INFORMATICA, 2005, Vol. 16, No. 1, 61–74

Multivariate Data Clustering for the Gaussian Mixture Model

Mindaugas KAVALIAUSKAS
Faculty of Fundamental Science, Kaunas University of Technology
Donelaičio 72, LT-3000 Kaunas, Lithuania
e-mail: snaiperiui@takas.lt

Rimantas RUDZKIS
Department of Applied Statistics, Institute of Mathematics and Informatics
Akademijos 4, 08663 Vilnius, Lithuania
e-mail: rudzkis@ktl.mii.lt

Received: June 2004

Abstract. This paper discusses a soft sample clustering problem for multivariate independent random data satisfying the mixture model of the Gaussian distribution. The theory recommends to estimate the parameters of model by the maximum likelihood method and to use "plug-in" approach for data clustering. Unfortunately, the calculation problem of the maximum likelihood estimate is not completely solved in multivariate case. This work proposes a new constructive a few stage procedure to solve this task. This procedure includes statistical distribution analysis of a large number of the univariate projections of observations, geometric clustering of a multivariate sample and application of EM algorithm. The results of the accuracy analysis of the proposed methods is made by means of Monte-Carlo simulation.

Key words: clustering, multivariate data, Gaussian mixture model, projection-based clustering, EM algorithm.

1. Introduction

This paper analyses a sample clustering problem in the case where distribution of the observed random vector satisfies a Gaussian mixture model. The clustering problem is closely connected with a mixture model identification problem. It is natural to use the MLE (*maximum likelihood estimator*) for calculation of parameters. Unfortunately, finding of the maximum likelihood estimate is complicated task in the multivariate case, because the dimension of parameter is large. Traditionally the EM (*expectation maximisation*) algorithm is used to accomplish this task. But this algorithm converges locally and precise enough initial value of parameter should be used to ensure convergence to the MLE, see (Boyles, 1983). There is no widely accepted good method for initializing the parameters, see (Vlassis and Likas, 2000), so the calculation of initial parameter value is the main problem in the clustering of the multivariate Gaussian mixture. Usually following initialization methods are used: random multistart method, k-means or some heuristic

method and hierarchical clustering start. Some procedures combines EM algorithm with incremental adding components to the mixture.

In spite of a popularity of clustering methods, most clustering software are based on the datamining approach or heuristic procedures. Only a few probabilistic model-based clustering packages designed for the Gaussian mixture model data are available: NORMIX, EMMIX and MCLUST. NORMIX uses a random start method, EMMIX uses a random start or k-means start from random centres and MCLUST package uses hierarchical model-based clustering for initialization of the EM algorithm, see (Wolfe, 1967), (McLachlan et al, 1999) and (Fraley and Raftery, 1999).

The authors of this paper propose the new probabilistic clustering algorithm for multivariate data satisfying a Gaussian mixture model. This constructive procedure is several-stage that embraces various ideas of data projection, geometric clustering and application of the EM algorithm. The large number of projection directions is used to avoid loss of data clustering structure.

Let us introduce some notation. Let $O(1), ..., O(n)$ be the observed objects. Clustering is partitioning of these objects into several homogeneous groups, in a certain sense. In the case of geometric clustering sample data are matrices $[\rho_{i,j}]_{i,j=1,...,n}$ of distances (or pseudo-distances) between the observed objects $O(i)$ and $O(j)$. By denoting $N \stackrel{def}{=} \{1, ..., n\}$, we will call the non-intersecting subsets $K_1, ..., K_q$, $\bigcup\limits_{j=1}^{q} K_j = N$ as clusters.

One of the most popular geometric clustering methods is minimization of the sum of mean distances inside the clusters:

$$\sum_{j=1}^{q} \frac{1}{\|K_j\|} \sum_{s,r \in K_j} \rho_{s,r} \longrightarrow \min . \tag{1}$$

Here and later we denote the number of elements of set A by $\|A\|$.

The probabilistic model is applied under the assumption that the observed objects belong to some aggregate that consists of q non-intersecting classes. If a d-dimensional features vector X corresponds to every object and ν denotes the number of a class to which the object belongs, then the vector X and number ν are random. When features $X(k)$ of the observed objects $O(k)$, $k = \overline{1,n}$ are known and the class numbers $\nu(k)$ are unknown, we have a clustering problem. We'll restrict to the case, where objects are selected independently. So, we have a sample $\mathbb{X} = \{X(1), ..., X(n)\}$ consisting of an independent copies of the random column-vector X. In the case of hard clustering, we have to estimate the unknown values $\nu(k) \in \{1, ..., q\}$, i.e., to obtain $\widehat{\nu}(k) = \widehat{\nu}(k, \mathbb{X})$, $k = \overline{1,n}$. In case of soft clustering, the a posteriori probabilities $\pi(j, X(k))$, $j = \overline{1,q}$, $k = \overline{1,n}$ are estimated. These probabilities are defined by the equality

$$\pi(j, x) = \mathbb{P}\{\nu = j | X = x\}. \tag{2}$$

Let Φ_j denote a conditional distribution of X given $\nu = j$. If Φ_j is a Gaussian distribution with density φ_j, $j = \overline{1,q}$, then the distribution density f of vector X satisfies the equality

$$f(x) = \sum_{j=1}^{q} p_j \varphi_j(x) \stackrel{def}{=} f(x,\theta), \quad x \in \mathbb{R}^d. \tag{3}$$

