Liet. matem. rink., T. 42, spec. nr., 523–528

Procedure of the removal of the outliers from the sample satisfying the multidimensional Gaussian mixture model

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

1. Introduction

Let us have a sample supposed to satisfy a multidimensional Gaussian mixture model with outliers (i.e., random noise). Let us consider the problem of estimating a posteriori probabilities from the sample.

The most common method for this problem is application of the EM algorithm to the initial sample. Besides that, it is known (see, e.g., [1]) that projection of a sample to a lower dimension subspace (so called discriminant subspace, see, e.g., [4]) can reduce errors of the estimates of a posteriori probabilities. This requires an additional step – estimating of the basic vectors of the discriminant subspace. The presence of a random noise makes this problem more complicate (see, e.g., [2]), especially because of non-robustness of the projection pursuit procedure, which is step-by-step procedure to find the basic vectors of the discriminant subspace.

R. Rudzkis and M. Radavičius in [3] proposed the sequential procedure of selection of initial values for the EM algorithm. This procedure has an additional cluster (background cluster). At each step of the procedure sample elements, which can be hardly assigned to any of Gaussian clusters are assigned to the background cluster. At the end of the procedure background cluster is deleted or replaced by the Gaussian cluster. We propose modification of the mentioned procedure, which is suitable to delete the outliers from the sample. At the end of the procedure we leave some positive noise level and the vast majority of the outliers are assigned to the noise cluster. This allows us to remove the outliers from the sample.

Computer simulation results show that properly selected noise level at the end of modified automatic procedure helps to assign the outliers to the noise cluster and easily remove them.

Theoretical background of the problem of classification of multidimensional Gaussian mixture is given, e.g., in [3]–[5]. The problem of robustness in cluster analysis of multivariate observations is analysed (especially, using minimum contrast estimates with truncated contrast function), e.g., in [6]. We are thankful to prof. R. Rudzkis who gave the idea and many constructive and valuable remarks. The introduction presents already known methods.

Main definitions. Let have q independent d-dimensional Gaussian random variables Y_i with different distribution densities $\varphi(\cdot; M_i, R_i) \stackrel{\text{def}}{=} \varphi_i$, where means M_i and covariance matrices R_i, $i = 1, 2, \ldots, q$, are unknown. 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. We observe d-dimensional r.v. $X = Y_\nu$. Each observation belongs to one of q classes depending on r.v. ν. Distribution density of r.v. X is therefore a Gaussian mixture density

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

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

We will consider the general classification problem of estimating *a posteriori* probabilities $\pi(i, x) = \mathbf{P}\{\nu = i | X = x\}$ from the sample $\{X_1, X_2, \ldots, X_N\} \stackrel{\text{def}}{=} X^N$ of i.i.d. random variables with distribution density (1). Under assumptions above

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

We define a set $K_i = \{X^N, \pi(i, \cdot)\}$ as a cluster corresponding to the ith class and the selected classification rule π.

The most common method to estimate a posteriori probabilities is based on the EM-algorithm (see, e.g., [3]). The EM algorithm is an iterative procedure, which starts either from given parameter θ or given a posteriori probabilities applying in turn formula (2) for calculation of a posteriori probabilities $\pi(i, X_j)$, $i = 1, 2, \ldots, q$, $j = 1, 2, \ldots, N$, from given parameter θ or the following formulae

$$p_i = \frac{1}{N} \sum_{j=1}^{N} \pi(i, X_j), \quad i = 1, 2, \ldots, q, \tag{3a}$$

$$M_i = \frac{1}{N} \sum_{j=1}^{N} \frac{\pi(i, X_j)}{p_i} X_j, \quad i = 1, 2, \ldots, q, \tag{3b}$$

$$R_i = \frac{1}{N} \sum_{j=1}^{N} \frac{\pi(i, X_j)}{p_i} (X_j - M_i)(X_j - M_i)^{\mathrm{T}}, \quad i = 1, 2, \ldots, q, \tag{3c}$$

for calculation of the parameter θ from given a posteriori probabilities $\pi(i, X_j)$, $i = 1, 2, \ldots, q$, $j = 1, 2, \ldots, N$. The EM algorithm usually ends after some predefined number of iterations.

