

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Mindaugas Kavaliauskas

The Projection-based Statistical Analysis of the Multivariate Gaussian Distribution Mixture

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01 P)

Vilnius, 2005

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Mindaugas Kavaliauskas

The Projection-based Statistical Analysis of the Multivariate
Gaussian Distribution Mixture

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01 P)

Vilnius, 2005

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics in 1999–2004.

Scientific supervisor

Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics, 01 P)

The dissertation is defended at the Joint Council of the Mathematical Sciences of Vilnius Gediminas Technical University and the Institute of Mathematics and Informatics.

Chairman

Prof Dr Habil Kęstutis KUBILIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics, 01 P)

Members:

Prof Dr Habil Romanas JANUŠKEVIČIUS (Vilnius Pedagogical University, Physical Sciences, Mathematics, 01 P)

Dr Paulius MIŠKINIS (Vilnius Gediminas Technical University, Physical Sciences, Physics, 02 P)

Prof Dr Habil Alfredas RAČKAUSKAS (Vilnius University, Physical Sciences, Mathematics, 01 P)

Assoc Prof Dr Marijus RADAVIČIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics, 01 P)

Opponents:

Prof Dr Habil Vydas ČEKANAVIČIUS (Vilnius University, Physical Sciences, Mathematics, 01 P)

Assoc Prof Dr Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics, 01 P)

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics at Room 203 of the Institute of Mathematics and Informatics at 1 p.m. on January 19, 2005. Address: Akademijos 4, LT-08663, Vilnius, Lithuania.

Phone +370 5 2109300, fax +370 5 2729209.

The summary of the doctoral dissertation was distributed on December 17, 2005.

The dissertation is available at the libraries of Vilnius Gediminas Technical University and the Institute of Informatics and Mathematics.

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Mindaugas Kavaliauskas

Daugiamačių Gauso skirstinių mišinio statistinė analizė, taikant
duomenų projektavimą

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01 P)

Vilnius, 2005

Disertacija rengta 1999–2004 metais Matematikos ir informatikos institute.

Mokslinis vadovas

prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika, 01 P).

Disertacija ginama Vilniaus Gedimino technikos universiteto ir Matematikos ir informatikos instituto jungtinėje Matematikos mokslo krypties taryboje.

Pirmininkas

prof. habil. dr. Kęstutis KUBILIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika, 01 P).

Nariai:

prof. habil. dr. Romanas JANUŠKEVIČIUS (Vilniaus pedagoginis universitetas, fiziniai mokslai, matematika 01 P),

dr. Paulius MIŠKINIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, fizika, 02 P),

prof. habil. dr. Alfredas RAČKAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika, 01 P),

doc. dr. Marijus RADAIVIČIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika, 01 P).

Oponentai:

prof. habil. dr. Vydas ČEKANAVIČIUS (Vilniaus universitetas, fiziniai mokslai, matematika, 01 P),

doc. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika, 01 P).

Disertacija bus ginama viešame Matematikos mokslo krypties gynimo tarybos posėdyje 2005 m. sausio 19 d. 13 val. Matematikos ir informatikos instituto 203 kambaryje. Adresas: Akademijos 4, LT-08663, Vilnius.

Tel. 8 5 2109300, faks. 8 5 2729209.

Disertacijos santrauka išsiuntinėta 2004 m. gruodžio 17 d.

Disertaciją galima peržiūrėti Matematikos ir informatikos instituto ir Vilniaus Gedimino technikos universiteto bibliotekose.

Introduction

Problem of the dissertation. The Gaussian random values are very common in practice, because if a random value depends on many additive factors, according to the Central Limit Theorem (if particular conditions are satisfied), the sum is approximately from Gaussian distribution. If the observed random value belongs to one of the several classes, it is from the Gaussian distribution mixture model. The mixtures of the Gaussian distributions are common in various fields: biology, medicine, astronomy, military science and many others.

The most important statistical problems are problems of mixture identification and data clustering. In case of high data dimension, these tasks are not completely solved. The new parameter estimation of the multivariate Gaussian distribution mixture model and data clustering methods are proposed and analysed in the dissertation. Since it is much easier to solve these problems in univariate case, the projection-based approach is used.

The aim of the dissertation. The aim of this work is the development of constructive algorithms for distribution analysis and clustering of data from the mixture model of the Gaussian distributions.

The goals are as follows:

1. to explore whether the univariate data projections can be used to estimate the distribution density and to cluster the data from the mixture model of the Gaussian distributions.
2. to propose the constructive procedure for calculating the initial value of parameter, that ensures convergence of the recurrent EM algorithm to the maximum likelihood estimate.
3. to explore whether the usage of univariate data projections, in case of a small sample size, can help to estimate parameter of the Gaussian distribution mixture more precisely than the multivariate maximum likelihood estimate.

Scientific novelty of the results. The proposed projection-based model parameter identification and data clustering methods of the multivariate Gaussian distribution mixture do not have analogues in the works of other authors.

1. The new non-parametric estimation method of univariate distribution density is proposed and analysed.
2. The new procedure for estimation of the Gaussian distribution density based on data projecting and the inversion formula is proposed and analysed.
3. The new procedures for estimation of the Gaussian distribution density based on data projecting and the least squares method are proposed and analysed.

4. The new geometric data clustering procedure based on pseudo-distances between sample points is proposed. The calculation of pseudo-distances are based on the analysis of the projected data.
5. The new identification methodology of the Gaussian distribution mixture is proposed and analysed. This methodology embraces the ideas of data projection, the proposed geometrical clustering, EM algorithm and the least squares method. In case of small sample sizes for some model parameter values, the estimates calculated using this methodology are more precise (by mean of analysed errors) than the maximum likelihood estimator.

The main results:

1. The proposed non-parametric univariate distribution density estimation method based on adaptive kernel estimator and the multiplicative bias correction procedure estimates the density of the Gaussian distribution mixture more precisely (by means of analysed errors) than the procedures implemented in a popular statistical packages if smoothness of the density depends on the argument value.
2. The maximum likelihood estimator is the best estimator of classification probabilities in most cases, however the geometric clustering procedure (G) often estimates the Gaussian mixture density better in case of analysed (small) sample sizes, if the precision measure is the error in space L_2 .
3. The several-stage procedure ($G - EM$), which embraces statistical estimation of the univariate data projections parameters, geometric clustering based on pseudo-distance, and the EM algorithm is constructive and reliable way for calculation of the maximum likelihood estimator.
4. In case of a small sample, the procedure described in the third item and improved by the least squares method for projected densities (LSD) is more precise (by mean of all analysed errors) than the maximum likelihood estimator, for some values of parameter.

Approval of the results. There are 10 scientific papers published and 15 conference talks are given on the topic of the dissertation.

Also talks in seminars of the Applied Mathematics Department of Kaunas University of Technology and the Applied Statistics Department of the Institute of Mathematics and Informatics were given.

The structure of the dissertation. The dissertation consists of introduction, four chapters, conclusions, references and appendices. The first chapter reviews the papers of the other authors on the topic of the dissertation. In the second chapter

the problem of adaptive estimation of the univariate distribution density function is analysed. The estimation method is proposed. In the third chapter the proposed statistical analysis procedures of the multivariate Gaussian distribution density are proposed. The simulation methodology and results are presented in the fourth chapter.

Definitions

Kernel density estimator. Let X be a random value with an unknown distribution density function $f(x)$ and $\mathbb{X} = (X_1, \dots, X_n)$ be the independent sample of X . Then a *kernel density estimator* is defined

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - X_i), \quad (1)$$

where

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right), \quad \int K(x) dx = 1 \quad (2)$$

is the *kernel shape function*, and h is *bandwidth* of the kernel. $\hat{f}_h$ is the *fixed bandwidth kernel density estimator* if $h = \text{const}$, and $\hat{f}_h$ is the *adaptive kernel density estimator* if $h = h(x)$. The main problem of density estimation is selection of bandwidth h .

The bias and variance of a kernel density estimator are

$$b_h(x) = \frac{c_1 f''(x) h^2}{2} + o(h^2) \quad \text{and} \quad (3)$$

$$\sigma_h^2(x) = \frac{c_2 f(x)}{nh} + O\left(\frac{1}{n}\right), \quad (4)$$

where c_1 and c_2 depend on the kernel shape function and $h = h(n) \xrightarrow{n \rightarrow \infty} 0$. Therefore, the variable bandwidth

$$h_{AS}(x) = \left(\frac{c_2 f(x)}{(c_1 f''(x))^2 n} \right)^{1/5} \quad (5)$$

and the fixed bandwidth

$$h_{AS} = \left(\frac{c_2 f(x)}{c_1^2 \|f''\|_2^2 n} \right)^{\frac{1}{5}} \quad (6)$$

are asymptotically optimal by means of loss function

$$\mathbb{E} \left(\hat{f}_h(x) - f(x) \right)^2 = b_h^2(x) + \sigma_h^2(x). \quad (7)$$

Gaussian distribution mixture model. The random vector $X \in \mathbb{R}^d$ is from the

Gaussian (normal) distribution mixture (*Gaussian mixture*, for short), if its distribution density function $f(x)$ satisfies the equality

$$f(x) = \sum_{j=1}^q p_j \varphi_j(x) \stackrel{\text{def}}{=} f(x, \theta), \quad p_j > 0, \quad \sum_{j=1}^q p_j = 1. \quad (8)$$

Here q is the number of clusters, and $\varphi_j(x)$ is the Gaussian *distribution density function* (*density*, for short) with mean M_j and covariance matrix R_j . Let us denote the parameter of the model by $\theta = (p_j, M_j, R_j, j = 1, \dots, q)$.

Let us denote the sample which consists of the independent copies of a random vector X by $\mathbb{X} = \{X_1, X_2, \dots, X_n\}$. One of the most important statistical problems is the estimation of model parameter θ . When dimension of data d increases, the dimension of model parameter θ increases rapidly, and estimation of parameter becomes a complicated task.

In case of the Gaussian mixture model, the problem of data clustering is also an important statistical problem. Let us expand the definition of the model. Let (X, ν) be a random pair, where X is the observed properties vector of an observed object and ν is the unobserved class number of that object. If conditional density of X given $\nu = j$ is Gaussian density, then unconditional density $f(x)$ of random vector X satisfies (8), i.e. X satisfies the Gaussian mixture model. $p_j = \mathbb{P}\{\nu(X) = j\}$ is a priori probability of the cluster.

The *soft clustering* problem is an estimation problem of a posteriori classification probabilities

$$\pi(j, x) = \mathbb{P}\{\nu(x) = j | X = x\}, \quad j = \overline{1, q} \quad (9)$$

The *hard clustering* problem is an estimation problem of an observed object class number ν .

Using the *plug-in* approach, i.e. by replacing the unknown parameter θ with estimate $\hat{\theta}$, we obtain the estimates of the soft clustering probabilities

$$\hat{\pi}(j, x) = \frac{\hat{p}_j \hat{\varphi}_j(x)}{f(x, \hat{\theta})} \stackrel{\text{def}}{=} \pi_{\hat{\theta}}(j, x) \quad (10)$$

from the Bayes' formula.

The EM algorithm. It is natural to use the *MLE* (*maximum likelihood estimator*) to calculate estimate $\hat{\theta}_{MLE}$ because this estimator is asymptotically efficient in the case of Gaussian mixture. Unfortunately, the calculation problem of θ_{MLE} is complicated if data dimension d is high. Usually the *EM* (*expectation maximization*) algorithm is used to solve this task. The recurrent EM algorithm recalculates

values of $\hat{\pi}$ and $\hat{\theta}$ using formula (10) and

$$\hat{p}_j = \sum_{x \in \mathbb{X}} \frac{\hat{\pi}(j, x)}{n}, \quad \hat{M}_j = \sum_{x \in \mathbb{X}} \frac{\hat{\pi}(j, x)}{\hat{p}_j n} x, \quad \hat{R}_j = \sum_{x \in \mathbb{X}} \frac{\hat{\pi}(j, x)}{\hat{p}_j n} x x' - \hat{M}_j \hat{M}_j'. \quad (11)$$

Unfortunately, the EM algorithm converges locally and to $\hat{\theta}_{MLE}$ only if the initial parameter of this recurrent algorithm is close enough to $\hat{\theta}_{MLE}$. Other authors suggest to use random start, k -mean method, hierarchical clustering or other heuristic method for the initial value selection. However, the selection of this value is a crucial problem and it is not completely solved. One of the dissertation goals is selection of the parameter initial value.

The projection-based distribution analysis. Large dimension of data results in: complicated selection of initial parameter value of EM algorithm, not sufficiently expressed asymptotical properties of MLE, unidentifiable model due to dimension of a model parameter being larger than the sample size. In these cases the projection-based approach should be used.

Let $X_\tau = \tau' X$ be a univariate projection of a random vector X to direction τ . Let us denote the density function of X_τ by f_τ . Since, one-to-one correspondence

$$f \leftrightarrow \{f_\tau, \tau \in \mathbb{R}^d\} \quad (12)$$

exists, it is natural to estimate the multivariate density f using density estimates $\hat{f}_\tau$ of the univariate data projections.

Notice that in the case of the Gaussian mixtures, the projection is also a mixture of the (univariate) Gaussian distributions and

$$f_\tau(x) = \sum_{j=1}^q p_j(\tau) \varphi_{j,\tau}(x) \stackrel{def}{=} f_\tau(x, \theta(\tau)), \quad (13)$$

where $\varphi_{j,\tau}(x) = \varphi(x; m_j(\tau), \sigma_j^2(\tau))$ is the univariate Gaussian density. These equations show the relation between parameter θ of the multivariate Gaussian distribution mixture and parameter of this mixture univariate projections

$$p_j(\tau) = p_j, \quad m_j(\tau) = \tau' M_j, \quad \sigma_j^2(\tau) = \tau' R_j \tau. \quad (14)$$

Let us denote $\theta(\tau) = (p_j(\tau), m_j(\tau), \sigma_j^2(\tau)), j = 1, \dots, q$.

The Proposed Methods of Univariate Density Estimation

Method based on the second derivative of density estimator. The expression of optimal bandwidth

$$\begin{aligned} h(x) &= \arg \min_h \left(\widehat{b}_h^2(x) + \widehat{\sigma}_h^2(x) \right) \\ &= \arg \min_h \left(\frac{c_1 \lambda(x, h) h^2}{2} + \frac{c_2 \widehat{f}_h(x)}{nh} \right) \end{aligned} \quad (15)$$

follows from the asymptotical properties of a kernel density estimator. Here $\lambda(x, h)$ denotes $\widehat{f}_h''(x)$. Since

$$\begin{aligned} Q(y) &= F(x + y) - F(x - y) - \frac{y}{h}(F(x + h) - F(x - h)) \\ &= \frac{f''(x)(y^3 - yh^2)}{3} + o(yh^3), \end{aligned} \quad (16)$$

we propose the following formula for calculation of $\lambda(x, h)$:

$$\begin{aligned} \lambda(x, h) &= \arg \min_{\lambda} \max_i |a_i + \lambda b_i|, \\ \text{where } a_i &= \widehat{Q}(y_i), \quad b_i = -\frac{y_i^3 - y_i h^2}{3}, \end{aligned} \quad (17)$$

where y_i is a discontinuity point of empirical distribution function $\widehat{F}(x)$ in interval $[x - h; x + h]$, and

$$\widehat{Q}(y) = \widehat{F}(x + y) - \widehat{F}(x - y) - \frac{y}{h} \left(\widehat{F}(x + h) - \widehat{F}(x - h) \right). \quad (18)$$

The plug-in method. Since

$$\begin{aligned} b_h(x) &= \int K(y) f(x + hy) dy - f(x) \\ &= \int (f(x + hy) - f(x)) K(y) dy, \end{aligned} \quad (19)$$

we propose the plug-in bias estimate

$$\widehat{b}_{h,\Delta}(x) = \int (\widehat{f}_{\Delta}(x + hy) - \widehat{f}_{\Delta}(x)) K(y) dy, \quad (20)$$

where $\hat{f}_\Delta$ is also a kernel density estimator with bandwidth $\Delta > h$. The optimal bandwidth value can be calculated

$$h(x) = \arg \min_h \left(\hat{b}_{h,\Delta}^2(x) + \hat{\sigma}_h^2(x) \right) = \arg \min_h \left(\hat{b}_{h,\Delta}^2(x) + \frac{c_2 \hat{f}_h(x)}{nh} \right). \quad (21)$$

The simulation analysis showed

$$\left| \hat{b}_{h,\Delta}(x) \right| = \int_0^1 \left| \hat{f}_\Delta(x + hy) - \hat{f}_\Delta(x - hy) + 2\hat{f}_\Delta(x) \right| K(y) dy \quad (22)$$

to be better than (20) for calculation of the bias estimate.

The proposed kernel density estimator with bandwidth selected using (21), was improved by bandwidth smoothing

$$\bar{h}(x) = \frac{\sum_{i=1}^n h(X_i) K_h(x - X_i)}{\sum_{i=1}^n K_h(x - X_i)}, \quad (23)$$

multiplicative bias correction

$$\tilde{f}(x) = \frac{\hat{f}_h(x)}{n} \sum_{i=1}^n \frac{K_h(x - X_i)}{\hat{f}_h(X_i)} \quad (24)$$

and density estimate smoothing

$$\bar{\bar{f}}_h(x) = \frac{\sum_{i=1}^n \hat{f}_h(X_i) K_{h^*}(x - X_i)}{\sum_{i=1}^n K_{h^*}(x - X_i)} \quad (25)$$

procedures, where $h^* = h^*(x) = \beta h(x)$, β is parameter of the method.

The Proposed Methods of Multivariate Gaussian Mixture Analysis

The inversion formula based method (I method). The density function f can be obtained from the characteristic function using the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-it'x} \psi(t) dt, \quad (26)$$

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

Let us denote the characteristic functions projected data by ψ_τ . Since there is a one-to-one correspondence between densities and characteristic functions, and

$$\psi(t) = \mathbb{E}e^{it'X} = \mathbb{E}e^{i|t|\tau'X} = \psi_\tau(|t|), \quad \text{where } \tau = t/|t|, \quad (27)$$

we obtain an equality

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:|\tau|=1} ds \int_0^\infty e^{-i u \tau' x} \psi_\tau(u) u^{d-1} du, \quad (28)$$

using the inversion formula (26) and performing the transformation to a spherical coordinate system. The first integral in expression (28) denotes the surface integral over unit sphere. By using the estimates of ψ_τ and replacing the surface integral by a sum, we obtain the formula for calculation of an estimate $\hat{f}$

$$\hat{f}(x) = \frac{c(d)}{\|T\|} \sum_{\tau \in T} \int_0^\infty e^{-i u \tau' x} \hat{\psi}_\tau(u) u^{d-1} e^{-h u^2} du, \quad (29)$$

where $\|T\|$ denotes a size of set T . The set T is a finite set of points, uniformly distributed on the unit sphere, $c(d) = \frac{d 2^{-d} \pi^{-\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$, where Γ denotes the Gamma function. The extra multiplier $e^{-h u^2}$ is used for additional smoothing of a density estimate.

The estimates of characteristic functions ψ_τ can be obtained from the parametric estimation of the univariate projected densities f_τ

$$\hat{\psi}_\tau(u) = \sum_{j=1}^{\hat{q}_\tau} \hat{p}_j(\tau) e^{i u \hat{m}_j(\tau) - u^2 \hat{\sigma}_j^2(\tau)/2}. \quad (30)$$

Formula (29) could be used to estimate not only a whole density $f(x)$, but also a density $\varphi_j(x)$ of a single mixture component. In this case, $\hat{\psi}_\tau(u)$ should be replaced by the characteristic function of corresponding component in the formula.

The least squares method for projected parameters (*LSP method*). Since equalities (14) exist, the parameter estimate of a multivariate mixture can be obtained from the parameter estimates of the univariate distributions using the least squares approach.

$$\begin{aligned}\hat{p}_j &: \sum_{\tau} (p_j - \hat{p}_j(\tau))^2 \longrightarrow \min, \\ \widehat{M}_j &: \sum_{\tau} (\tau' M_j - \widehat{m}_j(\tau))^2 \longrightarrow \min, \\ \widehat{R}_j &: \sum_{\tau} (\tau' R_j \tau - \widehat{\sigma}_j^2(\tau))^2 \longrightarrow \min.\end{aligned}\tag{31}$$

Estimates $\hat{p}_j$, $\widehat{M}_j$ and $\widehat{R}_j$ can be calculated using common technique.

Let us notice, that *I* and *LSP* methods can be used for data clustering only if cluster numbering is compatible among all projections.

The least squares method for projected densities (*LSD method*). Let us define parameter estimate what would minimize distance between the projected densities

$$\widehat{\theta}: \sum_{\tau} \left\| f_{\tau}(\cdot, \theta) - \widehat{f}_{\tau}(\cdot) \right\|_2^2 \longrightarrow \min,\tag{32}$$

where $f_{\tau}(\cdot, \theta)$ is a parametric expression of the density defined by (13), and $\widehat{f}_{\tau}(\cdot)$ is any estimate of projected data density.

The latter equation could be rewritten

$$\widehat{\theta}: \sum_{\tau} \int_{\mathbb{R}} \left(f_{\tau}^2(x, \theta) - 2f_{\tau}(x, \theta)\widehat{f}_{\tau}(x) \right) dx \longrightarrow \min.\tag{33}$$

Here $\widehat{f}_{\tau}^2(x)$ is omitted because it does not depend on minimization argument θ .

In the analysed case either the parametric estimate $\widehat{f}_{\tau}(\cdot)$ can be used or the integral of second component can be replaced by the sum over projections of observations. We selected the latter case, because in this case there is no need to estimate

the distribution parameters of the projected data. Hence, equation (33) is rewritten

$$\hat{\theta} : Q(\theta) = \sum_{\tau} \left(\int_{\mathbb{R}} f_{\tau}^2(x, \theta) dx - \frac{2}{n} \sum_{j=1}^n f_{\tau}(\tau' X_j, \theta) \right) \longrightarrow \min. \quad (34)$$

All unknown model parameters could be grouped into three groups

$$\theta = (\mathbf{p} \quad \mathbf{u} \quad \mathbf{v})', \quad (35)$$

where $\mathbf{p}$ is a vector of clusters probabilities, $\mathbf{u}$ is a vector of all mean parameters, and $\mathbf{v}$ is a vector of all parameters of covariance matrices from all clusters. Since equality

$$\int_{\mathbb{R}} \varphi(x; m_1, \sigma_1^2) \varphi(x; m_2, \sigma_2^2) dx = \varphi(m_1 - m_2; 0, \sigma_1^2 + \sigma_2^2), \quad (36)$$

exists, the function (34) has a shape

$$Q(\theta) = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) \varphi_{\mathbf{i}}(\mathbf{u}, \mathbf{v}) \longrightarrow \min, \quad (37)$$

where $\mathbf{i}$ is vector index consisting of projections $\tau \in T$, clusters $i, j \in \{1, \dots, q\}$ and sample points $l \in \{1, \dots, n\}$, i.e. $\mathbf{i} \in T \otimes \{1, \dots, q\} \otimes \{1, \dots, q\} \otimes \{1, \dots, n\}$. Let us notice, that

$$\varphi_{\mathbf{i}} = \frac{1}{\sqrt{\beta_{\mathbf{i}}}} \exp \left(-\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}} \right), \text{ where } \alpha_{\mathbf{i}} = ((\mathbf{u} - a_{\mathbf{i}})^T c_{\mathbf{i}})^2, \beta_{\mathbf{i}} = b_{\mathbf{i}}^T \mathbf{v}. \quad (38)$$

Here $a_{\mathbf{i}}$, $b_{\mathbf{i}}$ and $c_{\mathbf{i}}$ are vectors dependent only on sample data and projection directions, i.e. these vectors are constant when sample is fixed and projection directions are selected.

Let us calculate derivatives of function $\varphi_{\mathbf{i}}$:

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{u}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}} \alpha'_{\mathbf{i}} = A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}), \quad \text{where } A_{\mathbf{i}} = -\frac{\varphi_{\mathbf{i}}}{\beta_{\mathbf{i}}} [c_{\mathbf{i}} c_{\mathbf{i}}^T], \quad (39)$$

$$\frac{d\varphi_{\mathbf{i}}}{d\mathbf{v}} = \varphi_{\mathbf{i}} \left(\frac{\alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} - \frac{1}{2\beta_{\mathbf{i}}} \right) \beta'_{\mathbf{i}} = d_{\mathbf{i}} - B_{\mathbf{i}} \mathbf{v}, \quad (40)$$

$$\text{where } d_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}} \alpha_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} b_{\mathbf{i}}, \quad B_{\mathbf{i}} = \frac{\varphi_{\mathbf{i}}}{2\beta_{\mathbf{i}}^2} [b_{\mathbf{i}} b_{\mathbf{i}}^T].$$

Hence derivatives of Q are

$$\frac{dQ}{d\mathbf{u}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}}(\mathbf{u} - a_{\mathbf{i}}) = C\mathbf{u} - e, \quad (41)$$

$$\text{where } C = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}}, \quad e = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) A_{\mathbf{i}} a_{\mathbf{i}},$$

and

$$\frac{dQ}{d\mathbf{v}} = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) (d_{\mathbf{i}} - B_{\mathbf{i}}\mathbf{v}) = g - D\mathbf{v}, \quad (42)$$

$$\text{where } D = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) B_{\mathbf{i}}, \quad g = \sum_{\mathbf{i}} w_{\mathbf{i}}(\mathbf{p}) d_{\mathbf{i}}.$$

Here $a_{\mathbf{i}}, b_{\mathbf{i}}, c_{\mathbf{i}}, d_{\mathbf{i}}, e$ and g are vectors, $A_{\mathbf{i}}, B_{\mathbf{i}}, C$ and D are matrices. Since

$$\hat{\theta} : Q(\theta) \longrightarrow \min \quad \Rightarrow \quad \frac{dQ}{d\mathbf{p}}(\hat{\theta}) = 0, \quad \frac{dQ}{d\mathbf{u}}(\hat{\theta}) = 0, \quad \frac{dQ}{d\mathbf{v}}(\hat{\theta}) = 0, \quad (43)$$

the recurrent procedure could be used to calculate $\hat{\theta}$

$$\mathbf{u}^{(k+1)} = C^{-1}(\theta^{(k)}) e(\theta^{(k)}), \quad \mathbf{v}^{(k+1)} = D^{-1}(\theta^{(k)}) g(\theta^{(k)}). \quad (44)$$

The vector of probabilities $\mathbf{p}$ is calculated in each step using common least squares technique because $\frac{dQ}{d\mathbf{p}}$ is the linear function of $\mathbf{p}$.

Geometric clustering procedure (G method). 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 general case, this function does not satisfy the neither the triangular inequality nor $\rho(x, x) = 0$. Let us divide the set of observation numbers $N \stackrel{\text{def}}{=} \{1, \dots, n\}$ into non-intersecting subsets $\hat{K}_1, \dots, \hat{K}_q$ so as to minimize the value of the functional

$$Q(K_1, \dots, K_q) = \sum_{j=1}^q \frac{1}{\|K_j\|} \sum_{s,r \in K_j} \rho(X(s), X(r)). \quad (45)$$

The main task is selection of the distance function ρ . Since the problem of data clustering is being solved, the function ρ should express belonging of data to the same cluster, but not an Euclidean distance between the points. Let us use pseudo-distance function based on the estimates of classification probabilities $\pi_{\tau}(j, x) = \mathbb{P}\{\nu = j | \tau'X = \tau'x\}$ of projected data. These estimates can be calculated using

distribution parameters of sample projections

$$\hat{\pi}_\tau(j, x) = \frac{\hat{p}_j(\tau) \hat{\varphi}_{j,\tau}(x)}{f_\tau(x, \hat{\theta}(\tau))}. \quad (46)$$

We selected and analysed two functions of pseudo-distance between univariate data projections. A function

$$\rho_\tau(x, y) = \sum_{j=1}^{\hat{q}_\tau} |\hat{\pi}_\tau(j, x) - \hat{\pi}_\tau(j, y)| \quad (47)$$

and a function which estimates the probability of belonging to the same class

$$\rho_\tau(x, y) = 1 - \sum_{j=1}^{\hat{q}_\tau} \hat{\pi}_\tau(j, x) \hat{\pi}_\tau(j, y), \quad (48)$$

because

$$1 - \sum_{j=1}^q \pi_\tau(j, x) \pi_\tau(j, y) = \mathbb{P}\{\nu(X) \neq \nu(Y) | X = x, Y = y\}. \quad (49)$$

The pseudo-distance between a multivariate sample data is based on the defined pseudo-distance function of univariate data projections. Since clusters will overlap in some directions and the sample points of multivariate data will be assigned to the same cluster in a projection, such multivariate pseudo-distance function can be selected

$$\rho(x, y) = \max_{\tau} \rho_\tau(\tau'x, \tau'y). \quad (50)$$

However this definition of pseudo-distance is not robust. We suggest to calculate the maximum after discarding some α part of directions with the highest values of pseudo-distances between projected data.

Another way to make pseudo-distance more robust is to use average value

$$\rho^\beta(x, y) = \frac{1}{\|T\|} \sum_{\tau} \rho_\tau^\beta(\tau'x, \tau'y) \quad (51)$$

instead of maximum. In this case, β is the parameter of the method. It can be selected using the likelihood function.

Thus, the geometric hard clustering procedure consists of the following stages:

- selection of a projection directions set T ;
- calculation of the maximum likelihood estimates

$$\hat{\theta}(\tau) = \arg \max_{\theta(\tau)} \sum_{s=1}^n \log f_{\tau}(\tau' X(s), \theta(\tau)) \quad (52)$$

or their approximation for each $\tau \in T$;

- estimation of a posteriori probabilities by means of the "plug-in" technique

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

and calculation of pseudo-distance matrix;

- calculation of a set of clusters

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

Despite similarity of this proposed method and the k -mean method, the loss functional (45) depends only on distance values between data points. Distance values between data point and centre of the cluster is not used. This particularity of loss functional disabled us to use standard minimization procedure. Therefore functional minimization procedure was developed.

Methodology of Analysis and Results

The proposed methods were analysed by means of simulation. In this case, the theoretical value of distribution density and parameters are known, and the errors of estimates can be calculated.

The random sample was selected from the mixture of Gaussian distribution. Various values of parameter were selected to explore the properties of the proposed methods.

Analysis of the univariate density estimation methods. The univariate density estimators were compared by means of errors

$$\varepsilon_1(\hat{f}) = \|\hat{f} - f\|_1, \quad (53)$$

$$\varepsilon_2(\hat{f}) = \|\hat{f} - f\|_2. \quad (54)$$

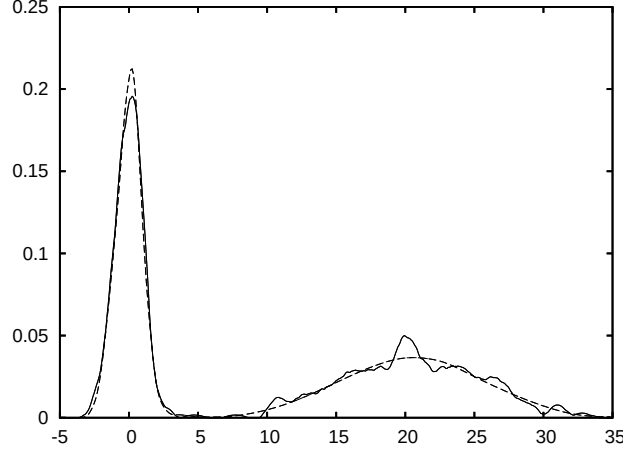


Figure 1. Comparison plot of the distribution density estimates. Solid line represents the plug-in estimator with additional procedures applied. Dashed line represents the fixed bandwidth estimator with bandwidth selected using well-known cross-validation method.

The Epanichnikov kernel shape function

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2), & \text{if } |x| < 1 \\ 0, & \text{if } |x| \geq 1 \end{cases} \quad (55)$$

was used.

The simulation analysis showed that the proposed univariate density estimation method is not stable but the plug-in estimate was precise enough. Additional procedures of bandwidth smoothing, multiplicative bias correction and density estimate smoothing increased precision of the proposed procedure significantly. In the analysed cases, after application of the mentioned procedures, error ε_1 decreased 1.680 times and error ε_2 decreased 1.519 times on the average. The most effective procedure was multiplicative bias correction.

Study of the multivariate Gaussian mixture analysis methods. Various combinations of the proposed methods were analysed, for example, $G - EM$, $G - LSP - EM$, $G - EM - LSD$, etc. The classification precision of statistical procedures and their combinations were measured by errors

$$\varepsilon_\pi = \mathbb{E} \sum_{i=1}^q |\hat{\pi}(j, X) - \pi(j, X)|, \quad (56)$$

$$\varepsilon_\nu = \mathbb{P}\{\hat{\nu}(X) \neq \nu(X)\}. \quad (57)$$

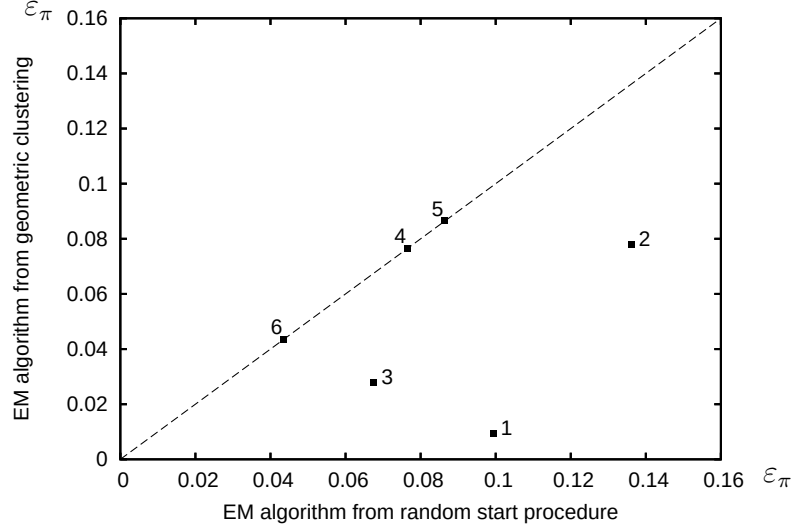


Figure 2. Plot of the random start procedure and geometric clustering procedure. Both methods were used to calculate an initial value of EM algorithm. The classification errors are presented. Mixtures 1, 2 and 3 represent the case of significantly different clustering probabilities¹.

The errors

$$\varepsilon_{L2} = \mathbb{E} \int_{\mathbb{R}^d} (f_{\hat{\theta}}(x) - f_{\theta}(x))^2 dx, \quad (58)$$

$$\varepsilon_{L2F} = \mathbb{E} \int_{\mathbb{R}^d} (f_{\hat{\theta}}(x) - f_{\theta}(x))^2 f(x) dx \quad (59)$$

were used to measure the precision of parameter estimation.

Simulation study showed that I method estimate of multivariate density is not smooth enough. Therefore, the additional smoothing multiplier e^{-hu^2} was used. The smoothing parameter h was selected using the likelihood function. Despite of smoothing, the value of ε_2 errors were larger than the errors of the other methods, but value of ε_{2F} errors were similar to the errors of the other methods. Hence, this method can be used to estimate the density in region where density value is high.

EM algorithm with initial value calculated using LSP method converges to the MLE estimator in most cases. But this method requires a compatible cluster numbering among all projections. Therefore, it can be used only in combination with other procedures in applications.

¹Probabilities of mixtures clusters. Mixture 1: $q = 2, p_1 = 0.92, p_2 = 0.08$. Mixture 2: $q = 3, p_1 = p_3 = 0.05, p_2 = 0.9$. Mixture 3: $q = 2, p_1 = 0.9, p_2 = 0.1$. Mixture 4: $q = 4, p_1 = p_3 = 0.2, p_2 = p_4 = 0.3$. Mixture 5: $q = 3, p_1 = p_3 = 0.35, p_2 = 0.3$. Mixture 6: $q = 4, p_1 = p_2 = p_3 = p_4 = 0.25$.

LSP method does not require the compatible cluster numbering, but it is a recurrent method and it requires the initial value of parameter (alike the EM algorithm). A combination of methods $G - EM - LSD$ is more precise than the MLE (by means of all calculated errors) in some analysed cases of a small sample.

G method neither requires a compatible cluster numbering nor an initial value of mixture parameter. This method calculates the initial data clustering and helps to obtain both: a compatible cluster numbering and an initial value of mixture parameter for the other methods. This is the base procedure for all other proposed methods. It is precise enough and in most cases G ensures convergence of EM algorithm to the MLE.

Since the problem of an initial parameter value selection exists, many statistical packages use the random start procedure to obtain an initial parameter value of the EM algorithm. The comparison study of the random start procedure and geometric clustering procedure was performed. Even if a very large number of random start points is used (we tested with 4000 start points), the EM algorithm does not converge to the MLE if the probabilities of the clusters differ significantly. The geometric clustering procedure improved by the EM algorithm does not have this drawback. See, Fig. 2.

Conclusions

1. The proposed non-parametric univariate distribution density estimation method estimates the density of the Gaussian distribution mixture more precisely (by means of analysed errors) than the procedures implemented in a popular statistical packages if smoothness of the density depends on the argument value.
2. The proposed Gaussian distribution mixture identification methodology based on data projection is constructive and more precise than the other constructive algorithms for calculation of the maximum likelihood estimator.
3. In case of a small sample, the projection-based statistical procedure which embraces geometric data clustering, the EM algorithm and the least squares method for projected densities enables to estimate the parameter of the Gaussian distribution mixture more precisely than the maximum likelihood estimator.
4. It is advisable to use a projection-based data clustering to solve the Gaussian distribution mixture identification and data clustering problems.

The List of Publications

Publications in journals of the journal list of the Institute of Scientific Information

1. Kavaliauskas M., Rudzkis R. (2005), Multivariate Data Clustering for the Gaussian Mixture Model, *Informatica*, **16** (1), Vilnius, MII. [ISSN 0868-4952]

Publications in journals of the journal list of the Department of Science and Study²

1. Kavaliauskas M. (1997), An Adaptive Choice of Smoothing Width in Statistical Estimation of Distribution Density, *Lietuvos matematikų draugijos XXXVII konferencijos darbai. Specialus Lietuvos matematikos rinkinio priedas*, Vilnius, Technika, pp. 198–203, (in Lithuanian). [ISBN 9986-05-350-1]
2. Rudzkis R., Kavaliauskas M. (1998), On Local Bandwidth Selection for Density Estimation, *Informatica*, **9** (4), Vilnius, MII, pp. 171–178. [ISSN 0868-4952]
3. Kavaliauskas M. (2000), Kernel Distribution Density Estimation Based on Cross-Validation, *Lietuvos matematikos rinkinys, spec. nr.*, **40**, Vilnius, MII, pp. 290–295, (in Lithuanian). [ISSN 0132-2818]
4. Kavaliauskas M., Rudzkis R. (2002), Projection-based Estimation of Multivariate Distribution Density, *Lietuvos matematikos rinkinys, spec. nr.*, **42**, Vilnius, MII, pp. 537–540. [ISSN 0132-2818]

Publications in the other peer reviewed international and foreign journals

1. Kavaliauskas M. (1999), Adaptive Density Estimation Method, *Proc. of 11th European Young Statisticians Meeting*, Marly-le-Roi, INRA, pp. 112–116.
2. Kavaliauskas M. (2001), An Adaptive Kernel Density Estimation Method, *Computer Data Analysis and Modelling: Robustness and Computer Intensive Methods. Proc. of 6th Intl. Conf.*, **1**, Minsk, BSU, pp. 171–178. [ISBN 985-445-490-8 (Vol.1), ISBN 985-445-489-4]

²Department of Science and Study is the department of the Ministry of Education and Science of the Republic of Lithuania.

3. Kavaliauskas M., Rudzkis R. (2004), The Projection-based Estimation of the Gaussian Mixture Parameters, *Computer Data Analysis and Modelling: Robustness and Computer Intensive Methods. Proc. of 7th Intl. Conf.*, **1**, Minsk, BSU, pp. 160–164. [ISSN 985-445-492-4]
4. Kavaliauskas M., Rudzkis R., Ruzgas T. (2004), The Projection-based Multivariate Distribution Density Estimation, *Acta et Commentationes Universitatis Tartuensis de Mathematica*, **8**, Tartu, Tartu University Press, pp. 1–7. [ISSN 1406-2283]

Publications in other journals

1. Kavaliauskas M., Ruzgas T. (2004), Application Nonparametric Methods for Clustering Gaussian Density Mixture Models, *Matematika ir matematinis modeliavimas. Konferencijos pranešimų medžiaga*, Technologija, Kaunas, pp. 14–19, (in Lithuanian). [ISBN 9955-09-616-0]

The List of Conference Talks

1. Kavaliauskas M., An Adaptive Choice of Smoothing Width in Statistical Estimation of Distribution Density, *Lietuvos matematikų draugijos XXXVIII konferencija*, VGTU, Vilnius, 1997, (in Lithuanian).
2. Kavaliauskas M., Rudzkis R., Adaptive Distribution Density Estimates, *Lietuvos matematikų draugijos XXXIX konferencija*, KMA, Kaunas, 1998, (in Lithuanian).
3. Kavaliauskas M., Smoothing Bandwidth Selection Methods in Estimation of the Distribution Density, *Lietuvos matematikų draugijos XL konferencija*, MII, Vilnius, 1999, (in Lithuanian).
4. Kavaliauskas M., Adaptive Density Estimation Method, *11th European Young Statisticians Meeting*, Marly-le-Roi, INRA, 1999.
5. Kavaliauskas M., Kernel Distribution Density Estimation Based on Cross-Validation, *Lietuvos matematikų draugijos XLI konferencija*, ŠU, Šiauliai, 2000, (in Lithuanian).

6. Kavaliauskas M., Estimation of the Multivariate Density Using Projections, *Lietuvos matematikų draugijos XLII konferencija*, KU, Klaipėda, 2001, (in Lithuanian).
7. Kavaliauskas M., An Adaptive Kernel Density Estimation Method, *Computer Data Analysis and Modelling: Robustness and Computer Intensive Methods*, BSU, Minsk, BSU, 2001.
8. Kavaliauskas M., Rudzkis R., Projection-based Estimation of Multivariate Distribution Density, *Lietuvos matematikų draugijos XLIII konferencija*, LKA, Vilnius, 2002.
9. Rudzkis R., Kavaliauskas M., Projection-based Multivariate Distribution Density Estimation, *8th Vilnius Conference on Probability Theory*, VGTU, Vilnius, 2002.
10. Kavaliauskas M., Rudzkis R., Ruzgas T., Estimation of the Multivariate Gaussian Distribution Mixture Parameters Using Univariate Projections, *Lietuvos matematikų draugijos XLIV konferencija*, VPU, Vilnius, 2003, (in Lithuanian).
11. Kavaliauskas M., Rudzkis R., Ruzgas T., The Projection-based Multivariate Distribution Density Estimation, *The 7th Tartu Conference on Multivariate Statistics. Satellite meeting of ISI 54th session in Berlin*, University of Tartu, Tartu, 2003.
12. Kavaliauskas M., Ruzgas T., Clustering of the Gaussian Distribution Mixture Using Nonparametric Methods, *Matematika ir matematinis modeliavimas*, Kauno technologijos universitetas, 2004, (in Lithuanian).
13. Kavaliauskas M., Rudzkis R., Usage of Univariate Projections in the Case of the Multivariate Gaussian Mixture Model, *Lietuvos matematikų draugijos XLV konferencija*, LŽŪU, Kaunas, 2004, (in Lithuanian).
14. Rudzkis R., Kavaliauskas M., Usage of Univariate Projections in Classification Algorithms Based on the Mixture Model of Multivariate Distributions, *6th Intl. School-Workshop on Multidimensional Analysis and Econometrics*, Tsakhkadzor, Armenia, 2004, (plenary talk, in Russian).
15. Kavaliauskas M., Rudzkis R., The Projection-based Estimation of the Gaussian Mixture Parameters, *Computer Data Analysis and Modelling: Robustness and Computer Intensive Methods*, BSU, Minsk, 2004.

About the Author

M. Kavaliauskas was born in Kaunas in 1975. In 1981, he entered the primary school in Kaunas. He was the winner of Lithuanian Young Mathematician Olympiad in 1991 and 1993, and a participant of the 33rd (Moscow) and 34th (Istanbul) International Mathematical Olympiads. In 1993, he graduated from Gymnasium of Kaunas University of Technology. He entered Kaunas University of Technology and graduated from the Faculty of Fundamental Sciences with Bachelor of Sciences degree in Applied Mathematics in 1997. He continued studies and graduated the same university with Master of Sciences degree in Applied Mathematics in 1999. From 1999 he continues studies at the Institute of Mathematics and Informatics to acquire a PhD degree in Mathematics.

Between 1997 and 1999, he was engineer at Statistics Centre of Vytautas Magnus University and the Institute of Mathematics and Informatics. From 2004 he is a scientific fellow of Kaunas University of Technology.

In scientific research his fields of interest are non-parametric density estimation and data clustering.

Mindaugas Kavaliauskas

**Daugiamačių Gauso skirstinių mišinio statistinė analizė,
taikant duomenų projektavimą**

(daktaro disertacijos reziumė)

Tiriama problema ir aktualumas. Gauso atsitiktiniai dydžiai ypač dažnai sutinkami praktikoje. Tais atvejais, kai stebimas Gauso atsitiktinis dydis priklauso vienai iš kelių klasių, stebėsime dydį, tenkinantį *Gauso skirstinių mišinio* (toliau, trumpumo dėlei *Gauso mišinio*) modelį. Gauso skirstinių mišiniai sutinkami įvairiose srityse: biologijoje, medicinoje, astronomijoje, karyboje ar net tekstilėje. Dėl dažno Gauso skirstinių mišinio panaudojimo taikomuosiuose klasterizavimo tyrimuose, šio modelio statistinio identifikavimo uždavinys yra ypač aktualus, tačiau iki šiol nėra iki galo išspręstas.

Parametrų įverčiams rasti galima taikyti *maksimalaus tikėtino metodo* (angl. *maximum likelihood method*, toliau *MTM*), deja, tokių įverčių apskaičiavimas yra rimta praktinė problema daugiamačiu atveju. Didėjant duomenų dimensijai ir klasterių kiekiui, mišinio parametrų matavimų skaičius greitai auga. Klasikinių optimizavimo metodų taikymas neduoda gerų rezultatų. Todėl dažniausiai mišinių analizėje parametrų vertinimui ir imties klasterizavimui naudojamas rekurentinis *tikėtino maksimalizavimo* (angl. *expectation maximization*, toliau *EM*) algoritmas. EM algoritmas maksimizuoja tikėtino funkciją, todėl gali būti naudojamas MTM įvertiniui apskaičiuoti. Deja, tikėtino funkcija turi daug lokalių ekstremumų, o EM algoritmas konverguoja lokaliai. Taigi MTM įvertinį galima pasiekti tik parinkus pradinę vertinamo parametro reikšmę pakankamai arti tikrosios mišinio parametro reikšmės. Būtent šios pradinės reikšmės parinkimas vis dar yra aktuali, galutinai neišspręsta problema, kuri ir buvo tiriama.

Parametrų įvertinimo ir duomenų klasterizavimo uždaviniai vienamačiam Gauso skirstinių mišiniui yra paprastesni ir juos galima konstruktyviai išspręsti. Todėl disertacijoje buvo taikomas duomenų projektavimas ir bandoma daugiamačio mišinio analizės uždavinius suvesti į suprojektuoto vienamačio mišinio analizės uždavinio sprendimą įvairiose projektavimo kryptyse.

Darbo tikslas: konstruktyvių algoritmų, skirtų daugiamačių duomenų, tenkinančių Gauso skirstinių mišinio modelį, pasiskirstymo analizei ir klasifikavimui, sukūrimas. *Darbo uždaviniai:* ištirti vienamačių duomenų projekcijų panaudojimo galimybę daugiamačių Gauso skirstinių mišinio pasiskirstymo tankiui vertinti bei duomenims klasterizuoti; sukurti procedūrą, pakankamai tiksliai parenkančią pradinę parametro reikšmę ir užtikrinančią rekurentinio EM algoritmo konvergavimą į MTM įvertį.

Tyrimų naujumas. Pasiūlytas ir ištirtas naujas adaptyvus neparametrinis vienamačio pasiskirstymo tankio įvertis ir keturios naujos procedūros, skirtos daugiamatį Gauso skirstinių mišinio parametrus identifikuoti ir duomenims klasterizuoti. Pasiūlyti daugiamatį Gauso skirstinių mišinio statistinės analizės metodai išsiskiria iš kitų žinomų metodų duomenų projektavimo panaudojimu.

Siūlomi statistiniai metodai. Disertaciniame darbe pasiūlyti ir ištirti šie nauji statistiniai metodai:

- **vienamačio pasiskirstymo adaptyvaus tankio įvertinimo procedūra;**
- **Gauso mišinio tankio įvertis, paremtas apvertimo formule;** Siūlomas pasiskirstymo tankio įvertis

$$\hat{f}(x) = \frac{c(d)}{|T|} \sum_{\tau \in T} \int_0^{\infty} e^{-iu\tau'x} \hat{\psi}_{\tau}(u) u^{d-1} e^{-hu^2} du, \quad (60)$$

čia $c(d)$ yra konstanta priklausanti nuo stebėjimų dimensijos d , T — pasirinktų projektavimo krypčių aibė, $\hat{\psi}_{\tau}(u)$ — duomenų projekcijos į kryptį τ charakteristinė funkcija, h — glodinimo parametras.

- **mažiausių kvadratų metodas projektuotų duomenų parametrų.** Siūlomas daugiamatį Gauso mišinio parametrų įvertis

$$\begin{aligned} \hat{p}_j &: \sum_{\tau} (p_j - \hat{p}_j(\tau))^2 \longrightarrow \min, \\ \hat{M}_j &: \sum_{\tau} (\tau' M_j - \hat{m}_j(\tau))^2 \longrightarrow \min, \\ \hat{R}_j &: \sum_{\tau} (\tau' R_j \tau - \hat{\sigma}_j^2(\tau))^2 \longrightarrow \min, \end{aligned} \quad (61)$$

čia $\hat{p}_j(\tau)$, $\hat{m}_j(\tau)$ ir $\hat{\sigma}_j^2(\tau)$ yra projektuotų į kryptį τ duomenų parametrų įvertiniai;

- **mažiausių kvadratų metodas projektuotų duomenų tankiams.** Siūlomas daugiamatį Gauso mišinio parametrų įvertis

$$\hat{\theta}: \sum_{\tau} \left\| f_{\tau}(\cdot, \theta) - \hat{f}_{\tau}(\cdot) \right\|_2^2 \longrightarrow \min, \quad (62)$$

čia f_{τ} ir $\hat{f}_{\tau}$ yra projektuotų į kryptį τ duomenų pasiskirstymo tankis ir jo įvertis;

- **geometrinė duomenų klasterizavimo procedūra.** Siūloma procedūra daugiamas duomenims klasterizuoti, kur klasės $K_1, \dots, K_q$ randamos

$$\hat{K}_1, \dots, \hat{K}_q : \sum_{j=1}^q \frac{1}{\|\hat{K}_j\|} \sum_{s,r \in K_j} \rho(X(s), X(r)) \longrightarrow \min, \quad (63)$$

čia pseudo-atstumas ρ randamas naudojantis projektuotų duomenų klasifikavimo tikimybėmis.

Metodų tyrimas. Pasiūlytos statistinės procedūros buvo tiriamos Monte-Karlo būdu. Buvo skaičiuojamos paklaidos, atspindinčios pasiskirstymo tankio ir skirstinio parametrų įvertinimo ir duomenų klasterizavimo tikslumą. Siūlomos statistinės procedūros buvo lyginamos su žinomais ir populiariais metodais.

Išvados. Iš rezultatų, gautų atlikus metodų tyrimą, buvo padarytos išvados:

1. Pasiūlyta neparametrinė vienamačio pasiskirstymo tankio statistinio įvertinimo procedūra tiksliau vertina Gauso skirstinių mišinio tankį nei žinomuose statistiniuose paketuose naudojamos neparametrinės vertinimo procedūros, jei tankio lokalaus glodumo charakteristikos labai priklauso nuo argumento.
2. Siūloma duomenų projektavimu paremta Gauso skirstinių mišinių identifikavimo metodika yra konstruktyvi ir apskaičiuoja MTM įvertį tiksliau negu kiti konstruktyvūs algoritmai.
3. Mažų imčių atveju, naudojant duomenų projektavimu pagrįstą procedūrą sudarytą iš geometrinio klasterizavimo, EM algoritmo ir mažiausių kvadratų metodo, kai sumuojamos projektuotų duomenų pasiskirstymo tankių įverčių paklaidos L_2 metrikoje, galima gauti tikslesnius Gauso skirstinių mišinio parametrų įverčius negu maksimalaus tikėtumo metodo įverčiai.
4. Sprendžiant daugiamas Gauso skirstinių mišinio identifikavimo ir duomenų klasterizavimo uždavinius tikslinga naudoti duomenų projektavimą.