Here p_j are a priori probabilities of the observed object belonging to classes K_j, i.e., $p_j = \mathbb{P}\{\nu = j\}$ and $\theta = (p_j, M_j, R_j,\ j = \overline{1,q})$ is the parameters vector of the multivariate model, where M_j and R_j denotes a mean and the covariance matrix corresponding to the normal density φ_j.

It follows from (3) and the Law of total probability that for every $j = 1, ..., q$, $x \in \mathbb{R}^d$

$$\mathbb{P}\{\nu = j | X = x\} = \frac{p_j \varphi_j(x)}{f(x,\theta)} \stackrel{def}{=} \pi_\theta(j,x). \tag{4}$$

In this case the "plug-in" approach could be applied in clustering:

$$\widehat{\pi}(\cdot) = \pi_{\widehat{\theta}}(\cdot), \tag{5}$$

where $\widehat{\theta}$ is a statistical estimate of multivariate parameter θ. It is natural to use the maximum likelihood approach to estimate this parameter

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \sum_{k=1}^{n} \log f(X(k),\theta). \tag{6}$$

But if the data dimension d is high, the calculation of estimate (6) is a complicated task in practice. Usually a recurrent EM algorithm is used to solve this problem. This algorithm recalculates estimates $\widehat{\pi}$ and $\widehat{\theta}$ using expression (5) and the formulas

$$\begin{aligned} \widehat{p_j} &= \sum_{x \in \mathbb{X}} \frac{\widehat{\pi}(j,x)}{n}, \\ \widehat{M_j} &= \sum_{x \in \mathbb{X}} \frac{\widehat{\pi}(j,x)}{\widehat{p_j} n} x, \\ \widehat{R_j} &= \sum_{x \in \mathbb{X}} \frac{\widehat{\pi}(j,x)}{\widehat{p_j} n} xx' - \widehat{M_j}\widehat{M_j}'. \end{aligned} \tag{7}$$

But the EM algorithm estimator converges to statistic (6) only if the initial value of estimate θ is close to $\widehat{\theta}_{ML}$. So, we have to get a precise enough value of initial θ or $\pi(\cdot)$ using some other method. This problem is partly solved in the univariate case. For example, in (Rudzkis and Radavicius, 1995) the recurrent components allocation procedure is proposed. This procedure combines the EM algorithm and non-parametric estimation of density f and it is effective enough. The usage of this procedure in a multivariate case

is rather complicated, especially if the dimension of data is high. Dimension can be reduced by using the projection pursuit technique, see (Rudzkis and Radavicius, 1999) and (Rudzkis and Radavicius, 2003), however a part of information on the data clusters is lost after passing on to lower dimensional space in general case. If we do not restrict us to one projection, but will project a lot of times to various subspaces of lower dimension, the loss of information would be avoided. We are going to discuss a multivariate clustering method, based on the analysis of univariate data projections – a kind of tomography approach.

2. Employing the Inversion Formula

The probability function f can be obtained from the characteristic function using the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int\limits_{\mathbb{R}^d} e^{-it'x} f^*(t)\,\mathrm{d}t, \tag{8}$$

where $f^*(t) = \mathbb{E}e^{it'X}$ denotes the characteristic function of random vector X.

For each $\tau \in \mathbb{R}^d$ the scalar product $\tau'X$ is a univariate random value. We denote the distribution density of this value by f_τ. Let define f^* and f_τ^* as the characteristic functions of f and f_τ respectively. Since there is a one-to-one correspondence between densities and characteristic functions and

$$f^*(t) = \mathbb{E}e^{it'X} = \mathbb{E}e^{i|t|\tau'X} = f_\tau^*(|t|), \quad \text{where } \tau = t/|t|, \tag{9}$$

the distribution density f is uniquely defined by set of densities $\{f_\tau, \tau \in B\}$ of univariate projections of the random vector X, where B denotes the unit sphere in space $\mathbb{R}^d$. Making use of the inversion formula (8) and replacing variables by a spherical coordinate system, we obtain an equality

$$f(x) = \frac{1}{(2\pi)^d} \int\limits_B ds \int\limits_0^\infty e^{-iu\tau'x} f_\tau^*(u) u^{d-1}\,\mathrm{d}u, \tag{10}$$

where $\int\limits_B \mathrm{d}s$ denotes the surface integral over $\tau \in B$. >From the parametric estimation of univariate projected densities f_τ, we also obtain the estimates of characteristic functions f_τ^*. Putting these estimates on the right side of expression (10) and replacing the surface integral by a sum we obtain the formula for calculating estimate $\widehat{f}$

$$\widehat{f}(x) = \frac{c(d)}{|T|} \sum_{\tau \in T} \int\limits_0^\infty e^{-iu\tau'x} \widehat{f}_\tau^*(u) u^{d-1}\,\mathrm{d}u. \tag{11}$$

