Are Exponentially Small". In: *Bernoulli* (2006), pp. 1003–1017.