



Tomas RUZGAS

**THE NONPARAMETRIC ESTIMATION OF
MULTIVARIATE DISTRIBUTION DENSITY
APPLYING CLUSTERING PROCEDURES**

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

1350

Vilnius **2007**
LEIDYKLA
TECHNIKA

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Tomas RUZGAS

**THE NONPARAMETRIC ESTIMATION OF
MULTIVARIATE DISTRIBUTION DENSITY
APPLYING CLUSTERING PROCEDURES**

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

Vilnius  2007
LEIDYKLA
TECHNIKA

Doctoral dissertation was prepared at Institute of Mathematics and Informatics in 2002–2006.

Scientific Supervisor

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

The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:

Chairman

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

Members:

Prof Dr Habil Mindaugas BLOZNELIS (Vilnius University, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Vyantas PAULAUSKAS (Vilnius University, Physical Sciences, Mathematics – 01P)

Assoc Prof Dr Jonas RIMAS (Kaunas University of Technology, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P)

Opponents:

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

Assoc Prof Dr Marijus RADAVIČIUS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P)

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics in the Senate Hall of Vilnius Gediminas Technical University at 2 p. m. on 2 March 2007.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail doktor@adm.vtu.lt

The summary of the doctoral dissertation was distributed on 2 February 2007

A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, Vilnius, Lithuania) and the Library of Institute of Mathematics and Informatics (Akademijos g. 4, Vilnius, Lithuania).

© Tomas Ruzgas, 2007

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Tomas RUZGAS

**DAUGIAMAČIO PASISKIRSTYMO
TANKIO NEPARAMETRINIS ĮVERTINIMAS
NAUDOJANT STEBĖJIMŲ
KLASTERIZAVIMĄ**

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01 P)

Vilnius  2007

Disertacija rengta 2002–2006 metais Matematikos ir informatikos institute.

Mokslinis vadovas

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

Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:

Pirmininkas

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

Nariai:

prof. habil. dr. Mindaugas BLOZNELIS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),

prof. habil. dr. Vygantas PAULAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika – 01),

doc. dr. Jonas RIMAS (Kauno technologijos universitetas, fiziniai mokslai, matematika – 01P),

prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Oponentai:

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

doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2007 m. kovo 2 d. 14 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.

Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.

Tel.: +370 5 274 4952, +370 5 274 4956; faksas +370 5 270 0112;

el. paštas doktor@adm.vtu.lt

Disertacijos santrauka išsiuntinėta 2007 m. vasario 2 d.

Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto (Saulėtekio al. 14, Vilnius, Lietuva) bei Matematikos ir informatikos instituto (Akademijos g. 4, Vilnius, Lietuva) bibliotekose.

VGTU leidyklos „Technika“ 1350 mokslo literatūros knyga

1. General Characteristic of the Dissertation

Topicality of the problem. Widespread are models of density estimation when the distribution family is known. The Gaussian distribution model is especially often met. However, the distribution family is often unknown when real distribution problems are being solved, and the precondition about Gaussian model frequently is wrong. The density estimation is divided into parametric and nonparametric depending on the fact whether the distribution family of the random variable is known or unknown. In parametric estimation one assumes that the density function $f(x)$ underlying the d -dimensional random vector X belongs to some rather restricted family of functions $f(\cdot; \theta)$ indexed by a small number of parameters. In nonparametric density estimation no parametric assumptions about $f(x)$ are made and one assumes instead that the function $f(x)$ has some continuous value (e.g. that the function has continuous derivative of the second order).

Application of many modern nonparametric estimation procedures in practice is used for the density statistical estimation of the multivariate values. The most important element of the kernel estimators of density is the smoothing bandwidth. Spline knots selection for spline estimators is also a difficult task. Although there are lots of adaptive procedures for selection of mentioned parameters, the efficiency of estimating for a small size sample is low as well as lists priorities that need to be followed while creating this method.

One of the possible ways to increase the nonparametric estimation accuracy is to reduce the problem of a multimodal density analysis to the estimation of unimodal densities by treating the density analysed as a mixture of unimodal densities. We suggest performing sample clustering at the first stage of analysis and estimating each component of distribution mixture separately at the second stage. Also we search for a more effective clustering procedure. Using the Monte-Carlo method we analyse the relationship between the estimation error, clustering type, and the number of clusters.

Aim and tasks of the work – to develop and analyse constructive algorithms for effective estimation of distribution in case of multimodality; to perform a comparative analysis of accuracy for popular statistical non-parametric estimators in case of multimodality; to explore data pre-clustering influence over the accuracy of multimodal density estimation; to compare efficiency of different data clustering procedures applied for density estimation.

Scientific novelty

- The comparative analysis of estimation accuracy is made using different methods for estimation of multivariate densities in case of multimodality. The density function estimators are selected as representatives of popular different technique estimators, which are analysed by other researchers experimentally. Present research distinguishes from the previous because of its attention to projection-based methods.
- The efficiency of multimodal density nonparametric approximation by Gaussian mixture models for moderate size samples is analysed.
- The peculiar methodic of multimodal density analysis based on treating the analysed density as a mixture of unimodal densities and on the sample projection is proposed. Efficiency of sample pre-clustering for density estimation is analysed.
- The various sample pre-clustering procedures are compared. The advantages of soft clustering to hard clustering for nonparametric estimation of multimodal density are shown.

Methodology of research. The proposed methods were analysed by means of simulation.

Practical value. Estimation of probability density functions receives more important attention through new fields of application: treatment of genetic information, analysis of astronomy objects, machine learning and many other fields of engineering.

Defended propositions

- It is demonstrated that the quality of multimodal distribution density estimation increases significantly if the sample is clustered, and later on, the density estimation methods are applied separately to each cluster.
- It is shown that the sample clustering by using the Gaussian distribution mixture model and the EM algorithm is evidently more superior if compared with the popular geometrical clustering methods. The soft clustering is more advantageous than the hard clustering when clustering results are applied to statistical estimation of the density mixtures. In addition, while the number of clusters increases moderately the estimation error reduces.
- The comparative analysis of the accuracy of a few popular nonparametric estimators is made. It shows that the most effective algorithm of multimodal density nonparametric estimation is obtained

by combining the sample pre-clustering (using the Gaussian distribution mixture model and the EM algorithm) with J. H. Friedman procedure. The rule is proposed for the selection of the number of clusters and its efficiency is analysed.

The scope of the scientific work. The scientific work consists of the general characteristic of the dissertation, three chapters, conclusions, list of literature, list of publications and addenda. The total scope of the dissertation – 93 pages, 8 pictures, 4 tables and 2 addenda.