Here T is a finite set of points, uniformly distributed on the sphere B, $c(d) = \frac{\pi^{\frac{d}{2}}}{d\,\Gamma\left(\frac{d}{2}+1\right)}$, where Γ denotes the Gamma function. >From equality (3) it follows, that densities f_τ are mixtures of the univariate Gaussian densities. Let $\varphi_{j,\tau}$ denote the density of the univariate Gaussian distribution $N(m_j(\tau), \sigma_j^2(\tau))$, where $m_j(\tau) = \tau' M_j$, $\sigma_j^2(\tau) = \tau' R_j \tau$. Then

$$f_\tau(u) = \sum_{j=1}^{q} p_j \varphi_{j,\tau}(u) \stackrel{def}{=} f_\tau\big(u, \theta(\tau)\big), \tag{12}$$

where $\theta(\tau) = (p_j, m_j(\tau), \sigma_j^2(\tau),\ j = 1, ..., q)$. After estimating $\theta(\tau)$, $\tau \in T$ by some univariate Gaussian mixture analysis method (e.g., the method proposed in (Rudzkis and Radavicius, 1995)), from (12), and replacing the distribution functions by characteristic functions, we obtain estimates $\widehat{f}_\tau^*$ and it remains only to make use of (11). Unfortunately, the estimate of distribution density $f(x)$ calculated this way is non-parametric and it can be used in data clustering only indirectly, together with other methods. We need to have estimates not only of the density $f(x)$, but also of its components φ_j for data clustering. The components can be also defined by expression (11) by replacing $\widehat{f}$ and $\widehat{f}_\tau^*$ with $\widehat{\varphi}_j$ and $\widehat{\varphi}_{j,\tau}^*$ respectively, however, the numbering compatibility problem of components $\varphi_{j,\tau}$ among various projection directions $\tau \in T$ can arise in practice. The mentioned procedure of consecutive component isolation (see (Rudzkis and Radavicius, 1995)) applied in the univariate projection $\tau' X$ analysis assigns the first number to the component whose mean is closest to the mode of f_τ. In the general case, while projecting in various directions, Gaussian distributions corresponding to different clusters can have this property. Just like in the case of using the EM algorithm for clustering multivariate data, the problem disappears if we succeed in obtaining precise enough data clustering which allows a compatible grouping of data projections $\tau' X(t)$, $t = \overline{1, n}$. Later these clusters of univariate data can be improved using the EM algorithm and the inversion formula is applied in the above described way. So, the core problem is obtaining of the initial estimate $\pi(\cdot)$. The authors suggests using the geometric clustering approach for this task. The pseudo-distances matrices are based on the clustering results of univariate projections.

3. Geometric Clustering Procedure

Let us describe the proposed hard data clustering procedure. Let $\rho(x, y)$ be a non-negative pseudo-distance function defined for all $x, y \in \mathbb{R}^d$. In the general case, this function does not satisfy the triangular inequality. Let us divide the set of observation numbers N into non-intersecting subsets $\widehat{K}_1, ..., \widehat{K}_q$ so as to minimize value of the functional

$$Q(K_1, ..., K_q) = \sum_{j=1}^{q} \frac{1}{\|K_j\|} \sum_{s,r \in K_j} \rho\big(X(s), X(r)\big). \tag{13}$$

The minimization algorithms are widely analysed and we will not detail in this paper. The main task is selection of the distance function ρ. Let us consider the observations

$x, y \in \mathbb{X}$ to be closer to one another, the more similar are the estimates of a posteriori classification probabilities of their projections. So $\rho(x,y) = \max\limits_{\tau \in T} \rho_\tau(x,y)$, where

$$\rho_\tau(x,y) = \sum_{j=1}^{q} |\widehat{\pi}_\tau(j,x) - \widehat{\pi}_\tau(j,y)|, \tag{14}$$

and $\widehat{\pi}_\tau(j,x)$ denotes the estimate of probability

$$\pi_\tau(j,x) = \mathbb{P}\{\nu = j | \tau' X = \tau' x\}.$$

Thus, preliminary hard clustering of a sample consists of the following stages

- selection of a projection directions set $T \subset B$;
- calculation of the maximum likelihood estimates

$$\widehat{\theta}(\tau) = \arg\max_{\theta(\tau)} \sum_{s=1}^{n} \log f_\tau(\tau' X(s), \theta(\tau)) \tag{15}$$

  or their approximation for each $\tau \in T$;
- estimation of a posteriori probabilities by means of the "plug-in" technique

$$\widehat{\pi}_\tau(j,x) = \frac{\widehat{p}_j(\tau)\widehat{\varphi}_{j,\tau}(\tau' x)}{f_\tau(\tau' x, \widehat{\theta}(\tau))}, \quad \tau \in T;$$

- calculation of a set of clusters

$$(\widehat{K}_1, ..., \widehat{K}_q) = \arg \min_{K_1,...,K_q} Q(K_1, ..., K_q)$$

Next, analogously as having the training samples, it is possible to calculate the initial multivariate estimate $\widetilde{\theta}$ of parameter θ defined by the equalities

$$\begin{aligned} \widetilde{p}_j &= \frac{\|K_j\|}{n}, \\ \widetilde{M}_j &= \frac{1}{\|K_j\|} \sum_{s \in K_j} X(s), \\ \widetilde{R}_j &= \frac{1}{\|K_j\|} \sum_{s \in K_j} X(s) X'(s) - \widetilde{M}_j \widetilde{M}_j', \end{aligned} \tag{16}$$

where $j = \overline{1,q}$. Substituting $X(s)$ for their projections $\tau' X(s)$ in (16), we define the preliminary compatible estimates $\widetilde{\theta}(\tau)$, $\tau \in T$.

Using the "plug-in" approach (5), from $\widetilde{\theta}$ we obtain a preliminary estimate of soft clustering $\widetilde{\pi}$. This estimate can be improved by a recurrent EM algorithm until a new estimate becomes stable. Let us denote this new estimate by $\widehat{\pi}_{EM}$.

Applying the inversion formula in clustering, we first must fix the estimates $\widetilde{p}_j$ and improve other elements of $\widetilde{\theta}(\tau)$ by the EM algorithm. We calculate the statistics $\widehat{\varphi}_j$ placing $\widehat{f}_\tau^*$ by the estimates of characteristic functions corresponding to the obtained estimates of $\widetilde{\theta}(\tau)$ in formula (11). These statistics define the estimate of π by equalities

$$\widehat{\pi}_I(j,x) = \frac{\widetilde{p}_j\widehat{\varphi}_j(x)}{\sum\limits_{i=1}^{q}\widetilde{p}_i\widehat{\varphi}_i(x)}, \quad j=\overline{1,q}. \tag{17}$$

4. Simulation Results

It is not difficult to make analysis of precision of the proposed estimates by means of simulation. We assume the mean error of soft classification

$$\Delta_\pi(\widehat{\pi}) = \mathbb{E}\sum_{i=1}^{q}\big|\widehat{\pi}(j,X)-\pi(j,X)\big|. \tag{18}$$

as a precision measure of estimation of a posteriori probabilities. Having selected the parameter of the model θ and sample size n, we generate independent random samples $\mathbb{X}^{(s)} = (X^{(s)}(1),...,X^{(s)}(n))$, $s=1,\ldots,r$, and obtain realizations $\widehat{\pi}^{(1)},...,\widehat{\pi}^{(r)}$ of the estimates analysed. The empirical analogue of functional (18) is the statistics

$$\delta_\pi(\widehat{\pi}) = \frac{1}{rn}\sum\big|\widehat{\pi}^{(s)}(j,X^{(s)}(t))-\pi(j,X^{(s)}(t))\big|, \tag{19}$$

where the sum is taken over $j=\overline{1,q}$, $t=\overline{1,n}$, $s=\overline{1,r}$. We used $r=10$ in all the experiments of this study.

In practice we can meet not only the clustering problem, but also the problem of estimation of a Gaussian distribution mixture parameters or distribution density. Using (7), from the classification probabilities $\widehat{\pi}$ we can obtain the estimate of the mixture parameter $\widehat{\theta}$. The error

$$\Delta_\theta(\widehat{\theta}) = \mathbb{E}\int\limits_{\mathbb{R}^d}\big(f_{\widehat{\theta}}(x)-f_\theta(x)\big)^2\,\mathrm{d}x, \tag{20}$$

will be regarded as the precision measure for this parameter as well as for the parametric density estimate and its with empirical analogue would be

$$\delta_\theta(\widehat{\theta}) = \frac{1}{r}\sum_s\int\limits_{\mathbb{R}^d}\big(f_{\widehat{\theta}^{(s)}}(x)-f_\theta(x)\big)^2\,\mathrm{d}x, \tag{21}$$

which is not so difficult to calculate, because the integral can be expressed analytically. The main purpose of this research is to compare the precision of estimates $\widetilde{\pi}$, $\widehat{\pi}_{EM}$ and $\widehat{\pi}_{ML}$. The latter estimate is obtained by applying the "plug-in" approach to the maximum

Table 1

Error dependence on the sample size[2].

Sample	Error	$\widehat{\pi}_{ML}$	$\widetilde{\pi}$	$\widehat{\pi}_{EM}$	$\widetilde{\pi}^*$	$\widehat{\pi}^*_{EM}$
$n = 50$	δ_π	0.01987	0.06111	0.07903	0.10053	0.11879
	$\delta_\theta \times 100$	0.41735	0.46795	0.47304	0.46489	0.48914
$n = 100$	δ_π	0.01478	0.02578	0.01504	0.02773	0.01504
	$\delta_\theta \times 100$	0.26139	0.20742	0.23366	0.21509	0.23366
$n = 200$	δ_π	0.01397	0.01814	0.01397	0.02315	0.01397
	$\delta_\theta \times 100$	0.14650	0.13488	0.14650	0.13682	0.14650
$n = 500$	δ_π	0.00959	0.01748	0.00959	0.01753	0.00959
	$\delta_\theta \times 100$	0.07321	0.08112	0.07321	0.08265	0.07321

likelihood estimate (6), i.e. $\widehat{\pi}_{ML}(\cdot) = \pi_{\widehat{\theta}_{ML}}(\cdot)$. When simulating, we calculated this estimate by means of the EM algorithm with the true value of θ as an initial value.

In some cases, at small sample values, not only the above mentioned estimates, but also the value of statistic (17) were calculated. The maximum likelihood estimator is asymptotically efficient, but for a small sample size its errors can be on the average greater than those of other estimators including $\widehat{\pi}_I$.

In the study of precision the maximum likelihood estimates of the parameters of a univariate Gaussian mixtures were used. The EM algorithm with the initial value equal to the true value of $\theta(\tau)$ was used to obtain estimates $\widehat{\theta}(\tau)$. Unfortunately, it is impossible to do in a real application, that's why we also used the software[1] for estimating $\theta(\tau)$ based on the mentioned technique described in (Rudzkis and Radavicius, 1995). The latter cases are denoted by "*". Table 1 presents errors of the methods based on the maximum likelihood estimates pseudoestimates and the real estimates.

In the simulation process, it has been noticed that clustering probability estimates obtained by geometric clustering procedure are more stable if the maximum in formula (14) is calculated rejecting a part of directions τ, with the highest value of sum $\sum_{t,k=1}^{n} \rho_\tau(X(t), X(k))$. The number of rejected directions can be selected using maximum likelihood criterion. This modification of the method is natural, because without it even one direction with very inaccurate estimates of classification probabilities $\widehat{\pi}_\tau(\cdot,\cdot)$ can determine the value of distance function ρ.

The proposed methods were analysed by a variety of mixtures having different properties: non-overlapping, partly overlapping or highly overlapping clusters; similar or significantly different cluster probabilities; small or large sample size; various structure of

[1]Software for estimation the parameter of Gaussian mixture developed in Institute of Mathematics and Informatics, Lithuania

[2]The parameters of the mixture are $d = 5$, $q = 4$, $p_1 = p_2 = p_3 = p_4 = 0.25$, $M_1 = (8, 8, 0, 0, 0)'$, $M_2 = (-8, 8, 0, 0, 0)'$, $M_3 = (-8, -8, 0, 0, 0)'$, $M_4 = (8, -8, 0, 0, 0)'$, $R_1 = \mathrm{diag}\,(16, 1, 1, 9, 4)$, $R_2 = \mathrm{diag}\,(9, 16, 4, 1, 4)$, $R_3 = \mathrm{diag}\,(25, 16, 4, 9, 1)$ and $R_4 = \mathrm{diag}\,(25, 4, 9, 1, 9)$.

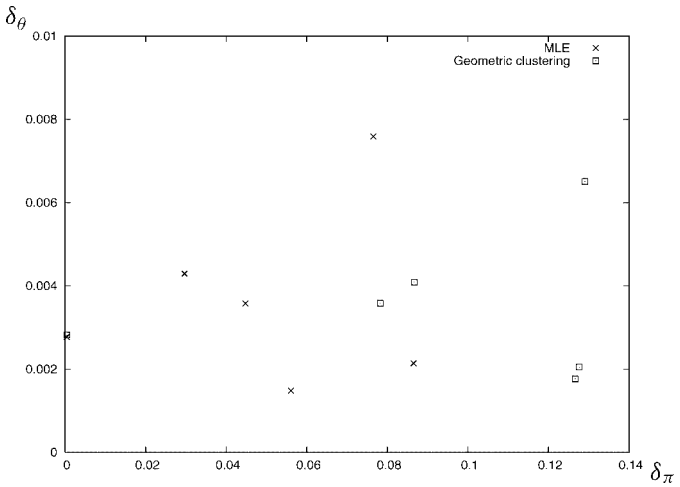

Fig. 1. Error plot of the maximum likelihood and geometric clustering procedures.

covariance matrices. Let us take one typical case of mixture with partly overlapping clusters and analyse the dependence of the method precision on the sample size n. The results of this research are given in the table. The parameters of the Gaussian mixture used are $d = 5$ and $q = 4$, i.e., $\dim\theta = 83$. With such a dimension of θ, we consider the sample where $n = 100$ to be very small, $n = 200$ – small, and $n = 500$ of medium size. The results for an extremely small size $n = 50$ will also be presented for comparison. The Monte-Carlo simulation study has shown that it is enough to use 200–500 univariate projections for geometric clustering of 5-dimensional data. Therefore we used a set of uniformly distributed directions T containing 500 directions in further research.

The comparison error plot of the maximum likelihood method and the proposed geometric clustering method $\widetilde{\pi}^*$ is presented in Fig. 1. The plot shows, that geometric clustering method has larger δ_π error than maximum likelihood estimator, but it is more precise by means of δ_θ criterion.

The conclusions follow from the given error table and error plot:

- the maximum likelihood method is the best method to estimate classification probabilities;
- in the case of a small and medium sample, the proposed geometric clustering method, improved by the EM algorithm yields the same errors as the maximum likelihood estimator (pseudo-estimator, to be precise, because it is calculated using the true values of parameter θ). This conclusion is also valid for the estimator $\widehat{\pi}_{EM}$ which has been calculated relying on the assumption that the MLE could be found for univariate data, as well as for the estimator $\widehat{\pi}^*_{EM}$ which is the sample function. Therefore the authors of the paper suggest using exactly this statistic combining the univariate data clustering method (Rudzkis and Radavicius, 1995), geometric clustering and the EM algorithm for the multivariate data;
- the clustering method for univariate data described in (Rudzkis and Radavicius, 1995) is precise enough and it can be used to calculate pseudo-distances in the

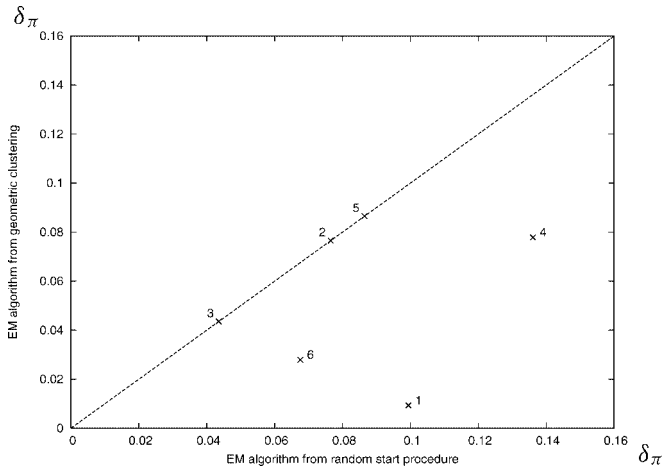

Fig. 2. Comparison plot of the EM algorithm initialized using random start procedure and the EM algorithm initialized using the proposed geometric clustering. The values of classification error δ_π are presented in the plot. The random start initialization failed to calculate the MLE estimate of parameter if sample size is small and the probabilities of the clusters are significantly different. The classification error are significantly greater in these cases. In the other cases, the EM algorithm converges to the same value using both initialization methods.[3]

geometric clustering procedure because the errors of estimators $\widetilde{\pi}$ and $\widetilde{\pi}^*$ differ but slightly;

- though the maximum likelihood estimator $\widehat{\pi}_{ML}$ is the best one to estimate classification probabilities, it is not best to estimate the distribution density. Statistic $\widetilde{\pi}$ and $\widetilde{\pi}^*$ often yields lower errors δ_θ than $\widehat{\pi}_{ML}$.

The simulation results have also indicated that the geometric clustering algorithm works properly even without the EM algorithm improvement when there exist direction in which the clusters of projected data are well separated.

The proposed methods were compared with the maximum likelihood estimator and with a popular in a statistical software the random start EM procedure also. Random start procedure uses the random partitioning of observations for an initial parameter of the EM algorithm. For large data samples the subsampling of data were used as suggested in (McLachlan et al, 1999). This is to limit the effect of the centre limit theorem witch would have the randomly selected starts being similar for each component in large samples. In many cases the random start method should be repeated 20–200 times to calculate the MLE of parameter. But if a prior probabilities of the clusters are significantly different, the random start procedure does not find a cluster with a small a prior probability value even after a large number of start points (our analysis is based on 4000 start points). The simulation results are displayed in Fig. 2. If sample size is small and the probabilities of the clusters are significantly different the proposed EM algorithm initialization method is better than the random start method. Using the proposed procedure improved by EM algorithm the MLE estimate is calculated.

[3]The parameters of the mixtures are given in appendix.

We will briefly review the results analysis based on the Inversion formula method. After the first experiments it has already been observed that the distribution density estimate defined by formula (11) is not smooth and additional smoothing is needed. The authors of the paper propose to introduce an additional multiplier e^{-hu} under the integral sign, where h is a small value, a so-called smoothing bandwidth. Such an expression of the smoothing multiplier is especially convenient in the case of Gaussian distribution mixtures, because we can still find the exact value of the integral (11). In practice, introduction of such a multiplier simply increases the estimates of cluster variance by the value $2h$. The simulation analysis indicated that it is optimal to select the smoothing bandwidth value dependent on scale parameter of the data. The reasonable choice is $c \cdot \lambda_{max}$, where λ_{max} is the maximum eigenvalue of the data covariance matrix. The parameter c can be selected using the maximum likelihood criterion. The Monte-Carlo research has also indicated that the same smoothing bandwidth value cannot be used for estimating density values of separate components of the mixture. It should be chosen adaptively, depending on the properties of the component. The smoothing bandwidth selection problem requires more comprehensive study.

5. Appendix

This appendix contains the parameters of analysed mixtures.

Mixture 1

$$d = 5,\ q = 2,\ n = 250,\ p_1 = 0.92,\ p_2 = 0.08,$$
$$M_1 = (0,0,0,0,0)',\ M_2 = (5,0,0,0,0)',$$
$$R_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0.07 \\ 0 & 4 & -0.5 & 0 & 0 \\ 0 & -0.5 & 7 & 0 & 0.4 \\ 0 & 0 & 0 & 4 & 0 \\ 0.07 & 0 & 0.4 & 0 & 3 \end{pmatrix},$$
$$R_2 = \begin{pmatrix} 6 & 0 & 0 & 0 & 0 \\ 0 & 4.6994 & 0.7182 & -2.8503 & -0.6473 \\ 0 & 0.7182 & 4.6341 & -1.9133 & -0.2569 \\ 0 & -2.8503 & -1.9133 & 5.6764 & -1.3610 \\ 0 & -0.6473 & -0.2569 & -1.3610 & 2.9899 \end{pmatrix}$$

Mixture 2

$$d = 5,\ q = 4,\ n = 400,\ p_1 = p_3 = 0.2,\ p_2 = p_4 = 0.3,$$
$$M_1 = M_2 = M_3 = M_4 = (0,0,0,0,0)',\ R_1 = \operatorname{diag}(1,1,1,1,1),$$

$$R_2 = \begin{pmatrix} 16 & 0 & 0 & 0 & 8 \\ 0 & 25 & -5 & 0 & 0 \\ 0 & -5 & 25 & 4 & 0 \\ 0 & 0 & 4 & 25 & 0 \\ 8 & 0 & 0 & 0 & 36 \end{pmatrix},$$

$$R_3 = \begin{pmatrix} 36 & 20 & 0 & 0 & 0 \\ 20 & 65 & -5 & 0 & 0 \\ 0 & -5 & 4 & 1 & 0 \\ 0 & 0 & 1 & 16 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$R_4 = \begin{pmatrix} 25 & 0 & 10 & 0 & 8 \\ 0 & 4 & -2 & 0 & 0 \\ 10 & -2 & 49 & 4 & 12 \\ 0 & 0 & 4 & 16 & 0 \\ 8 & 0 & 12 & 0 & 16 \end{pmatrix}.$$

Mixture 3

$d = 5,\ q = 3,\ n = 200,\ p_1 = 0.35,\ p_2 = 0.3,\ p_3 = 0.35,$
$M_1 = (-5, 0, 0, 0, 0)',\ M_2 = (0, 0, 0, 0, 0)',\ M_3 = (5, 0, 0, 0, 0)',$
$R_1 = \operatorname{diag}(1, 1, 1, 1, 1),\ R_2 = \operatorname{diag}(9, 1, 4, 1, 4)',\ R_3 = \operatorname{diag}(2, 16, 4, 25, 1).$

Mixture 4

$d = 5,\ q = 3,\ n = 220,\ p_1 = 0.05,\ p_2 = 0.9,\ p_3 = 0.05,$
$M_1 = (-5, 0, 0, 0, 0)',\ M_2 = (0, 0, 0, 0, 0)',\ M_3 = (5, 0, 0, 0, 0)',$
$R_1 = \operatorname{diag}(1, 1, 1, 1, 1),\ R_2 = \operatorname{diag}(9, 1, 4, 1, 4)',\ R_3 = \operatorname{diag}(2, 16, 4, 25, 1).$

Mixture 5

$d = 5,\ q = 4,\ n = 200,\ p_1 = p_2 = p_3 = p_4 = 0.25,$
$M_1 = (-5, 0, 0, 0, 0)',\ M_2 = (5, 0, 0, 0, 0)',\ M_3 = (-5, 0, 0, 0, 0)',\ M_4 = (5, 20, 0, 0, 0)',$
$R_1 = \operatorname{diag}(4, 9, 1, 4, 9),\ R_2 = \operatorname{diag}(6, 4, 25, 1, 9),\ R_3 = \operatorname{diag}(1, 16, 4, 64, 1),$
$R_4 = \operatorname{diag}(1, 25, 36, 4, 6).$

Mixture 6

$d = 5,\ q = 2,\ n = 200,\ p_1 = 0.9,\ p_2 = 0.1,$
$M_1 = (0, 0, 0, 0, 0)',\ M_2 = (15, 0, 0, 0, 0)',$
$R_1 = \operatorname{diag}(25, 1, 1, 1, 1),\ R_2 = \operatorname{diag}(1, 1, 1, 1, 1)$

References

Berkhin, P. (2002). Survey of clustering data mining techniques. *Technical report*, Accrue Software, San Jose, USA.

Boyles, R.A. (1983). On the convergence of the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, **45**, 47–50.

Fraley, C., and A.E. Raftery (1999). MCLUST: Software for model-based cluster analysis, *Journal of Classification*, **16**, 297–306.

Fraley, C., and A.E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of American Statistical Association*, **97**(458), 611–631.

Friedman, J.H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, **82**(397), 249–266.

Huber, P.J. (1985). Projection pursuit (with discussion). *The Annals of Statistics*, **13**, 435–475.

McLachlan, G.J., D. Peel, K.E. Basford, P. Adams (1999). The EMMIX software for the fitting of mixtures of normal and t-components, *Journal of Statistical Software*. **4**(2).

Rudzkis, R., and M. Radavicius. (1995). Statistical estimation of a mixture of Gaussian distributions. *Acta Applicandae Mathematicae*, **38**, 37–54.

Rudzkis, R., and M. Radavicius. (1999). Characterization and statistical estimation of a discriminant space for Gaussian mixtures. *Acta Applicandae Mathematicae*, **59**, 617–626.

Rudzkis, R., and M. Radavicius. (2003). Testing hypotheses on discriminant space in the mixture model of Gaussian distributions. *Acta Applicandae Mathematicae*, **79**, 105–114.

Vlassis, N., A. Likas (2000). A greedy EM algorithm for Gaussian mixture learning. *Neural Processing Letters*, **15**(1), 77–87.

Wolfe, J.H. (1967). NORMIX: computational methods for estimating the parameters of multivariate normal mixture distributions. *Technical Bulletin USNPRA SRM 68-2*, U.S. Naval Personnel Research activity, San Diego.

M. Kavaliauskas, born in 1975, in 1997 graduated with master degree from Kaunas University of Technology, Faculty of Fundamental Sciences. Currently continues the studies at Institute of Mathematics and Informatics to acquire a PhD degree. He is also scientific fellow of Kaunas University of Technology. His fields in scientific research interest are non-parametric density estimation and data clustering.

R. Rudzkis was born in Kaunas, in 1951. In 1973 he graduated from Kaunas Polytechnic Institute (speciality of computers). In 1978 he maintained a doctoral degree in mathematics at Vilnius University and in 1993 a degree of doctor habilius in mathematics at Institute of Mathematics and Informatics. He is a professor, expert-member of the Lithuanian Academy of Sciences. Currently he is a head of the Applied Statistics Department at Institute of Mathematics and Informatics. R. Rudzkis fields of scientific research interest are theoretical and applied statistics, econometrics.

Daugiamačių duomenų klasterizavimas Gauso skirstinio mišinio modelyje

Mindaugas KAVALIAUSKAS, Rimantas RUDZKIS

Straipsnyje nagrinėjamas didelio matavimo nepriklausomų atsitiktinių dydžių, tenkinančių Gauso skirstinių mišinio modelį, imties negriežto klasterizavimo (klasifikavimo be apmokymo) uždavinys. Pagrindinis dėmesys skiriamas algoritmams, kurie remiasi duomenų projektavimu į mažesnio matavimo erdves. Pasiūlyta nauja kelių etapų klasterizavimo procedūra, apimanti stebėjimų vienamačių projekcijų pasiskirstymo statistinę analizę, daugiamatės imties klasterizavimą geometriniais metodais ir EM algoritmo taikymą. Darbe pateikiami nagrinėjamų metodų tikslumo analizės rezultatai, gauti Monte-Karlo būdu.