2. Procedure of the removal of the outliers from the sample – computer simulation results

Procedure of the removal of the outliers from the sample is based on the sequential procedure of selection of initial values for the EM algorithm proposed by R. Rudzkis and M. Radavičius in [3]. Very briefly (for details see [3]), given some number $r \in \{0, 1, \ldots\}$ it is supposed that

$$f(x) = \sum_{i=1}^{r} \tilde{p}_i \tilde{\varphi}_i(x) + \tilde{p}_0 \tilde{h}(x) \stackrel{\text{def}}{=} \tilde{f}_r(x, \theta) + \tilde{p}_0 \tilde{h}(x), \quad x \in \mathbf{R}^d, \tag{4}$$

where $\tilde{h}(x)$ is supposed to be a non-Gaussian distribution density. We call the cluster $K_0 = \{X^N, \pi(0, \cdot)\}$ corresponding to the distribution density $\tilde{h}$ the background cluster. The procedure gives an initial estimate for the next component $\tilde{p}_{r+1}\tilde{\varphi}_{r+1}(x)$. Let $\hat{f}$ be a nonparametric estimate of d.d. f. Denote

$$\Delta_r(X) = \Delta_r(X, \theta) = \left[\hat{f}(X) - \tilde{f}_r(X, \theta)\right]_+, \quad X \in X^N. \tag{5}$$

The a posteriori probabilities are calculated using formulae

$$\pi(0, X) = \pi_\theta(0, X) = \Delta_r(X)/\hat{f}(X), \quad X \in X^N, \tag{6a}$$

$$\pi(i, X) = \pi_\theta(i, X) = [1 - \pi(0, X)] \frac{\tilde{p}_i \tilde{\varphi}_i(X)}{\tilde{f}_r(X, \theta)}, \quad i = 1, 2, \ldots, r, \quad X \in X^N. \tag{6b}$$

The probability of the background cluster is calculated using formula

$$\tilde{p}_0 = \frac{1}{N} \sum_{j=1}^{N} \pi(0, X_j). \tag{7}$$

As r becomes equal to q we set $\pi(0, X) \equiv 0$ in (6a) and thus $\tilde{p}_0 = 0$.

This method is implemented in the software (i.e., DLL's written in Pascal language) created in Institute of Mathematics and Informatics (Vilnius) by R. Rudzkis, M. Radavičius, G. Jakimauskas and J. Sušinskas. The main advantage of the software is a possibility of completely automatic estimation of the parameters of the Gaussian mixture.

We propose the modification of the mentioned automatic procedure, which is suitable to remove the outliers from the sample. At the end of the sequential procedure (also in all intermediate steps) we leave some $\tilde{p}_0 \geqslant \varepsilon$, where the choice of ε depends on the actual noise level. Also we omit further procedures of joining clusters, refinement of the parameters, etc., that are present in the full automatic procedure.

Suppose we get some value of the probability of the background cluster $\tilde{p}_0$ using the formula (7). Denote $\tilde{p}_0' = \max(\tilde{p}_0, \varepsilon)$ and in (4) recalculate probabilities $\tilde{p}_i$, $i = 1, 2, \ldots, r$, proportionally, so that for recalculated probabilities the equation

$\sum_{i=0}^{r} \tilde{p}_i' = 1$ will hold. Let θ' be the corresponding parameter. Denote $\Delta_r^* = \Delta_r / \|\Delta_r\|_{L_1}$. We apply the EM procedure instead of (6a) using the formula

$$\pi(0, X) = \pi_{\theta'}(0, X) = \frac{\tilde{p}_0' \cdot \Delta_r^*(X, \theta')}{\tilde{f}_r(X, \theta') + \tilde{p}_0' \cdot \Delta_r^*(X, \theta')}, \quad X \in X^N. \tag{8}$$

Then we recalculate the probability of the background cluster using formula (7). To avoid calculation errors we repeat the procedure of recalculation of the probabilities. Clearly the inequality $\tilde{p}_0 \geqslant \varepsilon$ will hold.

After the Bayesian classification we can suppose that the vast majority of the outliers are assigned to the background (noise) cluster.

In practice we have a problem of selecting the proper value of ε. Presented computer simulation results show that the proposed method is not very sensitive to the selection of ε. We use Bayesian classification rule to assign the sample elements to the noise cluster or to the one of Gaussian clusters.

We have done computer simulation of the random noise by adding Gaussian component with small probability, zero mean and unit covariance matrix multiplied by sufficiently big constant to the various Gaussian mixture models, especially considered in [1] and [2].

In this paper we present three examples. In all examples we simulated the sample X^N with the sample size $N = 1000$ and applied the modified automatic procedure with the various values of ε. In the first two examples as basic mixture model we selected 3-dimensional Gaussian mixture model with three clusters with means $(-5, 0, 0)$, $(5, 5, 0,)$, $(0, -5, 0)$, equal probabilities and unit covariance matrices. In the third example as basic mixture model we selected 5-dimensional Gaussian mixture model with three clusters with means $(-5, 0, 0, 0, 0)$, $(5, 5, 0, 0, 0)$, $(0, -5, 0, 0, 0)$, equal probabilities and unit covariance matrices. In the first and the third example we selected noise level equal to 0.04 and corresponding covariance matrix equal to unit matrix multiplied by 40. In the second example we selected noise level equal to 0.01 and corresponding covariance matrix equal to unit matrix multiplied by 100. Given results are averages over 100 independent realizations.

We give number of extracted outliers (in per cent) and number of sample elements, which belong to the regular Gaussian clusters, which were classified as outliers (in per cent).

We can see that in the sufficiently wide range of ε the number of extracted outliers is close to the maximum available value. Beginning from the some value of ε (depending on mixture model and actual noise level) the number of extracted outliers increases very slowly. This can be a recommendation for choosing an appropriate ε in practice. Number of non-outliers classified as outliers is comparatively very small, so we loose a little information estimating Gaussian mixture parameters from the sample with removed sample elements classified as outliers. Note that some irregularities in the presented examples are caused not only by pure random factors but also by the different behaviour of the automatic procedure with different values of ε.

Fig. 1. **Number of extracted outliers** (per cent)

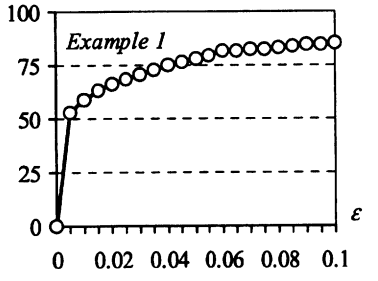


Fig. 2. **Number of non-outliers classified as outliers** (per cent)

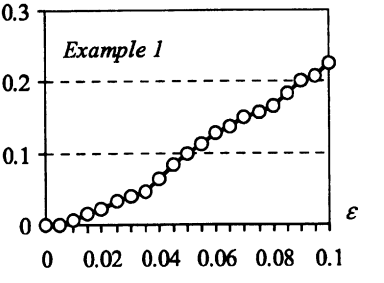


Fig. 3. **Number of extracted outliers** (per cent)

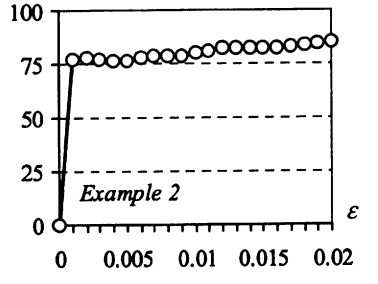


Fig. 4. **Number of non-outliers classified as outliers** (per cent)

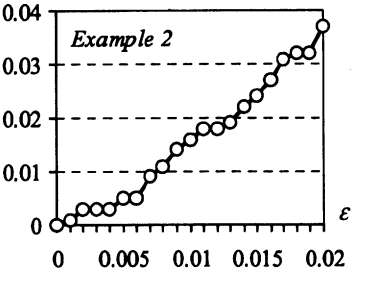


Fig. 5. **Number of extracted outliers** (per cent)

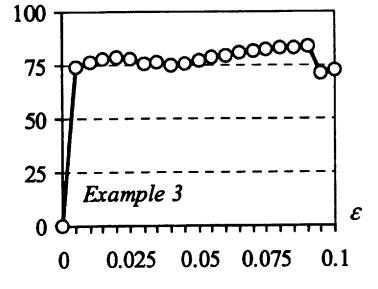
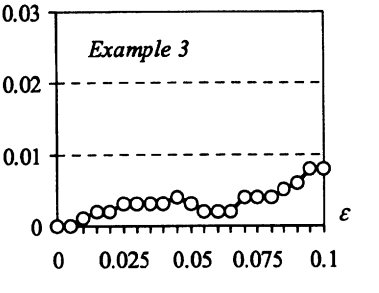


Fig. 6. **Number of non-outliers classified as outliers** (per cent)



Computer simulation results show that properly selected probability of the background cluster at the end of modified automatic procedure helps to assign the outliers to the noise cluster and easily remove them.

References

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

[2] G. Jakimauskas, Influence of the outliers to the classification of multidimensional Gaussian mixtures, *Liet. Matem. Rink.*, **41** (spec. nr.), 438–443 (2001).

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

[4] 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).

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

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

Procedūra išsiskiriančių stebėjimų išskyrimui iš imties, patenkinančios daugiamačio Gauso mišinio modelį

G. Jakimauskas

Nagrinėta procedūra išsiskiriančių stebėjimų išskyrimui iš imties, patenkinančios daugiamačio Gauso mišinio modelį. Straipsnis paremtas Gauso mišinių statistinio vertinimo matematiniais metodais, pasiūlytais R. Rudzkio ir M. Radavičiaus (1995), ypač nuoseklia pradinių reikšmių EM algoritmui parinkimo procedūra su papildomu tarpiniu triukšmo klasteriu. Pateikiami kompiuterinio modeliavimo rezultatai rodo, kad tam tikra minėtos procedūros modifikacija yra gerai tinkama išsiskiriančių stebėjimų išskyrimui iš imties.