2. The Analysed Density Estimation Algorithms

A large number of non-parametric methods designed for statistical estimation of the density function of random vectors are used in the modern data analysis. The kernel density estimators are the most common ones (J. N. Hwang, S. R. Lay and A. Lippman 1994, W. Härdle and M. Müller 2000). Spline (C. Kooperberg 1998, T. Takada 2001) and semiparametric (A. R. Gallant and D. W. Nychka 1987, F. Hoti, L. Holmström 2004) algorithms are also popular. Application of many popular non-parametric estimation procedures in practice encounters a problem of optimal parameter selection. The most important element of the kernel density estimators is the smoothing bandwidth. Spline knots selection for spline estimators is also a difficult task. Though there exist a lot of adaptive procedures for selection of mentioned parameters (C. Gu and C. Qiu 1993, W. Härdle and M. Müller 2000, D. W. Scott 1992, C. J. Stone, M. Hansen, C. Kooperberg and Y. K. Truong 1997), the efficiency is low in the case of a small sample size. It is advisable to apply the data projection technique (J. H. Friedman, W. Stuetzle, A. Schroeder 1984, J. H. Friedman 1987, M. Kavaliauskas, R. Rudzkiš 2002) in this case, because the parameter selection problem becomes more difficult when the dimension of the observed random vectors increases.

Let X be a d -dimensional random vector with an unknown distribution density function $f(x)$ and $\mathbf{X} = (X(1), \dots, X(n))$ be the independent sample of X . The comparative analysis of estimation accuracy was made using different methods. The density function estimators were selected as representatives of popular different technique estimators, which were studied experimentally by other researchers. The following statistical estimators of the distribution density are considered:

1. Silverman's adaptive kernel density estimator. A separate bandwidth is used for each observation.

2. The semi-parametric kernel density estimator analysed by Hoti and Holmström, who decomposed a random vector into two subvectors. The density of one of these vectors is estimated by the kernel density estimator, while the density of the other is approximated by the normal density function.
3. The log-spline density estimator proposed by Kooperberg and Stone. The logarithm of the analysed density is approximated by the sum of cubic B-splines.
4. The method based on projection pursuit and sequential gaussianisation of projections proposed by Friedman.
5. The inversion formula-based density estimator, which is proposed by Rudzkis.

Adaptive kernel density estimator (AKDE). Let $Z(1), \dots, Z(n)$ be standartized observations with density function $f(z)$. The kernel density estimator with the variable bandwidth is defined by the following expression

$$\hat{f}(z) = \frac{1}{n} \sum_{t=1}^n \frac{1}{h_t^d} K\left(\frac{z - Z(t)}{h_t}\right), \text{ where } \frac{z}{h} = \left(\frac{z_1}{h}, \dots, \frac{z_d}{h}\right)'. \quad (1)$$

We use the same procedure as in (J. N. Hwang, S. R. Lay, A. Lippman 1994): the standard Gaussian function is used as the kernel function K and the bandwidth h_t is determined by the equality $h_t = h(\tilde{f}(Z(t))/q)^{-v}$, where $h = (4/(2d+1)n)^{\frac{1}{d+4}}$, $\tilde{f}(\cdot)$ is the kernel density estimator (1) obtained by substituting h for h_t , $q = \exp\left(\frac{1}{n} \log \sum_{t=1}^n \tilde{f}(Z(t))\right)$ and v is the sensitivity parameter. The authors in (J. N. Hwang, S. R. Lay, A. Lippman 1994) propose to select values of this parameter from the set $\{0.2, 0.4, 0.6, 0.8\}$ by the cross-validation method (see B. W. Silverman 1986).

Semiparametric kernel density estimator (SKDE). The observed d -dimensional random vector X is decomposed into two s and $(d-s)$ -dimensional subvectors $X = (Y, Z)'$. Thus, the sample is decomposed $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})'$. The density function $f_X(\cdot)$ is presented as the product of the density function of random vector Y and a conditional density function of random vector Z , i.e.,

$$f_X(x) = f_Y(y)f_{Z|Y}(z|y), \text{ where } x = (y, z)' \in \mathbf{R}^d. \quad (2)$$

The density function $f_Y(y)$ is estimated using the kernel method, analogous to (1), with the constant kernel bandwidth matrix H , that is chose as diagonal. Subvector Y is selected by the cross-validation method as suggested in (F. Hoti, L. Holmström 2004). The conditional density $f_{Z|Y}(\cdot|y)$

is approximated by the Gaussian distribution $N(m(y), C(y))$. The conditional mean $m(y)$ and the conditional covariance matrix $C(y)$ of the random vector Y are defined by the equalities

$$\hat{m}(y) = \sum_{t=1}^n W_{H_{2,t}}(y, \mathbf{Y}) Z(t) \quad (3)$$

and

$$\hat{C}(y) = \sum_{t=1}^n W_{H_{2,t}}(y, \mathbf{Y}) (Z(t) - \hat{m}(y))(Z(t) - \hat{m}(y))', \quad (4)$$

where the weights

$$W_{H_{2,t}}(y, \mathbf{Y}) = \frac{K_{H_2}(y - Y(t))}{\sum_{j=1}^n K_{H_2}(y - Y(j))}, \quad (5)$$

by recommendation in (F. Hoti, L. Holmström 2004) suggested to chose $H_2 = 2H$.

Log spline density estimator (LSDE). The log spline density estimator approximates the logarithm of the multivariate density function by the sum of splines (see C. J. Stone, M. Hansen, C. Kooperberg and Y. K. Truong 1997)

$$\hat{f}(x) = \exp \left(\sum_{j=1}^s \beta_j B_j(x) - C(\boldsymbol{\beta}) \right), \quad (6)$$

for the given set of basis functions $B_1, \dots, B_s$ with the coefficient vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_s)$ and the normalizing factor $C(\boldsymbol{\beta})$. The procedure proposed by Kooperberg and Stone applies the cubic B-splines to estimate univariate densities. The spline knots are selected using the Akaike information criterion, and the spline coefficients are calculated using the maximum likelihood method. The estimate of the multivariate density function is the product of univariate spline density estimates.

Projection pursuit density estimator (PPDE). Standartized sample is analysed as in the case of AKDE. J. H. Friedman proposed the iterative algorithm based on the sequential search for univariate projections of data having the distribution the most different from Gaussian (P. J. Huber 1985) and subsequent transformation of these projections to the Gaussian random values. The value Z is transformed after each step, $Z^{(k)} = Q_k(Z)$, $k = 1, 2, \dots$. Let us define $Z^{(0)} = Z$. $Z^{(k)}$ is obtained from $Z^{(k-1)}$ by the following procedure. Let $g_k(u)$, $u \in \mathbf{R}^1$ denote the density function of univariate projection $\tau'Z^{(k-1)}$, where the direction vector $\tau = \tau(k)$ is selected so that g_k differs most from the standard normal density ϕ . Let us denote the corresponding distribution functions by G_k and Φ . We define

$$Z^{(k)} = Z^{(k-1)} + [\Phi^{-1}(G_k(\tau'Z^{(k-1)})) - \tau'Z^{(k-1)}]\tau.$$

Thus, the random vector $Z^{(k-1)}$ is transformed in such a way that the projection of $Z^{(k)}$ onto the direction τ would have the distribution function Φ , and the projection to the direction orthogonal to τ would remain unchanged Friedman has proved (J. H. Friedman, W. Stuetzle, A. Schroeder 1984) that the random vector $Z^{(k)}$ converges in distribution to the standard Gaussian random vector as $k \rightarrow \infty$. Thus, for large enough M ,

$$f(z) \cong \varphi(z^{(M)}) \prod_{k=1}^M \frac{g_k(\tau'(k)z^{(k-1)})}{\varphi(\tau'(k)z^{(k)})}, \quad (7)$$

where $z^{(k)} = Q_k(z)$. From expression (7), we obtain Friedman's statistics by substituting statistical estimators for the unknown univariate distribution densities g_k in the right-hand side of expression.

The projective estimator, on the basis of the Legendre orthogonal polynomial, was used for estimating densities g_k of the univariate projections. This estimator is identical to that used by Friedman. Let $\xi_1, \dots, \xi_n$ be univariate random values with a density function $g(u)$. Applying the transformation $\eta_t = 2\Phi(\xi_t) - 1$, $y = 2\Phi(u) - 1$, we obtain random values $\eta_1, \dots, \eta_n$ with the density $g^*(y) = g(u)/(2\varphi(u))$, which is concentrated in the interval $[-1, 1]$. Using the expansion on the basis of Legendre polynomials $\{\psi_j\}_{j=0}^\infty$

$$g^*(y) = \sum_{j=0}^{\infty} b_j \psi_j(y)$$

and replacing the coefficients $b_j = (j + 1/2)E\psi_j(\eta_t)$ by their empirical counterparts, we obtain the final estimator

$$\hat{g}(u) = \varphi(u) \sum_{j=0}^s \frac{2j+1}{n} \sum_{t=1}^n \psi_j(\eta_t) \psi_j(y). \quad (8)$$

According to the recommendations (J. N. Hwang, S. R. Lay and A. Lippman 1994), the order of expansion (8) was assumed to be $s \leq 6$. The projection directions $\tau(k)$ proposed by Friedman that maximize the value of the estimated projection index

$$\tau(k) = \arg \max_{\tau} \{\hat{I}(\tau)\}, \quad (9)$$

are selected. Here $I(\tau)$ is defined by the equalities

$$I(\tau) = \int_{-1}^1 g^{*2}(r) dr - 1/2 = \sum_{j=1}^{\infty} \frac{2j+1}{2} E^2[\psi_j(r)], \quad r = 2\Phi(\tau'Z) - 1, \text{ and its}$$

estimator is obtained from (8) with $\eta_t = 2\Phi(\tau'Z(t)) - 1$, $t = 1, \dots, n$.

Inversion formula based density estimator (IFDE). The density function f can be obtained from the characteristic function using the inversion formula

$$f(x) = \frac{1}{(2\pi)^d} \int_{\mathbf{R}^d} e^{-it'x} \psi(t) dt, \quad (10)$$

where $\psi(t) = \mathbf{E}e^{it'X}$ 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

$$\psi(t) = \mathbf{E}e^{it'X} = \mathbf{E}e^{i|t|\tau'X} = \psi_\tau(t), \text{ where } \tau = t/|t|, \quad (11)$$

we obtain an equality

$$f(x) = \frac{1}{(2\pi)^d} \int_{\tau:|\tau|=1} ds \int_0^\infty e^{-iu\tau'x} \psi_\tau(u) u^{d-1} du, \quad (12)$$

using the inversion formula (10) and performing the transformation to a spherical coordinate system. The first integral in expression (12) 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 $\hat{f}$

$$\hat{f}(x) = \frac{c(d)}{\#T} \sum_{\tau \in T} \int_0^\infty e^{-iu\tau'x} \hat{\psi}_\tau(u) u^{d-1} e^{-\lambda u^2} du, \quad (13)$$

where $\#T$ denotes a size of set T . The set T is a finite set of points, uniformly distributed on the unit sphere, $c(d) = d2^{-d} \pi^{-\frac{d}{2}} \Gamma(\frac{d}{2} + 1)$, where Γ denotes the Gamma function. The extra multiplier $e^{-\lambda u^2}$ is used for additional smoothing and $\hat{\psi}_\tau(\cdot)$ is the Fourier transform of the univariate projection $\tau'X$ density function estimate $\hat{f}_\tau$.

Let the estimate $\hat{f}_\tau$ is obtained by adaptive kernel density estimator procedure with the Gaussian kernel function. This enables us to calculate the integral on the right-hand side of the expression (13) analytically. For each $\tau \in T$,

$$\hat{f}_\tau(v) = \frac{1}{n} \sum_{t=1}^n \varphi\left(\frac{v - \tau'X(t)}{h_t}\right) / h_t, \quad h_t = h_t(\tau) \quad (14)$$

and

$$\hat{\psi}_\tau(u) = \frac{1}{n} \sum_{t=1}^n \exp(iu\tau'X(t) - h_t^2 u^2 / 2). \quad (15)$$

The smoothing parameter λ was selected using the cross-validation method, for $\#T = 10000$.

3. Algorithms of Sample Clustering and Determining the Number of Clusters

The term *cluster analysis* (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. An agglomerative hierarchical clustering procedure based by Johnson's (1967) is one of the most common (P. H. A. Sneath and R. R. Sokal 1973). *k*-means (J. B. MacQueen 1967) and *k*-nearest-neighbor (I. Gitman 1973, D. H. Huizinga 1978) procedures are also popular. We describe more exhaustively the sample clustering method based on density approximation by a mixture of Gaussian distribution densities.

Clustering using the EM algorithm. If the density function of the random vector X has q maxima, a mixture of q unimodal densities can approximate it

$$f(x) = \sum_{k=1}^q p_k f_k(x). \quad (16)$$

Let the distribution of the random vector X depend on a random variable v which take on the values $1, \dots, q$. It is interpreted as the number of class the observed object belongs to. In the classification theory quantities $p_k = \mathbf{P}\{v = k\}$ are called *a priori* probabilities of the observed object that belongs to k^{th} class, and quantities $\pi_k(x) = \mathbf{P}\{v = k | X = x\}$ are *a posteriori* probabilities. Function f_k is treated as the conditional density of X when $v = k$. By the term soft clustering of a sample we refer to the estimation of the values $\pi_k(X(t))$ for the all $k = 1, \dots, q$, $t = 1, \dots, n$. A sample is hard clustered if estimators $\hat{v}(1), \dots, \hat{v}(n)$ of $v(1), \dots, v(n)$ are indicated where $v(t)$ denotes the class number of the observation $X(t)$.

