2122. Assessment and comparison of likely density distributions in the cases of thickness measurement of skin tumours by ultrasound examination and histological analysis

Indre Drulyte[1], **Tomas Ruzgas**[2], **Renaldas Raisutis**[3], **Skaidra Valiukeviciene**[4]
[1, 3]Prof. K. Baršauskas Ultrasound Research Institute, Kaunas University of Technology, Kaunas, Lithuania
[2]Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences,
Kaunas University of Technology, Kaunas, Lithuania
[3]Department of Electrical Power systems, Faculty of Electrical and Electronics Engineering,
Kaunas University of Technology, Kaunas, Lithuania
[4]Department of Skin and Venereal Diseases, Lithuanian University of Health Sciences, Kaunas, Lithuania
[1]Corresponding author
E-mail: [1]*drulyte.indre@inbox.lt*, [2]*tomas.ruzgas@ktu.lt*, [3]*renaldas.raisutis@ktu.lt*,
[4]*skaidra.valiukeviciene@lsmuni.lt*

Received 9 May 2016; received in revised form 14 June 2016; accepted 21 June 2016
DOI http://dx.doi.org/10.21595/jve.2016.17183

Abstract. Ultrasonic diagnostic methods are used to estimate the structural changes and to measure parameters of lesions of the human tissue. Nowadays, the special algorithms of medical data analysis are able to perform diagnosis and monitor the progress of treatment, efficiency of treatment methods, also to estimate the health status and to make prognosis of the diseases evolution. The aim of the presented research is to check the goodness of fit test for thicknesses of the skin tumours measured in two different ways (ultrasound examination and histological analysis) and to compare the compatibility of likely density of histological thicknesses distribution of the skin tumours and density of Normal distribution. As a result, the study has showed that thicknesses of the skin tumours measured by ultrasonic method are strongly similar to histological values, which means that the density of ultrasonic thicknesses distribution and density of Normal distribution are closely interconnected. Therefore, the obtained results show the sufficient level of reliability in the case of application of non-invasive ultrasonic thickness measurement comparing with reference invasive technique based on biopsy and histological thickness evaluation.

Keywords: skin tumour, thickness measurement, goodness of fit test, kernel method, nonparametric density estimator, Monte Carlo method.

1. Introduction

The storage of medical information and its statistical analysis are being carried out since the middle of ages. The first known statistical journal of medicine was published in London, in 1662 [1]. In 1863, F. Nightingale, the pioneer of nowadays nursing, raised the problem about the lack of medical statistics records and non-systematic storage in hospitals, as a consequence of treatment effectiveness and costs limited analysis. In 1977 the US Congress published a study “Medical information systems practitioner’s consequences” [2]. It states, that medical information systems can be a useful tool for training, also to help medicine and health care specialists leading to higher quality of facilities and optimization of health care institution activity. The authors of study have confirmed that medical information system will be a useful tool for researches and health governing institutions. Since 2000, the active global implementation of regional and national electronic health records systems started. The aim of these systems is to save all important patients medical records. Lithuanian health sector also applies information technologies, creating a national electronic health services and information system for cooperation infrastructure, also subsystem for national medical images archiving and exchange. Health care institutions implement and improve information systems of hospitals, systems for radiological images preview and archiving, information systems of laboratories [3]. Information system of the health care keeps a structured information about the patient, such as diagnosis, demographic patient data, vital functions, test results and etc. These data analysis and mining are very important for all patients. The smart analysis of patient records helps to solve tasks as a faster diagnostic, choosing of optimal treatment, prediction of treatment period and results, to identify the risk of complications, resources optimization of the health care institutions. Last decade, data mining research in biomedicine is highly considered [4, 5]. Data mining methods and algorithms can be useful if researches clearly understand scopes, types of the data and peculiarities. The most common tasks mentioned in literature are classification, clustering, prediction, association, visualization, identification of deviations and analysis of internal links. For these data mining tasks, we need to choose a suitable algorithm. Choosing of method or an optimal algorithm depends on aims of task analysis and data characteristics. Over the last decade, there are found enough methods of data mining application in medicine. In diagnosis there are widely applied neural networks, decision trees, decision rules [6], methods for search of associative rules (for costs analysis) [7], prediction of patient health and treatment probability, also very popular to use combinations of prediction algorithms [5]. In 2014, N. Esfandiari et al. [4], carried out a literature review, there are described applications of data mining in medicine based on analysis of the structured data. There are stated that classification (neural networks, decision trees, decision rules, support vector model), clustering (k-means, hierarchical clustering) and associative search (a priori associative rules search) models are the most popular in medicine. Lalayants et al. [8] have said that the solution of successful medical data mining is to identify the right activity of health care institution or to find the clinical problem. Data mining methods are usually used in biomedical data analysis and visualization tasks in order to facilitate decision-making [9]. If the data mining process would be enough simple, the management of information problems would be already solved long time ago (R. Bellazzi, B. Zupan [5]). Practical data mining application in medicine has some obvious barriers as technological problems, trans-disciplinary communication, ethics and patient data security [7, 9, 10]. Medical research leads to a lot of data characterizing the condition of patient. All these data are dynamically changing and depend on patient illness, patient biological condition, environment, the quality of life, related diseases and other actually reasons, those can be described as a random factor. The change of medical statistics observations is described by primary statistics analysis. The results lead to further instructions of medical research and affect hardly choosing and application of the appropriate statistical method. The reliability of above mentioned methods usually depends on the assumption of data distribution – normal, binomial and etc. Therefore, at first, it is necessary to check the appropriate assumption. This paper will present a simple, effective method of nonparametric statistics and some hypothesis criteria about the variable distribution and identity checking of two distributions. These hypotheses are called goodness of fit test hypothesis. The purpose of research is to determine the connection between thicknesses of the skin tumours measured by non-invasive ultrasonic technique and after a surgical intervention measured histologically by optical microscope. Also, to compare the compatibility of likely density of histological thicknesses distribution of the skin tumours and Normal distribution density this method is effective for structured big data matrix and simple to use. There is no problem to check the sample which is distributed by well-known theoretical distribution, because these cases are already examined both theoretically as well as empirically. The biggest challenge is to check the identity between two samples. As a solution the most commonly used techniques relies on differences of density distribution. Even in nowadays data analysis there are a lot of evaluation methods of density distribution, but in practice it is not easy to find the effective evaluation procedure if the data distribution is multimodal and the volume of sample is small. Kernel smoothing is the most frequently used nonparametric estimation method (see Jones, et al., 1996 [11], Marron and Wand, 1992 [12]; Silverman, 1986 [13]). Thus far, there is no generally accepted method for kernel estimation, which outperforms the other in all cases. Although many adaptive selection procedures have been proposed (Bashtannyk and Hyndman, 1998 [14]; Jones, 1992 [15]; Zhang et al., 2004 [16]), their efficiency has not been well established yet, especially for samples of a moderate size. On the basis of Lithuanian cancer register data, Lithuania established more than 250 melanomas cases every year. Even Lithuania is not included in the list of biggest melanomas risk country, but the statistics shows that the number of melanomas cases in Lithuania is increasing every year. The main reason is too late diagnosis. Usually melanoma is diagnosed in 2-4 stages. The mortality of melanoma in Lithuania is bigger than in other Europe countries [17, 18]. Melanoma is a rapidly growing and spreading malignant tumor, rarely amenable to treat through the spread of time. In the absence of effective treatment of metastatic melanoma, a key factor of survival of melanoma is early diagnosis and urgent surgical removal of the primary tumor. The earlier diagnosis of melanoma can be prevented by regularly checking of nevus and disposal nevus, those can be malignant. Surgical removal of melanoma having thickness of 1 mm increase the probability of survival, for 10 the years survival rate is 90-97 percent [19, 20].

The paper consists of 6 sections. Section 2 reviews the kernel density estimator and kernel functions; Section 3 proves optimal selection of smoothing parameter; Section 4 describes the simulation experiment and contains the simulation results; Section 5 shows the analysis in an empirical context using the retrospective observations of thicknesses of the skin tumour for goodness of fit tests; the concluding remarks are presented in Section 6.

2. Kernel density estimator

A d-dimensional random vector $X \in R^d$ satisfies a mixture model if its distribution density function $f(x)$ is given by the equality:

$$f(x) = \sum_{k=1}^{q} p_k f_k(x) = f(x, \theta). \tag{1}$$

The parameter q is the number of components in the mixture. The component weights p_k are called a priori probabilities and satisfy the conditions:

$$p_k > 0, \quad \sum_{k=1}^{q} p_k = 1. \tag{2}$$

Function $f_k(x)$ is the distribution density function of the kth component and θ is the vector of parameters of mixture model Eq. (1). Suppose a simple sample $\mathbf{X} = \left(X(1), \ldots, X(n)\right)$ of size n from $\mathbf{X}$ is given. The estimation of the distribution density of an observed random vector is one of the main statistical tasks.

A histogram is one of the simplest and the oldest density estimators. This graphical representation was first introduced by Karl Pearson in 1891 (Scott, 1992 [21]). For the approximation of density $f(x)$, the number of observations $X(t)$ falling within the range of Ω is calculated and divided by n and the volume of area Ω. The histogram produced is a step function and the derivative either equals zero or is not defined (when at the cut off point for two bins). This is a big problem if we are trying to maximize a likelihood function that is defined in terms of the densities of the distributions.

It is remarkable that the histogram stood as the only nonparametric density estimator until the 1950's, when substantial and simultaneous progress was made in density estimation and in spectral density estimation. In 1951, in a little-known paper, Fix and Hodges [22] introduced the basic algorithm of nonparametric density estimation; an unpublished technical report was published formally as a review by Silverman and Jones in 1989 [23]. They addressed the problem of statistical discrimination when the parametric form of the sampling density was not known. During the following decade, several general algorithms and alternative theoretical modes of analysis were introduced by Rosenblatt in 1956 [24], Parzen in 1962 [25], and Cencov in 1962

[26]. Then followed the second wave of important and primarily theoretical papers by Watson and Leadbetter in 1963 [27], Loftsgaarden and Quesenberry in 1965 [28], Schwartz in 1967 [29], Epanechnikov in 1969 [30], Tarter and Kronmal in 1970 [31] and Kimeldorf and Wahba in 1971 [32]. The natural multivariate generalization was introduced by Cacoullos in 1966 [33]. Finally, in the 1970's the first papers focusing on the practical application of these methods were published by Scott et al. in 1978 [34] and Silverman in 1978 [35]. These and later multivariate applications awaited the computing revolution.

The basic kernel estimator $\hat{f}(x)$ with a kernel function K and a fixed (global) bandwidth h for multivariate data $X \in \mathbf{R}^d$ may be written compactly as:

$$f(x) = \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{x - X(t)}{h}\right). \tag{3}$$

The kernel function $K(u)$ should satisfy the condition:

$$\int_{-\infty}^{+\infty} K(u)\,du = 1. \tag{4}$$

Usually, but not always, $K(u)$ will be a symmetric probability density function $K(u) = K(-u)$ for all values of u (see Silverman, 1986 [13]).

At first, the data is usually prescaled in order to avoid large differences in data spread. A natural approach (see Fukunaga, 1972 [36]) is first to standardize the data by a linear transformation yielding data with zero mean and unit variance. As a result, Eq. (3) is applied to the standardized data. Let Z denote the sphered values of random X:

$$Z = S^{\frac{-1}{2}} * (X - \bar{X}), \tag{5}$$

where $\bar{X}$ is the empirical mean, and $S \in R^{d\times d}$ is the empirical covariance matrix. Applying the kernel density estimator to the standardized data $Z = \big(Z(1), \dots, Z(n)\big)$ yields the following estimator of density function $f(x)$:

$$f_Z(z) = \frac{1}{nh^d}\sum_{t=1}^{n} K\left(\frac{z - Z(t)}{h}\right), \tag{6}$$

$$f(x) = \frac{(\det S)^{\frac{-1}{2}}}{nh^d}\sum_{t=1}^{n} K\left(S^{\frac{-1}{2}}\frac{x - X(t)}{h}\right). \tag{7}$$

The comparative analysis of estimation accuracy was made for four different types of kernels. The first three kernels are classical, whereas the last one is new.

The Gaussian kernel is consistent with the distribution of normal $\varphi(x)$ (see Gasser et al., 1985 [37], Marron and Nolan, 1988 [38]) selection:

$$K_G(x) = \varphi(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} e\left(\frac{-x^T x}{2}\right). \tag{8}$$

The Epanechnikov kernel is the second order polynomial, corrected to satisfy the properties of the density function (see Epanechnikov, 1969 [30], Sacks and Ylvisaker, 1981 [39]):

$$K_E(x) = \frac{d+2}{2V_d}(1 - x^T x)\mathbf{1}_{\{|x^T x \le 1|\}}, \tag{9}$$

where $V_d = \pi^{d/2}/\Gamma(d/2+1)$ is the volume of the d-dimensional unit sphere, and $\Gamma(u) = \int_0^\infty y^{u-i}\, e^{-y} dy.$

The Triweight kernel proposed by Tapia and Thompson in 1978 [40] has better smoothness properties and finite support. It was investigated in detail by Hall in 1985 [41]:

$$K_T(x) = \frac{(d+4)(d+6)}{24}\frac{(d+2)}{2V_d}(1 - x^T x)^3\mathbf{1}_{\{|x^T x \le 1|\}}. \tag{10}$$

The new kernel K_{New} has lighter tails than Gaussian distribution density and was introduced by the authors of this article:

$$K_{New}(x) = \varphi\left(|u|^{\frac{1}{\alpha}}\right)\frac{1}{\alpha^d}\left(\left|\prod_{i=1}^{d} x_i\right|^{\frac{1}{d}}\right)^{1-\alpha}. \tag{11}$$

This kernel function depends on parameter α. In simulations, the chosen values of the parameter were 0.25, 0.5, and 0.75. The first two values produce worse accuracy results in comparison with the value of 0.75. Therefore, only the results obtained for $\alpha = 0.75$ are reported here.

3. Optimal bandwidth selection

There are three parameters in kernel density estimator: the sample size n, the kernel function $K(\cdot)$ and the bandwidth h. Quite typically we cannot do anything about the sample size and we have to make the best out of the situation by choosing an appropriate kernel and a suitable bandwidth. It is well known that the bandwidth selection is the most crucial step in order to obtain a good estimate (see Wand and Jones, 1995 [42]). Unfortunately, bandwidth selection is the most difficult problem in kernel density estimation and a definite and unique solution to this problem does not exist.

It is rather surprising that the most effective bandwidth selection method is a visual assessment by the researcher. The researcher visually compares different density estimates, based upon a variety of bandwidths and then chooses the bandwidth that corresponds to the subjectively optimal estimate. The unfortunate part is that such bandwidths are non-unique; this method will yield different bandwidths when performed by different researchers. This method can also be very time consuming.

The approach based on mathematical analysis is to quantify the discrepancy between the estimate and the target density by evaluated error criterion. The optimal bandwidth will then be the bandwidth value that minimizes the error measured by the error criterion. Such a method is objective and can be time-efficient as computers can solve it numerically.

A global measure of precision is the asymptotic mean integrated squared error (AMISE):

$$AMISE\left(\hat{f}(x)\right) = \frac{K_\nu^2(K)}{(\nu!)^2}R(\nabla^\nu f)h^{2\nu} + \frac{R(K)^d}{nh^d}, \tag{12}$$

where $\nabla^\nu f(x) = \sum_{k=1}^{d} \partial^\nu/\partial x_k^\nu f(x)$ and $R(g) = \int_{-\infty}^{\infty} g(u)^2 du$ is the roughness of a function. The order of a kernel, ν, is defined as the order of the first non-zero moment $\kappa_j(K) = \int_{-\infty}^{\infty} u^j K(u) du$. For example, if $\kappa_1(K) = 0$ and $\kappa_2(K) > 0$ then K is a second-order kernel and $v = 2$. If $\kappa_1(K) = \kappa_2(K) = \kappa_3(K) = 0$ but $\kappa_4(K) > 0$ then K is a fourth-order kernel and $v = 4$.The order of a symmetric kernel is always even. Symmetric non-negative kernels are second-order kernels. A kernel is higher-order kernel if $v > 2$. These kernels will have negative parts and are not probability densities.

The optimal bandwidth is:

$$h_0 = (\frac{(v!)^2 dR(K)^d}{2vK_v^2(K)R(\nabla^v f)})^{1/(2v+d)} n^{-1/(2v+d)}. \tag{13}$$

The optimal bandwidth depends on the unknown quantity $R(\nabla^{(v)} f)$. For a rule-of-thumb bandwidth, Silverman proposed that it is possible to try the bandwidth computed by replacing f in the optimal formula by g_0 where g_0 is a reference density – a plausible candidate for f, and $\hat{\sigma}$ is the sample standard deviation (see Bruce E. Hansen, 2009 [43]). The standard choice is a multivariate normal density. The idea is that if the true density is normal, then the computed bandwidth will be optimal. If the true density is reasonably close to the normal, then the bandwidth will be close to optimal. Calculation of that is proceeded according to:

$$R(\nabla^v \varphi) = \frac{d}{\pi^{\frac{d}{2}} 2^{d+v}} \left((2v-1)!! + (d-1)\left((v-1)!!\right)^2\right), \tag{14}$$

where the double factorial means $(2s+1)!! = (2s+1)(2s-1) \ldots 5 \cdot 3 \cdot 1$. Making this substitution, we obtain:

$$h_0 = C_v(K,d) n^{-1/(2v+d)}, \tag{15}$$

where:

$$C_v(K,d) = \left(\pi^{\frac{d}{2}} 2^{d+v-1} (v!)^2 R(K)^d / vK_v^2(K) \left((2v-1)!! + (d-1)\left((v-1)!!\right)^2\right)\right)^{1/(2v+d)},$$

and this assumed that variance is equal to 1. Rescaling the bandwidths by the standard deviation of each variable, we obtain the rule-of-thumb bandwidth for the ith variable is:

$$h_i = \hat{\sigma}_i \, C_v(K,d) n^{-\frac{1}{2v+d}}. \tag{16}$$

Table 1. Normal reference rule-of-thumb constants ($C_v(K,d)$) for the multivariate second-order kernel density estimator

Kernel	$d = 1$	$d = 2$	$d = 3$	$d = 4$	$d = 5$	$d = 6$	$d = 7$	$d = 8$	$d = 9$	$d = 10$
Gaussian	1.059	1.000	0.969	0.951	0.9340	0.933	0.929	0.927	0.925	0.925
Epanechnikov	2.345	2.191	2.120	2.073	2.044	2.025	2.012	2.004	1.998	1.995
Triweight	3.155	2.964	2.861	2.800	2.762	2.738	2.723	2.712	2.706	2.702
New	1.142	1.079	1.045	1.025	1.014	1.007	1.002	1.000	0.998	0.998

Table 1 provides the normal reference rule-of-thumb constants ($C_v(K,d)$ in (Eq. (15)) for the second-order d-variate kernel density estimator. We point out several striking features. First, in the common setting of a second order kernel ($v = 2$) the rule-of-thumb constants are decreasing as d increases. Scott (1992 [21]) notes that these reach a minimum when $d = 11$. The $v = 2$ case is the only one he considers. When $v > 2$, it is possible to show that the rule-of-thumb constants are increasing in the dimensionality of the problem. The basic idea behind this is given that higher-order kernels reduce bias; larger bandwidths are needed to minimize AMISE. However, note that the increase is not uniform over v.

4. The analysis of estimation accuracy

A comprehensive simulation study was conducted with the aim to compare the kernel functions described before. The main attention is paid to the case where the density of independent d-dimensional observations is GMM:

$$f(x)=\sum_{i=1}^{q}p_i\,\varphi_i\,(x)=f(x,\theta),\quad x\in\boldsymbol{R}^d,\tag{17}$$

where $\theta=(p_i,M_i,R_i,i=1,2,\dots,q)$. Univariate, bi-variate, and quinta-variate GMMs, from a suggested collection was used in comparative analysis as the benchmark densities:

1) Gaussian
$p_1=1, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
2) skewed unimodal
$p_1=1/5, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_2=1/5, M_2=(1/2{,}0,\dots,0), R_2=diag([(2/3)^2,\dots,(2/3)^2])$
$p_3=3/5, M_3=(13/12{,}0,\dots,0), R_3=diag([(5/9)^2,\dots,(5/9)^2])$
3) strongly skewed
$p_n=1/8, M_n=(3((2/3)^n-1),0,\dots,0), R_n=diag([(2/3)^{2n},\dots,(2/3)^{2n}]), n=0,\dots,7$
4) kurtotic unimodal
$p_1=2/3, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_2=1/3, M_2=(0,\dots,0), R_2=diag([(1/10)^2,\dots,(1/10)^2])$
5) outlier
$p_1=1/10, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_2=9/10, M_2=(0,\dots,0), R_2=diag([(1/10)^2,\dots,(1/10)^2])$
6) bimodal
$p_1=1/2, M_1=(-1{,}0,\dots,0), R_1=diag([(2/3)^2,\dots,(2/3)^2])$
$p_2=1/2, M_2=(1{,}0,\dots,0), R_2=diag([(2/3)^2,\dots,(2/3)^2])$
7) separated bimodal
$p_1=1/2, M_1=(-3/2{,}0,\dots,0), R_1=diag([(1/2)^2,\dots,(1/2)^2])$
$p_2=1/2, M_2=(3/2{,}0,\dots,0), R_2=diag([(1/2)^2,\dots,(1/2)^2])$
8) skewed bimodal
$p_1=3/4, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_2=1/4, M_2=(3/2{,}0,\dots,0), R_2=diag([(1/3)^2,\dots,(1/3)^2])$
9) trimodal
$p_1=9/20, M_1=(-6/5{,}0,\dots,0), R_1=diag([(3/5)^2,\dots,(3/5)^2])$
$p_2=9/20, M_2=(6/5{,}0,\dots,0), R_2=diag([(3/5)^2,\dots,(3/5)^2])$
$p_3=1/10, M_3=(0,\dots,0), R_3=diag([(1/4)^2,\dots,(1/4)^2])$
10) claw
$p_1=1/2, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_n=1/10, M_n=(n/2-1{,}0,\dots,0), R_n=diag([(1/10)^2,\dots,(1/10)^2]), n=0,\dots,4$
11) double claw
$p_1=49/100, M_1=(-1{,}0,\dots,0), R_1=diag([(2/3)^2,\dots,(2/3)^2])$
$p_2=49/100, M_2=(1{,}0,\dots,0), R_2=diag([(2/3)^2,\dots,(2/3)^2])$
$p_n=1/350, M_n=(n-3/2{,}0,\dots,0), R_n=diag([(1/100)^2,\dots,(1/100)^2]), n=0,\dots,6$
12) asymmetric claw
$p_1=1/2, M_1=(0,\dots,0), R_1=I=diag([1,\dots,1])$
$p_n=2^{1-n}/31, M_n=(n+1/2{,}0,\dots,0), R_n=diag([(2^{-n}/10)^2,\dots,(2^{-n}/10)^2]),$
$n=-2,\dots,2$
13) asymmetric double claw

$p_j = 46/100, M_j = (2j-1,0,\dots,0), R_j = diag([(2/3)^2,\dots,(2/3)^2]), j = 0, 1;$
$p_n = 1/100, M_n = (-n/2,0,\dots,0), R_n = diag([(1/100)^2,\dots,(1/100)^2]), n = 1, 2, 3;$
$p_k = 1/100, M_k = (k/2,0,\dots,0), R_k = diag([(1/100)^2,\dots,(1/100)^2]), k = 1, 2, 3.$
14) smooth comb: $p_n = 2^{5-n}/63, M_n = (65-96(1/2)^n/21,0,\dots,0),$
$R_n = diag([32/63/2^{2n},\dots,32/63/2^{2n}]), n = 0,\dots, 5.$
15) discrete comb
$p_n = 2/7, M_n = (12n-15/7,0,\dots,0), R_n = diag([(2/7)^2,\dots,(2/7)^2]), n = 0, 1, 2;$
$p_k = 1/21, M_k = (2k/7,0,\dots,0), R_k = diag([(1/21)^2,\dots,(1/21)^2]), k = 8, 9, 10.$

These densities have been carefully chosen because they thoroughly represent many different types of challenges to curve estimators. The first five represent different types of problems that can arise for unimodal densities. The rest of the densities are multimodal. Densities from 6 to 9 are mildly multimodal and one might hope to be able to estimate them fairly well with a data set of a moderate size.

The remaining densities are strongly multimodal and for moderate sizes it is difficult to recover even their shape. Yet, they are well worth studying because the issue of just how many of them can be recovered is an important one. The claw density, 10, is of special interest as this is where the surprising result of local minima in the mean integrated square error occurs. The double claw density, 11, is essentially the same as 6, except that approximately 2 % of the probability mass appears in the spikes. The asymmetric claw and double claw densities, 12 and 13, are modifications of 10 and 11, respectively. The smooth and discrete comb densities, 14 and 15, are enhancements of the basic idea of separated bimodal, 7. Both of these are shown because they have much different Fourier transform properties, since 14 has essentially no periodic tendencies, while 15 has two strong periodic components.

Note that univariate case of this set of models is similar to collection suggested by Marron and Wand in 1992 [12].

In the simulation study, low-size and moderate-size samples (16, 32, 64, 128, 256, 512, 1024) were used. 10000 replications were generated in each case. The conclusions presented below are based on the analysis of these medians and minimums. The estimation accuracy is measured by the mean absolute percentage error:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{f\big(X(t)\big)-\hat{f}\big(X(t)\big)}{f\big(X(t)\big)}\right| \cong \int\left|f(x)-\hat{f}(x)\right|dx. \tag{18}$$

5. Results of the study

The results of univariate kernel density estimation are examined in detail by Ruzgas and Drulyte, 2013 [1]. The experimental research showed that some of kernel density functions used with multiple distributions mixtures lead to particularly good results. For example, Triweight kernel density function is characterized as one of the most effective when the study is done by using “Discrete comb” mixture with sample size bigger than 256, and dimension equal to 2. The results obtained with Epanechnikov kernel density function have shown that this function is appropriate to be used when the calculations are carried out with the average sample size by using “Bimodal”, “Separated bimodal” and “Smooth comb” mixtures with dimension equal to 2. In addition, the new kernel density function proposed by authors of this research has also shown unexpected results. The smallest median errors for all sample sizes when dimension is equal to 2, are obtained by using the mean average percentage error (MAPE) even at five different mixtures: “Gaussian”, “Skewed unimodal”, “Strongly skewed”, “Kurtotic unimodal”, “Outlier”. Meanwhile, when the sample size is less or equal to 256, the smallest median errors are obtained with “Bimodal”, “Separated bimodal”, “Smooth comb” and “Discrete comb” mixtures. Another important point is that the new kernel density function gives us the smallest median errors with all mixtures of Gaussian distribution and all sample sizes when the dimension is equal to five. The second effective function is Gaussian kernel density function.

The errors dependences on sample size and selected dimension are shown in Fig. 1. Here the Gaussian, Epanechnikov, Triweight and new kernel density functions are marked as G, E, T and N show the results of estimation accuracy. Medians and minimums of mean average percentage errors are marked by solid and dashed lines. The results of errors dependences on sample size results got by using "Skewed bimodal" mixture and different dimensions are shown in Fig. 1.

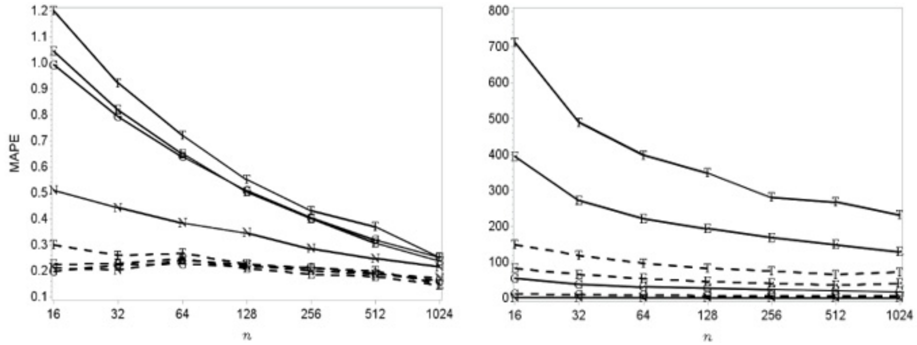


Fig. 1. Estimation accuracy based on MAPE for skewed bimodal bi-variate and quinta-variate densities (here MAPE means the mean absolute percentage error; n is the sample size; the Gaussian, Epanechnikov, Triweight and new kernel density functions are marked as G, E, T and N)

When the dimension is increasing the smallest errors are getting by using new kernel density function. Meanwhile, the Gaussian kernel density function is respectively appropriate to be used when the dimension is smaller or equal to 4 and smaller than 3 in the case of the Epanechnikov and Triweight functions. The effectiveness of the Gaussian kernel density function is shown in Fig. 2.

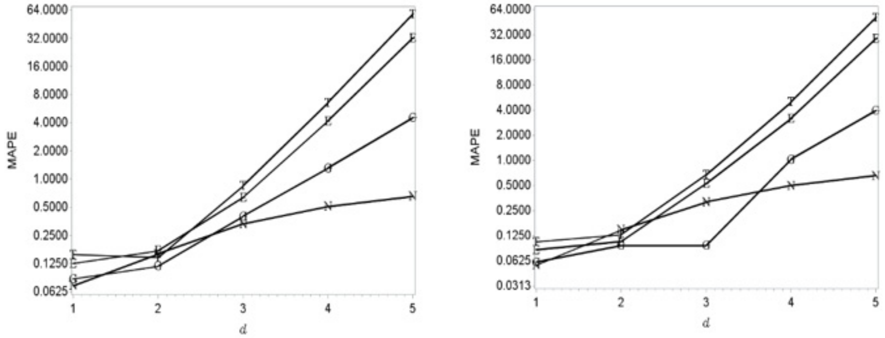


Fig. 2. The relationship between number of dimension and MAPE (Gaussian densities with sample sizes 512 and 1024). Here MAPE means the mean absolute percentage error; d is the dimension; the Gaussian, Epanechnikov, Triweight and new kernel density functions are marked as G, E, T and N

6. The application of goodness fit of test

The set of real clinical data was used as an empirical example (see Fig. 3). Within this section a set of values (the sample size was equal to 52 observations) of the skin lesions previously used for clinical decision support by non-invasive ultrasonic measurements in vivo and histological evaluation ex vivo of their thickness and malignancy after surgical excision has been obtained and compared. The analysis was performed retrospectively in an empirical context in order to estimate the goodness of fit tests.

Histological and ultrasonic data have been collected at the Department of Skin and Venereal Diseases of Lithuanian University of Health Sciences (LUHS). The study was approved by regional ethics committee; the collection of all data was approved by the institutional review board after patients' informed consent was obtained in accordance with the Declaration of Helsinki Protocols. The data used in the empirical example were acquired on 52 suspicious melanocytic skin tumours (MST) which included 46 melanocytic nevi and 6 melanomas. Inclusion criteria of the study covered size of the tumour up to 1 cm in diameter and histological thickness of ≤1.5 mm.

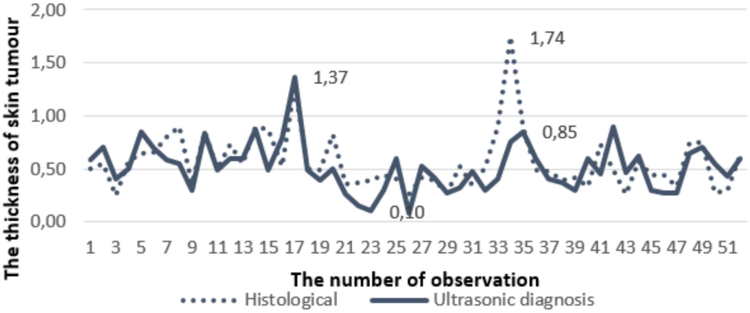


Fig. 3. The results of histological measurements and measurements made by ultrasonic diagnosis

During non-invasive ultrasonic measurements of human skin DUB-USB ultrasound system ("Taberna pro medicum") of 22 MHz was used for transmission and reception of ultrasonic waves. The immersion experimental set-up with mechanically scanned ultrasonic transducer was employed. The transducer was focused at the surface of the skin. In addition, the system was used for acquisition, digitization and transfer to personal computer the received A-scan ultrasonic signals. The set of acquired A-scan signals were used for reconstruction of the B-scan image. Finally, the maximal thickness of the skin lesion was manually evaluated by a well-experienced dermatologist measuring the distance between the lower edge of the entry echo and the deepest point of the posterior margin of the hypoechoic zone. During the evaluation of thickness, the value of ultrasound velocity was assumed to be 1580 m/s.

After a surgical excision and during the routine histopathology the vertical distance from the uppermost level of the stratum granulosum in the epidermis to the lowest point of the lesion without infiltrate (histological tumour thickness, Breslow index) was independently evaluated by two pathologists and averaged.

More details about ultrasonic examinations in dermatology and comparison with histological data are provided by Jasaitiene et al. in 2011 [44] and Kučinskienė et al. in 2014 [45].

For the goodness of fit we are using some tests based on kernel density estimators described above. Let $X_1, \ldots, X_n$ be a sample of independent observations of a random variable X with unknown probability density function $f(x)$, $x \in R$. For the given sample it is required to test the hypothesis mentioned in publication made by Rudzkis and Bakshaev in 2013 [46]:

H_0: $f(x) = f_0(x)$, against alternative H_1: $f(x) = (1 - \epsilon) f_0(x) + \epsilon g(x)$.

Here $f_0(x)$ is a given probability density function, ϵ is negligible and $\epsilon g(x)$ is an arbitrary distribution, where $\sigma_g^2 \leq \sigma_{f_0}^2$ and σ_f^2 is a variance of distribution f.

In this study five tests of goodness of fit have been tried: Pearson's chi-squared test, Rudzkis-Bakshaev's test, Kolmogorov–Smirnov test, Cramer von Mises test and Kuiper's test for four different kernel functions. One of the steps leading to the main result was to check the goodness of fit between the density of ultrasonic thicknesses distribution and density of histological thicknesses distribution of the skin tumours. The next step was to compare the compatibility of likely density of histological thicknesses distribution of the skin tumours and Normal distribution density. If two mentioned checked conditions are satisfied, as a result it is clear that the density of ultrasonic thicknesses distribution and Normal distribution density are interconnected. All results of the goodness of fit between the density of ultrasonic thicknesses distribution and density of histological thicknesses distribution of the skin tumours (denoted as U H) and the goodness of fit between the density of histological thicknesses distribution and Normal distribution density (denoted as H N) are shown in Table 2.

Table 2. The results of goodness fit of test based on kernel functions

Goodness of Fit Test	Kernel function				
		Normal	Epanechnikov	Triweight	New proposed
Pearson's chi-squared χ^2	U H	~1	~1	~1	~1
	H N	0.4474	0.0063	0.1220	0.0087
Rudzkis-Bakshaev	U H	0.9930	0.9970	0.9970	0.9900
	H N	0.9730	0.9670	0.9780	0.8990
Kolmogorov-Smirnov	H N	0.8826	0.9251	0.9079	0.9124
	H N	0.9997	0.9999	0.9999	0.9998
Cramer von Mises	U H	0.6851	0.7246	0.7040	0.7186
	H N	0.8973	0.8909	0.8984	0.8859
Kuiper's	U H	0.9998	0.9999	0.9999	0.9999
	H N	~1	~1	~1	~1

7. Conclusions

Within the performed study the check of the goodness of fit test for thicknesses of the skin tumours measured in two different ways (non-invasive ultrasound examination and invasive histological analysis). The performed simulation study leads to the kernel K which has shown a better performance for Gaussian mixtures with considerably overlapping components and multiple peaks (double claw distribution). In addition, its accuracy decreases more slowly than the other kernels, when the random vector dimension increases. The empirical study has shown that Pearson's chi-squared test is the most sensitive of all used tests. The main reason is the differences between empirical and theoretical distributions due to heavy tails of the empirical distributions. As a result, the Kuiper's test has the lowest sensitivity criteria and was the most powerful in performed comparative analysis. The obtained results have shown that the density of ultrasonic thicknesses distribution is similar to the Normal distribution density more than 90 percent. Hence, the reliability of ultrasonic thickness measurement of the skin tumour is completely covered by high similarity to the histological thickness measurement, which is known as a golden standard in the field of dermatology. Also the application of goodness fit of test has shown that p-value of all criteria's with all kernel functions are approximately 2 times bigger than Pearson's chi-squared test. Therefore, it proves that the application of non-invasive ultrasonic technique (at least of 22 MHz) for thickness estimation of the melanocytic skin lesions (tumours and nevus) possesses high reliability and is suitable to be used in daily clinical practise.

References

[1] **Ruzgas T., Drulytė I.** Kernel density estimators for Gaussian mixture models. Lithuanian Journal of Statistics (Lietuvos Statistikos Darbai), Lithuanian Statistical Association, Vilnius, Vol. 52, Issue 1, 2013, p. 14-21.

[2] Policy Implications of Medical Information Systems. Report by the US Congress Office of Technology Assessment, http://digital.library.unt.edu/ark:/67531/metadc39374/, 1977.

[3] Ministry of Health of the Republic of Lithuania, http://sam.lrv.lt/en/.

[4] **Esfandiari N., Babavalian M. R., Moghadam A. M., Tabar V. K.** Knowledge discovery in medicine: current issue and future trend. Expert Systems with Applications, Elsevier, Vol. 41, Issue 9, 2014, p. 4434-4463.

[5] **Bellazzi R., Zupan B.** Predictive data mining in clinical medicine: current issues and guidelines. International Journal of Medical Informatics, Elsevier, Vol. 77, Issue 2, 2008, p. 81-97.

[6] **Houston A. L., Chen H., Hubbard S. M., Schatz B. R., Ng T. D., Sewell R. R., Tolle K. M.** Medical data mining on the internet: research on a cancer information system. Artificial Intelligence Review, Springer, Vol. 13, Issue 5, 1999, p. 437-466.

[7] **Silver M., Sakata T., Su H. C., Herman C., Dolins S. B., O'Shea M. J.** Case study: how to apply data mining techniques in a healthcare data warehouse. Journal of Healthcare Information Management, Wiley, Vol. 15, Issue 2, 2001, p. 155-164.

[8] **Lalayants M., Epstein I., Auslander G. K., Chan W. C. H., Fouché C., Giles R., Joubert L., Rosenne H., Vertigan A.** International social work. Sage Journals, Vol. 56, Issue 6, 2013, p. 775-797.

[9] **Wasan S., Bhatnagar V., Kaur H.** The impact of data mining techniques on medical diagnostics. Data Science Journal, Ubiquity Press, Vol. 5, 2006, p. 119-126.

[10] **Cios K. J., Moore G. W.** Uniqueness of medical data mining. Artificial Intelligence in Medicine, Elsevier, Vol. 26, Issues 1-2, 2002, p. 1-24.

[11] **Jones M. C., Marron J. S., Sheather S. J.** A brief survey of bandwidth selection for density estimation. Journal of the American Statistical Association, Vol. 91, 1996, p. 401-407.

[12] **Marron J. S., Wand M. P.** Exact mean integrated squared error. Annals of Statistics, Vol. 20, 1992, p. 712-736.

[13] **Silverman B. W.** Density Estimation for Statistics and Data Analysis. Chapman and Hall, London, 1986.

[14] **Bashtannyk D. M., Hyndman R. J.** Bandwidth Selection for Kernel Conditional Density Estimation. Technical Report, Department of Econometrics and Business Statistics, Monash University, 1998.

[15] **Jones M. C.** Potential for automatic bandwidth choice in variations of kernel density estimation. Statistics and Probability Letters, Vol. 13, 1992, p. 351-356.

[16] **Zhang X., King M. L., Hyndman R. J.** Bandwidth selection for multivariate kernel density estimation using MCMC. Computational Statistics and Data Analysis, Vol. 50, 2004, p. 3009-3031.

[17] **Smailyte G., Jasilionis D., Kaceniene A., Krilaviciute A., Ambrozaitiene D., Stankuniene V.** Suicides among cancer patients in Lithuania: a population-based census-linked study. Cancer Epidemiology, Elsevier, Vol. 37, