Liet. matem. rink., T. 41, spec. nr., 438–443

Influence of the outliers to the classification of multidimensional Gaussian mixtures

Gintautas JAKIMAUSKAS (MII)
e-mail: gnt@ktl.mii.lt

1. Introduction

We consider the problem of estimating a posteriori probabilities from the multidimensional sample supposed to satisfy multidimensional Gaussian mixture model with the outliers (i.e. random noise). It is known (see, e.g., [6] and [7]) that in the case of pure Gaussian mixture model projection to lower dimension subspace can reduce errors of estimates of a posteriori probabilities. Suppose that outliers in the mixture model satisfy Tukey-Huber distortion model (see [11]). We present computer simulation results which demonstrate the influence of noise level and noise type to the estimates of a posteriori probabilities.

In general we face a dilemma: classification of the initial sample or projection to lower dimension subspace and then classification of the projected sample. Presence of the outliers makes this problem more complicate. One of possible methods to select whether to project the data to lower dimension subspace or not is the bootstrap method. We simulate realizations with some preliminary parameters obtained from the sample, then compare errors of estimates of a posteriori probabilities, assuming that preliminary parameters are true parameters of the sample. One of the main difficulties is to obtain sufficiently good preliminary parameters in the presence of the outliers. Also the outliers make impact on the results of the projection pursuit procedure finding basic vectors of the discriminant subspace.

Theoretical background of the problem of classification of multidimensional Gaussian mixture is given, e.g., in [8]–[10]. The problem of presence of the outliers is given, e.g., in [11]. We are thankful to prof. R. Rudzkis who gave the idea and many constructive and valuable remarks.

The introduction presents already known methods. Description of the EM algorithm and the projection pursuit algorithm is given, e.g., in [8]–[9].

Main definitions. Let $\varphi(\cdot; M_i, R_i) \stackrel{def}{=} \varphi_i$, $i = 1, 2, \ldots, q$ be different d-dimensional Gaussian distribution densities, where means M_i and covariance matrices R_i, $i = 1, 2, \ldots, q$, are unknown. Let we have q independent d-dimensional random variables Y_i^* with distribution densities

$$\varphi_i^*(x) = (1 - \varepsilon_i)\varphi_i(x) + \varepsilon_i h_i(x), \quad x \in \mathbb{R}^d, \tag{1}$$

where h_i is "distortion" probability density for the qth class, ε_i is outlier probability of the qth class. Let ν^* be random variable (r.v.) independent of Y_i^*, $i = 1, 2, \ldots, q$, and taking on values $1, 2, \ldots, q$ with unknown probabilities $p_i^* > 0$, $i = 1, 2, \ldots, q$, respectively. In this paper we assume that number of classes q is known. We observe d-dimensional r.v. $X = Y_{\nu^*}^*$. Each observation belongs to one of q classes (including outliers) depending on r.v. ν^*. Distribution density of r.v. X is therefore a Gaussian mixture density with Tukey-Huber distortions

$$f(x) = \sum_{i=1}^{q} p_i^* \varphi_i^*(x) \stackrel{def}{=} f(x, \theta), \quad x \in \mathbb{R}^d, \tag{2}$$

where $\theta = (p_i^*, M_i, R_i,\ i = 1, 2, \ldots, q)$ is an unknown multidimensional parameter to be estimated. Probabilities $p_i^* = \mathbf{P}\{\nu^* = i\}$ are a *priori* probabilities for r.v. X to belong to ith class (including probability of the outlier). Denote

$$p_0 = \varepsilon = \sum_{i=1}^{q} p_i^* \varepsilon_i, \tag{3}$$

$$h(x) = \sum_{i=1}^{q} \frac{p_i^* \varepsilon_i}{p_0} h_i(x), \tag{4}$$

and let all the outliers belong to the class 0. Now let we have q independent d-dimensional Gaussian random variables Y_i with distribution densities $\varphi_i(x)$ and independent d-dimensional random variable Y_0 with distribution density $h(x)$. Let ν be random variable (r.v.) independent of Y_i, $i = 0, 1, 2, \ldots, q$, and taking on values $0, 1, 2, \ldots, q$ with unknown probabilities $p_i > 0$, $i = 0, 1, 2, \ldots, q$, respectively. Probabilities $p_i = \mathbf{P}\{\nu = i\}$ are a *priori* probabilities for r.v. X to belong to ith class. We can rewrite equality (2) as follows:

$$f(x) = \sum_{i=1}^{q} p_i \varphi_i(x) + p_0 h(x). \tag{5}$$

We will consider the general classification problem of estimating *a posteriori* probabilities $\pi(i, x) = \mathbf{P}\{\nu = i | X = x\}$ from the sample $\{X_1, X_2, \ldots, X_N\} \stackrel{def}{=} X^N$ of i.i.d. random variables with distribution density (5). Under assumptions above

$$\pi(i, x) = \pi_\theta(i, x) = \frac{p_i \varphi_i(x)}{f(x, \theta)}, \quad i = 1, 2, \ldots, q, \quad x \in \mathbb{R}^d. \tag{6}$$

The most common method to estimate a posteriori probabilities is based on the EM-algorithm (see, e.g., [8]).

Let $V = \operatorname{cov}(X, X)$ be the covariance matrix of r.v. X. Define the scalar product of arbitrary vectors $u, h \in \mathbb{R}^d$ as $(u, h) = u^T V^{-1} h$ and denote by u_H the projection of arbitrary vector $u \in \mathbb{R}^d$ to a linear subspace $H \subset \mathbb{R}^d$. Discriminant space H is defined as

a linear subspace $H \subset \mathbb{R}^d$ with the property $\mathbf{P}\{\nu = i|X = x\} = \mathbf{P}\{\nu = i|X_H = x_H\}$, $i = 1, 2, \ldots, q$, $x \in \mathbb{R}^d$, and the minimal dimension. Denote $k = \dim H$. It is known that for Gaussian mixture densities with equal covariance matrices we have $k < q$. Clearly, if $k < q$ and H is known, then it is better to estimate a posteriori probabilities from the projected sample rather than initial sample. Unfortunately, in practice H is not known and must be estimated and we get additional estimation errors. As shown in [6] and [7] in many cases despite additional errors the projected sample allows to decrease errors of estimation of a posteriori probabilities.

2. Computer simulation results

We present computer simulation results on the influence of the outliers to the statistical procedure of the selection one of the two methods of estimating of a posteriori probabilities:

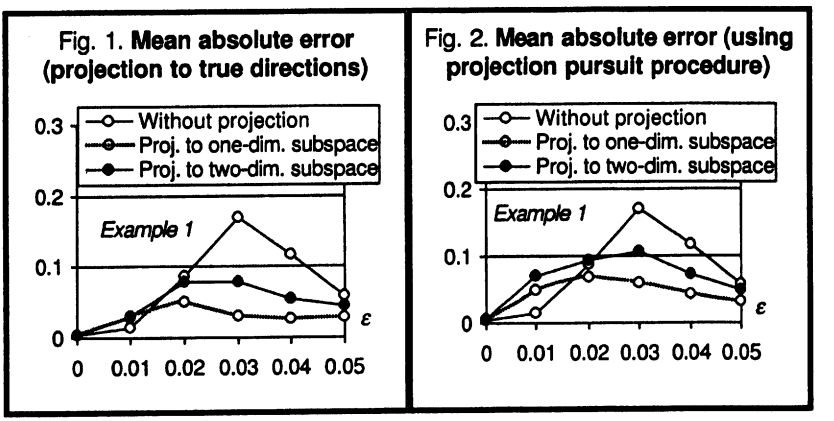
1) Method based on application of the EM algorithm to the initial sample from $\mathbb{R}^d$. This method is implemented in software created in Institute of Mathematics and Informatics (see [5]). This software does not require user intervention, because initial parameter estimates are selected from the sample;

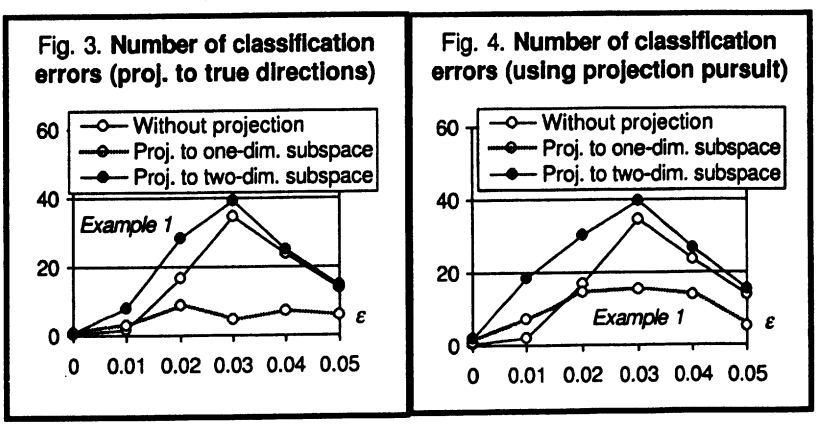
2) Two stage estimation method, where in the first stage we estimate k-dimensional space H from the sample (see [9], [10]). In the second stage a posteriori probabilities are estimated using first method to the projected sample.

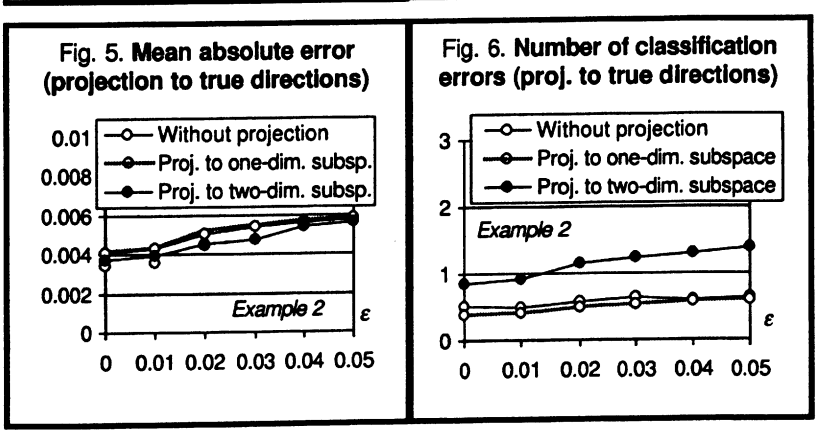
Computer simulation was done as follows. We simulate the sample X^N with the selected mixture model and the selected sample size (in our case $N = 300$). As basic mixture model we selected 5-dimensional Gaussian mixture model with three clusters with means $(-5, -a, 0, 0, 0)$, $(0, 2a, 0, 0, 0)$, $(5, a, 0, 0, 0)$, equal probabilities and unit covariance matrices. At the next step we obtain parameters for bootstrap using the completely automatic procedure, which starts from no information about cluster structure. Bootstrap begins with simulating selected number (in our case 10 realizations) of independent realizations with obtained parameters now supposed to be known. To each realization we apply the procedure of calculating accuracy of estimation of a posteriori probabilities without projection, with projection to one-dimensional subspace and with projection to two-dimensional subspace (as in [7]).

In this paper we present two examples of computer simulation results that demonstrate influence of the outliers to the of the automatic parameter estimation procedure and bootstrap methods. In these examples parameters for bootstrap are obtained from theoretical parameters (using EM algorithm) that were used for simulation of the sample X^N excluding the noise cluster.

We studied accuracy of estimation of a posteriori probabilities and number of Bayesian classification errors (i.e., classification using estimated parameters vs. classification using theoretical parameters). Accuracy of estimation of a posteriori probabilities is measured as mean absolute distance $l(\hat{\pi}^N, \pi^N)$ between the estimated a posteriori probabilities $\hat{\pi}^N$ and the theoretical a posteriori probabilities π^N. We compare distance $l(\hat{\pi}^N, \pi^N)$







and $l(\hat{\pi}_H^N, \pi^N)$ where $\hat{\pi}^N$ are obtained from MLE in the initial space and $\hat{\pi}_H^N$ are obtained from MLE in the discriminant subspace H. Number of Bayesian classification errors is measured as percentage of differences in Bayesian classification comparing classification using known theoretical parameter versus classification using estimated parameter.

In Example 1 (see Figs. 1–4, in all Figs. on x axis we have noise level ε) we use basic mixture model (parameter $a = 0$) with one noise cluster with center at zero point and covariance matrix 100 times unit covariance matrix (this component is added to the standardized basic mixture model). In the case of projection to true directions we project data to x and y axes. In other case we project data to the directions obtained using projection pursuit procedure. We compare accuracy of estimation of a posteriori probabilities, number of Bayesian classification errors.

In Example 2 (see Figs. 5–6) we use basic mixture model (parameter $a = 0$) with two symmetrical noise clusters with centers $(0, -3, 0, 0, 0)$ and $(0, 3, 0, 0, 0)$ and unit covariance matrices added to the standardized basic mixture model.

Computer simulation results show that large deviation of the outliers makes the big influence on the estimates of a posteriori probabilities even for low noise level. At the same time in the case of projection to the lower dimension subspace we obtain better accuracy of estimates of a posteriori probabilities. Influence of the outliers to the projection pursuit procedure results is comparatively small. Reversely, small deviation of the outliers (situated far from the main clusters) makes little influence on the estimates of the a posteriori probabilities even for high noise level. This is because in this case automatic procedure finds noise clusters well separated from the main clusters. However, outliers situated outside discriminant subspace for the high noise level makes impact to the results of the projection pursuit procedure. From the computer simulation results we can draw a conclusion that some modification is required to the procedures to make them more robust to the presence of the outliers.

References

[1] S.A. Aivazyan, Mixture approach to clustering via maximum likelihood, criteria of model complexity and projection pursuit, In *Data Science, Classification and Related Methods, Abstracts of 5th IFCS Conference*, IFCS, Cobe, 1(36) (1996).

[2] S.A. Aivazyan, V.M. Buchstaber, I.S. Yenyukov, L.D. Meshalkin, Applied statistics, Classification and reduction of dimensionality, *Finansy i Statistika*, M. (1989) (in Russian).

[3] J.H. Friedman, Exploratory projection pursuit, *J. Amer. Statist. Assoc.*, **82**, 249–266 (1987).

[4] J.H. Friedman, J.W. Tukey, A projection pursuit algorithm for exploratory data analysis, *IEEE Trans. Comput.*, C-21, 881–889 (1974).

[5] G. Jakimauskas, Efficiency analysis of one estimation and clusterization procedure of one-dimensional Gaussian mixture, *Informatica*, **8**(3), 331–343 (1997).

[6] G. Jakimauskas, R. Krikštolaitis, Appropriateness of a projection pursuit in classification, *LMD mokslo darbai*, MII, Vilnius, **3**, 370–374 (1999).

[7] G. Jakimauskas, R. Krikštolaitis, Influence of of projection pursuit on classification errors: computer simulation results, *Informatica*, **11**(2), 115–124 (2000).

[8] R. Rudzkis, M. Radavičius, Statistical estimation of a mixture of Gaussian distributions, *Acta Applicandae Mathematicae*, **38**, 37–54 (1995).

[9] R. Rudzkis, M. Radavičius, Projection pursuit in Gaussian mixture models preserving information about cluster structure, *Liet. Matem. Rink.*, **37**(4), 550–563 (1997) (in Russian).

[10] R. Rudzkis, M. Radavičius, Characterization and statistical estimation of a discriminant space for Gaussian mixtures, *Acta Applicandae Mathematicae*, **58**, 279–290 (1999).

[11] E.E. Zhuk, Yu.S. Kharin, *Robustness in Cluster Analysis of Multivariate Observations*, Belarus State University, Minsk, (1998) (in Russian).

Išsiskiriančių stebėjimų įtaka daugiamačių Gauso mišinių klasifikavimui

G. Jakimauskas

Nagrinėtas apriorinių tikimybių statistinio įvertinimo uždavinys, kai stebėjimai tenkina daugiamačio Gauso mišinio modelį su išsiskiriančiais stebėjimais. Tiriamas išsiskiriančių stebėjimų įtaka bootstrap metodo taikymui, kuomet yra parenkamas vienas iš dviejų metodų: EM algoritmo taikymas pirminiams duomenims arba projektuotiems į mažesnės dimensijos erdvę.