The mixture of Gaussian distributions is the most popular model in the clusterisation theory and practice. Therefore in that follows we assume that $f_k(x)$ are Gaussian densities with means M_k and covariance matrixes R_k . Let $f(x, \theta)$ denote the right-hand side of the equation (16) where $\theta = (p_k, M_k, R_k)$, $k = 1, \dots, q$. Since

$$\pi_k(x) = \frac{p_k f_k(x)}{f(x, \theta)}, \quad k = \overline{1, q} \quad (17)$$

the estimators of *a posteriori* probabilities are obtained as usual by "plug-in" method which replaces unknown parameter vector θ on the right-hand side of (17) with its maximum likelihood estimate $\hat{\theta}^* = \arg \max_{\theta} L(\theta)$,

$$L(\theta) = \prod_{t=1}^n f(X(t), \theta). \quad (18)$$

The EM algorithm, an iterative procedure most frequently used to find this estimation, is applied in this study.

Let $\hat{\pi}_k = \hat{\pi}_k^{(r)}$ be the estimates obtained after r cycles of the iterative procedure. Then a new estimate $\hat{\theta} = \hat{\theta}^{(r+1)}$ is defined by the equalities

$$\begin{aligned} \hat{p}_k &= \frac{1}{n} \sum_{t=1}^n \hat{\pi}_k(X(t)), \quad \hat{M}(k) = \frac{1}{n \hat{p}_k} \sum_{t=1}^n \hat{\pi}_k(X(t)) \cdot X(t), \\ \hat{R}(k) &= \frac{1}{n \hat{p}_k} \sum_{t=1}^n \hat{\pi}_k(X(t)) \cdot [X(t) - \hat{M}(k)] \cdot [X(t) - \hat{M}(k)]' \end{aligned}$$

for all $k = 1, \dots, q$. Inserting $\hat{\theta}^{(r+1)}$ into the right-hand side of expression (17) we get $\hat{\pi}^{(r+1)}(X(t))$, $k = \overline{1, q}$, $t = \overline{1, n}$. Using this iterative procedure the non-decreasing sequence $L(\hat{\theta}^{(r)})$ is obtained, however its convergence to the global maximum depends on the initial value $\hat{\theta}^{(0)}$ (or $\hat{\pi}^{(0)}$). The simplest solution of the initial value selection problem is a random start technique: the EM procedure is repeated many times from the random starting value $\hat{\pi}^{(0)}$. The result with the maximal value of $L(\hat{\theta})$ is selected as final. The methodology of consecutive extraction of the mixture components (R. Rudzkiš, M. Radavičius 1995) can be also applied.

For choosing number of clusters q various tests of model adequacy can be used. As a rule, parametric criteria, most often the likelihood ratio criterion or other tests, based on the likelihood ratio are used for testing the model adequacy or the correctness of the chosen number of components. Let us describe the proposed criteria for testing the model adequacy.

Maximum likelihood function increases criteria. Beginning with $q = 1$, the consecutive values of the parameter q further are increased till hypothesis about model adequacy will not be rejected (e.g. with the significance level $\alpha = 0.05, 0.1$). To test this hypothesis, criteria based on the increment of the maximum likelihood function can be also applied. Let

$\hat{\theta}(q)$ be an estimator of θ obtained by the EM algorithm with number of clusters equal to q . Let

$$\psi = L(\hat{\theta}(q+1)) - L(\hat{\theta}(q)), \quad (19)$$

and let $G(u)$ denote estimated distribution of ψ obtained by making use of parametric bootstrap under the condition that mixture model of Gaussian distributions is valid. Then hypothesis of model adequacy is not rejected if

$$1 - G(\psi) \geq \alpha. \quad (20)$$

It meets with difficulties in practice when computed value of likelihood function is zero through a large number of observations and little values of density function into it. Therefore we can use function instead of $L(\theta)$ which is defined in (18):

$$\tilde{L}(\theta) = \prod_{t=1}^n (\aleph + f(X(t), \theta)), \quad (21)$$

where $\aleph = \max \{u : \sum_{t=1}^n \mathbf{1}_{\{f(X(t), \theta) \leq u\}} \leq 0.05n\}$. Then equality (19) can be written

as

$$\tilde{\psi} = \tilde{L}(\hat{\theta}(q+1)) - \tilde{L}(\hat{\theta}(q)). \quad (22)$$

The condition $1 - G(\tilde{\psi}) \geq \alpha$ is verified using $\tilde{\psi}$ criterion.

The number of clusters obtained in this way is denoted by q^* .

4. Conditional Density Estimators when Sample Is Soft Clustered

We discuss the application of few methods in estimating the conditional density in case the sample has been clustered. As noted above, we assume that the observed random vector X depends on a latent random variable v that takes on the values $1, \dots, q$, and that the conditional density f_i of the random vector X given as $\{v = i\}$ is unimodal for $i = \overline{1, q}$. Since the formula (16) holds, we use equality

$$\hat{f}(x) = \sum_{i=1}^q \hat{p}_i \hat{f}_i(x), \quad (23)$$

to estimate the density $f(x)$ after the sample clustering. Replacing *a posteriori* classification probabilities $\pi_i(x) = \mathbf{P}\{v = i | X = x\}$ by their estimates $\hat{\pi}_i(x)$, and using expression $p_i = \mathbf{E}\pi_i(X)$ we have

$$\hat{p}_i = \frac{1}{n} \sum_{t=1}^n \hat{\pi}_i(X(t)), \quad i = 1, \dots, q.$$

When estimating the component $f_i(x)$ of mixture (23) using the adaptive kernel density estimation algorithm, observations $Z(t)$ like $X(t)$ are investigated with weights $\hat{\pi}_i(t) = \hat{\pi}_i(Z(t)) = \hat{\pi}_i(X(t))$, $t = 1, \dots, n$. So, the formula (1) is replaced by

$$\hat{f}_i(z) = \frac{1}{\hat{p}_i n} \sum_{t=1}^n \frac{\hat{\pi}_i(t)}{h_t^d} K\left(\frac{z - Z(t)}{h_t}\right). \quad (24)$$

Similarly, employing the semiparametric algorithm in the estimation of $f_i(x)$, weights (5), which are used for estimation conditional mean (3) and covariance matrix (4), is replaced by

$$W_{i,H_{2,t}}(y, \mathbf{Y}) = \frac{\hat{\pi}_i(t) K_{H_2}(y - Y(t))}{\sum_{j=1}^n \hat{\pi}_i(j) K_{H_2}(y - Y(j))}, \quad (25)$$

Analogous, when estimating the component $f_i(x)$ using the projection pursuit algorithm, the weight $\hat{\pi}_i(t) = \hat{\pi}_i(Z(t))$, $t = 1, \dots, n$ is assigned to the observation $\xi_t = \tau'Z(t)$, $t = 1, \dots, n$. Therefore, when applying equality (7) the estimate the density $f_i(x)$, formula (8) is replaced by

$$\hat{g}_i(u) = \varphi(u) \sum_{j=0}^s \frac{2j+1}{\hat{p}_i n} \sum_{t=1}^n \hat{\pi}_i(t) \psi_j(\eta_t) \psi_j(y). \quad (26)$$

In this way, the estimator of the projection index is replaced by

$$\hat{I}_i(\tau) = \sum_{j=1}^s \frac{2j+1}{2\hat{p}_i^2 n^2} \sum_{t=1}^n \left(\hat{\pi}_i(t) \psi_j(r_t) \right)^2. \quad (27)$$

Similarly, using the inversion formula based algorithm, in the estimation of $f_i(x)$, equation (14) for each projection $\tau \in T$

$$\hat{f}_\tau(v) = \frac{1}{\hat{p}_i n} \sum_{t=1}^n \frac{\hat{\pi}_i(t)}{h_t} \varphi\left(\frac{v - \tau'X(t)}{h_t}\right), \quad h_t = h_t(\tau) \quad (28)$$

by corresponding characteristic function $\hat{\psi}_\tau$.

The same formulas are valid in the case of hard clustering with $\hat{\pi}_i(t) = \mathbf{1}_{\{\hat{v}(t)=i\}}$, $\hat{p}_i = \sum_t \hat{\pi}_i(t)$.

5. 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. Since a preliminary comparative analysis made in this research has shown that probability clustering methods were evidently more advantageous than geometrical

ones (k -means, etc.), only the sample clustering methods based on density approximation by a mixture of Gaussian distribution densities were considered in the sequel.

Mixtures of the multivariate ($d = \overline{2, 5}$) Gaussian and Cauchy distributions with independent components were used. The multidimensional Cauchy density is defined as a product of the univariate Cauchy densities. Thus,

$$f(x) = \sum_{i=1}^q p_i f_C(x, m_i, u_i), \text{ where } f_C(x, m_i, u_i) = \prod_{j=1}^d \frac{u_{ij}}{\pi[u_{ij}^2 + (x_j - m_{ij})^2]}.$$

For comparison, samples of the components that correspond to Gaussian mixtures with a unit covariance matrix are generated. In the latter case,

$$f(x) = \sum_{i=1}^q p_i f_N(x, m_i, \sigma_i), \text{ where } f_N(x, m_i, \sigma_i) = \frac{1}{\prod_{j=1}^d \sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{1}{2} \sum_{j=1}^d \frac{(x_j - m_{ij})^2}{\sigma_{ij}^2}\right\}.$$

The comparative analysis of the mentioned density estimation methods has been made exploring the data that was used by J. N. Hwang, S. R. Lay and A. Lippman in their paper 1994:

The very first data generated are of unimodal distribution with the following parameters: for the Cauchy distribution,

$$p=1, m=(0.0, 0.0, 0.0, 0.0, 0.0)', u=(0.84, 1.02, 0.70, 1.20, 0.96)';$$

for the Gaussian distribution,

$$p=1, m=(0.0, 0.0, 0.0, 0.0, 0.0)', \sigma^2=(0.84, 1.02, 0.70, 1.20, 0.96)'.$$

In case $d = \overline{2, 4}$, the parameters are defined by the first d elements of the given 5-dimensional parameter.

Data of the second type are of slightly overlapping bimodal distribution with the following parameters: for the Cauchy distribution,

$$p_1 = 0.65, m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)', u_1 = (0.42, 0.51, 0.35, 0.60, 0.48)',$$

$$p_2 = 0.35, m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)', u_2 = (0.33, 0.46, 0.53, 0.43, 0.45)';$$

for the Gaussian distribution,

$$p_1 = 0.65, m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)', \sigma_1^2 = (0.42, 0.51, 0.35, 0.60, 0.48)',$$

$$p_2 = 0.35, m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)', \sigma_2^2 = (0.33, 0.46, 0.53, 0.43, 0.45)'.$$

Data of the third type are of highly overlapping bimodal distribution with the following parameters: for the Cauchy distribution,

$$p_1 = 0.65, m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)', u_1 = (0.84, 1.02, 0.70, 1.20, 0.96)',$$

$$p_2 = 0.35, m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)', u_2 = (0.66, 0.92, 1.06, 0.86, 0.90)';$$

for the Gaussian distribution,

$$p_1 = 0.65, m_1 = (0.0, 0.0, 0.0, 0.0, 0.0)', \sigma_1^2 = (0.84, 1.02, 0.70, 1.20, 0.96)',$$

$$p_2 = 0.35, m_2 = (2.0, 2.0, 2.0, 2.0, 2.0)', \sigma_2^2 = (0.66, 0.92, 1.06, 0.86, 0.90)'.$$

Various values of a parameter are selected to explore the properties of the estimators. The samples with the standard scale parameters from the following mixtures have been considered

1. $q = 2, p_1 = (1-p_2), p_2 = 0.1, 0.3, 0.5, m_1 = (0.5k, 0.5k, 0.5k, 0.5k, 0.5k)',$
 $m_2 = (0, 0, 0, 0, 0)'$ (here and below the index k takes on the values 1, ..., 6).
2. $q = 3, p_1 = p_2 = (1-p_3)/2, p_3 = 0.1, 1/3, 0.8, m_1 = (0.5k, 0.5k, 0.5k, 0.5k, 0.5k)',$
 $m_2 = (0, 0, 0, 0, 0)', m_3 = (0.5k, 0.5k, 0, 0, 0)'.$
3. $q = 4, p_1 = p_2 = p_3 = (1-p_4)/3, p_4 = 0.1, 0.25, 0.7,$
 $m_1 = (0.5k, 0.5k, 0.5k, 0.5k, 0.5k)', m_2 = (0, 0, 0, 0, 0)',$
 $m_3 = (0.5k, 0.5k, 0, 0, 0)', m_4 = (0, 0, 0.5k, 0.5k, 0.5k)'.$
4. $q = 2, p_1 = (1-p_2), p_2 = 0.1, 0.3, 0.5, m_1 = (0.5k, 0.5k)', m_2 = (0, 0)'.$
5. $q = 3, p_1 = p_2 = (1-p_3)/2, p_3 = 0.1, 1/3, 0.8, m_1 = (0.5k, 0.5k)', m_2 = (0, 0)',$
 $m_3 = (0.5k, 0)'.$
6. $q = 4, p_1 = p_2 = p_3 = (1-p_4)/3, p_4 = 0.1, 0.25, 0.7, m_1 = (0.5k, 0.5k)',$
 $m_2 = (0, 0)', m_3 = (0.5k, 0)', m_4 = (0, 0.5k)'.$

For each model, samples of different sizes 50, 100, 200, 400, 800, 1600 and 3200 are generated.

The accuracy of densities estimation was measured by

$$\delta_1 = \frac{1}{n} \sum_{t=1}^n |f(X(t)) - \hat{f}(X(t))| \cong \int |f(x) - \hat{f}(x)| f(x) dx, \quad (29)$$

$$\delta_2 = \frac{1}{n} \sum_{t=1}^n \left| \frac{f(X(t)) - \hat{f}(X(t))}{f(X(t)) + \hat{f}(X(t))} \right| \cong \frac{1}{2} \int |f(x) - \hat{f}(x)| dx, \quad (30)$$

and measure proposed in (J. N. Hwang, S. R. Lay, A. Lippman 1994):

$$\delta_3 = Err / Var. \quad (31)$$

Here $Err = \frac{1}{n} \sum_{t=1}^n (\hat{f}(X(t)) - f(X(t)))^2$ stands for the mean square error

and $Var = \frac{1}{n} \sum_{t=1}^n (f(X(t)) - \bar{f}(X(t)))^2$, where $\bar{f}(X(t))$ signifies the average of $f(X(1)), \dots, f(X(n))$.

The simulation analysis showed (part of the averaged errors δ_3 (in bold) and their standard deviations are presented in Tables 1–3, for other data models the accuracy analysis results are similar to the presented ones) that the results for the data were used by J. N. Hwang, S. R. Lay and A. Lippman, obtained by means of adaptive kernel and projection pursuit density estimation, are similar to those obtained by J. N. Hwang, S. R. Lay and A. Lippman, i.e., in the case of small sample sizes and heavy tails (Cauchy samples), it is better to use the kernel density estimators, in the case of large data dimensions and large sample sizes (400 and more observations), or in the case of the Gaussian distribution, better

Table 1. Single mode 4-dimensional distributions

Method	Cauchy distribution				Gaussian distribution			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.
AKDE	0.1994	0.1787	0.1283	0.1052	0.2055	0.1178	0.1455	0.0845
	0.0056	0.0324	0.0014	0.0051	0.0289	0.0121	0.0169	0.0169
PPDE	0.1804	0.1115	0.0445	0.0323	0.126	0.0668	0.0457	0.0243
	0.0034	0.0109	0.0073	0.0031	0.0088	0.0141	0.0061	0.0061
IKDE	0.2099	0.19	0.0777	0.0719	0.1764	0.1661	0.126	0.1159
	0.0087	0.0278	0.0094	0.0118	0.0063	0.0178	0.0037	0.0059
LSDE		0.1729		0.0729		0.1208		0.1099
		0.0107		0.0045		0.0066		0.0126
SKDE		0.1908		0.0647		0.0993		0.0541
		0.0032		0.0071		0.0124		0.0015

Table 2. Bimodal slightly overlapping 4-dimensional mixtures

Method	Cauchy distribution				Gaussian distribution			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.
AKDE	0.2173	0.1755	0.1706	0.1257	0.2963	0.2531	0.2495	0.1882
	0.1644	0.0466	0.0861	0.0214	0.055	0.0166	0.0706	0.0178
PPDE	0.2106	0.2027	0.1834	0.1057	0.2219	0.0928	0.059	0.0328
	0.0804	0.0598	0.0266	0.0224	0.0242	0.0137	0.0229	0.0148
IKDE	0.227	0.2124	0.1851	0.1732	0.253	0.2531	0.1841	0.1766
	0.0685	0.0037	0.0847	0.0214	0.0621	0.0017	0.0316	0.0017
LSDE		0.213		0.1378		0.1281		0.0824
		0.0283		0.0077		0.0148		0.0136
SKDE		0.2011		0.1418		0.1393		0.0759
		0.0313		0.0201		0.0107		0.0147

results are obtained using the projection pursuit density estimator. In the case of the 5-dimensional Gaussian distribution, quite good results are obtained using the method, based on the inversion formula. The preliminary data clustering into homogeneous groups, using automatic EM algorithm, enabled us to reduce errors by 2–3 times in the case of a small sample, and up to 5 times in some other cases. For large samples ($n = 1600, 3200$), the error reduction ratio equals 1.05–2. A conclusion can be drawn that unambiguously, *projection pursuit is the best estimator*. For unimodal Gaussian and Cauchy distributions, estimation errors decrease (especially in the case of small samples) up to 4.6 times, provided the preliminary data clustering is applied. Semiparametric density estimations are good enough

Table 3. Bimodal highly overlapping 4-dimensional mixtures

Method	Cauchy distribution				Gaussian distribution			
	$n = 400$		$n = 1600$		$n = 400$		$n = 1600$	
	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.	without cluster.	with cluster.
AKDE	0.2478	0.1946	0.1416	0.1341	0.2526	0.1049	0.2039	0.0629
	0.0889	0.0123	0.0434	0.0109	0.0471	0.0058	0.0729	0.0094
PPDE	0.1879	0.1628	0.1403	0.0912	0.1684	0.0512	0.0591	0.0412
	0.0078	0.0429	0.0165	0.0021	0.0278	0.0106	0.005	0.0063
IKDE	0.2496	0.2239	0.1455	0.1427	0.2563	0.2321	0.1808	0.1644
	0.0258	0.0518	0.0373	0.0213	0.0122	0.0018	0.005	0.0052
LSDE		0.2184		0.1352		0.1772		0.1213
		0.0299		0.0106		0.0061		0.0055
SKDE		0.2193		0.1245		0.0801		0.0809
		0.0047		0.0056		0.0078		0.0097

in the case of unimodal Gaussian densities. For 5-dimensional mixture densities, inversion formula based density estimator turned out to be a good estimator either. It has been found out that, in some cases, very accurate results could be obtained by the logspline method, however, in the cases with outliers, logspline density estimator yielded great errors that increased the overall averaged error of this method. The inversion formula based algorithm is very slow in comparison with other methods.

The results of 1-6 models obtained corroborate the previous conclusion about the expediency of the clustering procedure. The dependency of errors δ_1 and δ_2 on the size of sample, the type of clustering, the number of clusters, and the distance between the modes of densities is of rather a similar nature (in term of quality). The clustering procedure substantially reduces the error of estimation. The errors are 5–10 % smaller in the case of soft clustering as compared to the hard clustering. q^* stands for the optimal number of clusters (see the proposed definition at the Section 3). A larger number of clusters yields a very slight improvement, while computational time increases drastically.

The two figures below illustrate the functional dependency of the estimation errors on the sample size and the number of clusters for adaptive kernel and projection pursuit density estimators. The results for Cauchy mixture in case of model 2 are presented in Fig's 1–2 (the symbols A and P specify AKDE and PPDE algorithms respectively; the density estimators without clustering is plotted by solid lines, the density estimators with hard clustering by the short-dashed lines, and density estimators with soft clustering by the long-dashed lines). The same behaviour is established in the case of other models.

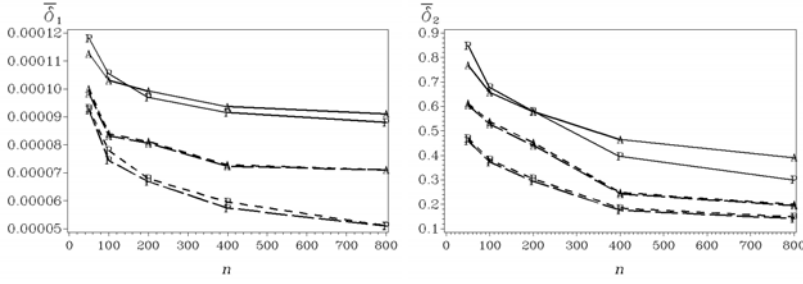


Fig 1. Relationship between the sample size and errors

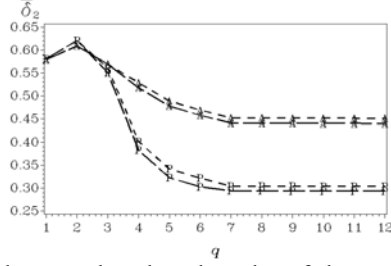


Fig 2. Relationship between the selected number of clusters and errors ($n = 400$)

6. Conclusions

Scientific work is applied for statistical estimation of multimodal distribution density of random vectors. There were formulated the following conclusions:

1. The quality of multimodal distribution density statistical estimation increases significantly if the sample is pre-clustered (i.e., the multimodal density function is approximated by a mixture of unimodal densities), and later on, the density estimation methods are applied to each cluster respectively.
2. The highest efficiency of densities estimations in the analyzed cases is reached when primary sample clustering is done, and later on, the iterative procedure proposed by J. H. Friedman is applied for estimating each density component.
3. Since the clustering results are applied to statistical estimation of multimodal densities, it is shown, that the soft clustering based on approximation of the sample distribution by the mixture of Gaussian densities and EM algorithm is more advantageous than the hard clustering or other popular geometrical ones.
4. The number of clusters, selected using bootstrap method, is near optimal.

Published works on the topic of the dissertation

Publications in journals of the databases of the Institute of Scientific Information

1. RUDZKIS, R.; RUZGAS, T. Clustering Effect on the Statistical Estimation Accuracy of Distribution Density. *Acta Applicandae Mathematicae*, Dordrecht Kluwer Academic Publishers, 2007. ISSN 0167-8019. (Accepted)

Publications in journals of the international databases from confirmed list of the Lithuania Science Council

2. KAVALLIAUSKAS, M.; RUDZKIS, R.; RUZGAS, T. The Projection-based Multivariate Distribution Density Estimation. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, Vol 8, Tartu, Tartu University Press, 2004, p. 135–141. ISSN 1406-2283.
3. RUZGAS, T.; RUDZKIS, R.; KAVALLIAUSKAS, M. Application of Clustering in the Non-Parametric Estimation of Distribution Density. *Nonlinear Analysis: Modelling and Control*, Vol 11, No 4, Vilnius, 2006, p. 393–411. ISSN 1392-5113.

Publications in the other reviewed scientific journals

4. RUZGAS, T. Comparison of Various Clustering Algorithms Efficiency. *Lietuvos matematikos rinkinys*, spec. release, Vol 42, Vilnius, MII, 2002, p. 571–576 (in Lithuanian). ISSN 0132-2818.
5. RUZGAS, T.; KAVALLIAUSKAS, M. Nonparametric Density Estimation Using a Multidimensional Mixture Model of Gaussian Distributions. *Lietuvos matematikos rinkinys*, spec. release, Vol 45, Vilnius, MII, 2005, p. 369–374 (in Lithuanian). ISSN 0132-2818.
6. ŠMIDTAITĖ, R.; RUZGAS, T. Research of Nonparametric Density Estimation Algorithms by Applying Clustering Methods. *Lietuvos matematikos rinkinys*, spec. release, Vol 46, Vilnius, MII, 2006, p. 273–279 (in Lithuanian). ISSN 0132-2818.

About the author

T. Ruzgas was born in Kaunas in 1977. In 1983, he started his studies at the secondary school in Kaunas. He was the laureate of Kaunas Young Mathematician Olympiad in 1994. In 1995, he graduated from Atžalynas secondary school in Kaunas. He entered Kaunas University of Technology and graduated from the Faculty of Fundamental Sciences with Bachelor of Sciences degree in Applied Mathematics in 1999. He continued studies and

graduated the same university with Master of Sciences degree in Applied Mathematics in 2001. From 2002 he continues studies at the Institute of Mathematics and Informatics to acquire a PhD degree in Mathematics.

Between 1999 and 2001, he was a laboratory assistant at Applied Mathematics Department of Kaunas University of Technology. Between 2001 and 2002, he was a junior researcher at Applied Statistics Department of Institute of Mathematics and Informatics. From 2001 he is an assistant at Applied Mathematics Department of Kaunas University of Technology.

DAUGIAMAČIO PASISKIRSTYMO TANKIO NEPARAMETRINIS ĮVERTINIMAS NAUDOJANT STEBĖJIMŲ KLASTERIZAVIMĄ

Mokslo problemos aktualumas. Nagrinėjama problema yra glaudžiai susijusi su daugiamačių stebinių pasiskirstymo analize – viena iš esminių duomenų analizės šakų, kuria remiasi daugelis kitų sprendžiamų (diskriminantinė analizė, vaizdų atpažinimas) uždavinių. Pasiskirstymo tankių vertinimo metodologija susilaukia vis didesnio dėmesio dėl naujai atsiradusių taikymo sričių: genetinės informacijos apdorojimo, astronomijos tyrimų objektų analizės, kompiuterinės technikos bei jos periferijos duomenų tyrimo ir t.t.

Gana plačiai paplitę nagrinėjami modeliai, kuomet skirstinio šeima yra žinoma. Ypač dažnai sutinkamas Gauso skirstinio modelis. Sprendžiant realius pasiskirstymo uždavinius dažnai nežinoma skirstinio šeima (pvz., prielaida apie modelio gausiškumą neretai būna klaidinga). Pagal tai, ar žinoma stebimo atsitiktinio dydžio skirstinio šeima, tankių vertinimas skirstomas į parametrinį ir nelinearinį vertinimą. Parametriniame vertinime daroma prielaida, kad tankio funkcija $f(x)$, apibūdinanti d -matį atsitiktinį vektorių X , priklauso tam tikrai gan siaurai funkcijų šeimai $f(\cdot; \theta)$, priklausančiai nuo nedidelio kiekio parametrų. Nelineariniam tankio vertinimui jokios parametrinės prielaidos apie $f(x)$ nėra reikalingos, tačiau vietoj to daromos kitos prielaidos, pavyzdžiui, apie funkcijos $f(x)$ tolydumą.

Šiuolaikinėje duomenų analizėje naudojama daugybė nelinearinių metodų, skirtų daugiamačių atsitiktinių dydžių pasiskirstymo tankio statistiniam vertinimui. Ypač plačiai paplitę branduoliniai įvertiniai, gana populiarūs ir splaininiai bei pusiau parametriniai algoritmai. Taikant daugumą populiarių nelinearinio vertinimo procedūrų, praktikoje susiduriama su jų parametrų optimalaus parinkimo problema. Branduolinių įvertinių konstrukcijos svarbiausiu elementu yra glodinimo plotis, splaininiams įvertinims nelengva parinkti mazgus ir t.t. Nors yra sukurta nemažai adaptivių minėtų parametrų parinkimo procedūrų, tačiau jos nėra

pakankamai efektyvios kai imties tūris nėra didelis, ypač jei stebinių dimensija yra didoka. Pastaruoju atveju tikslinga taikyti duomenų projektavimą, nes parametų parinkimo uždavinys tuo sunkesnis, kuo didesnė stebimų atsitiktinių vektorių dimensija.

Vienas iš būdų mėginti padidinti neparametrinių įverčių tikslumą yra daugiamodalinio tankio analizės suvedimas į vienamodalinių tankių vertinimą, traktuojant tiriamą tankį kaip vienamodalinių tankių mišinį. Pagrindinė darbo dalis skirta šios idėjos įgyvendinimui. Siūloma pirmame tyrimų etape imtį klasterizuoti ir po to kiekvieną klasterį atitinkančius skirstinių mišinio komponentus įvertinti atskirai.

Darbo tikslas ir uždaviniai – sukurti ir ištirti daugiamachio pasiskirstymo tankio neparametrinio vertinimo algoritmus, kurie būtų efektyvūs daugiamodališkumo atveju; atlikti populiarių neparametrinių tankių statistinių įverčių tikslumo lyginamąją analizę daugiamodaliniu atveju; ištirti pirminio duomenų klasterizavimo poveikį daugiamodalinio tankio statistinio vertinimo tikslumui; palyginti įvairių klasterizavimo procedūrų taikymo efektyvumą pasiskirstymo tankių vertinime.

Mokslinis naujumas

- Atlikta skirtingų tipų statistikų, skirtų daugiamachų tankių vertinimui, lyginamoji analizė daugiamodaliniu atveju. Tyrimui atrenkant konkrečias pasiskirstymo tankių vertinimo procedūras, buvo siekiama, kad jos atstovautų populiarių įverčių klasėms ir jau būtų eksperimentiškai tirtos kitų tyrėjų. Pateikiamas darbas išsiskiria ypatingu dėmesiu projektavimo metodams, kurių efektyvumas daugiamodaliniu atveju menkai tirtas.
- Ištirtas neparametrinės daugiamodalinio tankio aproksimacijos Gauso mišinių modeliais tikslingumas esant nuosaikiam imties dydžiui.
- Pasiūlyta originali daugiamodalinio tankio analizės metodika, paremta tiriamo tankio traktavimu vienamodalinių tankių mišiniu ir imties projektavimu. Ištirtas pirminio imties klasterizavimo efektyvumas tankio vertinime.
- Palygintos įvairios pirminio imties klasterizavimo procedūros ir parodytas negriežto klasterizavimo pranašumas prieš griežtą vertinant neparametrinį daugiamodalinį tankį.

Tyrimų metodika apima pasiūlytų statistinių procedūrų tirimą Monte-Karlo būdu. Buvo skaičiuotos paklaidos, atspindinčios pasiskirstymo tankio ir skirstinio parametų vertinimo bei duomenų klasterizavimo tikslumą. Siūlomos statistinės procedūros buvo lyginamos su žinomais ir populiariais metodais.

Praktinė vertė. Gauti tyrimų rezultatai praplečia pasiskirstymo tankių vertinimo metodologiją ir gali būti naudojami sprendžiant diskriminantinės analizės, vaizdų atpažinimo ir panašius uždavinius įvairiose taikomosiose srityse, pavyzdžiui apdorojant genetinę informaciją, analizuojant astronomijos tyrimų objektus ir kt.

Ginamieji teiginiai

- Atliekant populiarių neparametrinių tankio įverčių tikslumo palyginamąją analizę daugiamodaliskumo atveju, parodyta, kad įvertinimo rezultatai ženkliai pagerėja, jei stebiniai pirmiausiai klasterizuojami.
- Parodyta, kad imties skaidymas į klasterius taikant Gauso skirstinių mišinio modelį yra akivaizdžiai pranašesnis už populiarius geometrinius klasterizavimo metodus, o negriežtas klasterizavimas naudingesnis nei griežtas.
- Pasiūlyta klasterių skaičiaus nustatymo taisyklė ir ištirtas jos efektyvumas.
- Parodyta, kad daugiamodalinio neparametrinio tankio įvertinimo algoritmas gautas apjungiant pirminę imties klasterizaciją su J. H. Friedman procedūra, naudojančia duomenų projektavimą, yra efektyvesnis nei tirti kiti populiarūs vertinimo būdai.

Darbo apimtis. Darbą sudaro bendra darbo charakteristika, 3 skyriai, išvados, literatūros sąrašas, publikacijų sąrašas ir priedai. Bendra disertacijos apimtis – 93 puslapiai, 8 iliustracijos, 4 lentelės ir 2 priedai.

Darbo išvados

Iš rezultatų, gautų atlikus metodų tyrimą, suformuluotos šios išvados:

1. Daugiamodalinių pasiskirstymo tankių vertinimo rezultatai labai pagerėja, jei stebiniai pirmiausia klasterizuojami (traktuojant jų daugiamodalinį tankį kaip vienamodalinių tankių mišinį), o tankių vertinimo metodai yra taikomi kiekvienam klasteriui atskirai.

2. Daugeliu atvejų didžiausias vertinimo efektyvumas buvo pasiekiamas, kai po pirminio imties suskaidymo kiekvieną klasterį atitinkantys tankio komponentai buvo įvertinti J. H. Friedman pasiūlyta rekurentine procedūra.

3. Parodyta, kad negriežtas imties klasterizavimas, kuris remiasi nagrinėjamo tankio aproksimacija Gauso pasiskirstymo tankių mišiniu ir EM algoritmu, yra pranašesnis nei griežtas ar kitos populiarios geometrinio klasterizavimo procedūros, kai klasterizavimo rezultatai taikomi daugiamodaliniams tankiams statistiškai vertinti.

4. *Bootstrap* metodu nustatomas klasterių skaičius yra artimas optimaliam.

Tomas Ruzgas

**THE NONPARAMETRIC ESTIMATION OF
MULTIVARIATE DISTRIBUTION DENSITY
APPLYING CLUSTERING PROCEDURES**

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

Tomas Ruzgas

**DAUGIAMAČIO PASISKIRSTYMO
TANKIO NEPARAMETRINIS ĮVERTINIMAS
NAUDOJANT STEBĖJIMŲ KLASTERIZAVIMĄ**

**Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01 P)**

2007 01 22. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, LT 10223 Vilnius
Spausdino Matematikos ir informatikos instituto
Informacijos ir leidybos skyrius, Akademijos g. 4, LT 08412 Vilnius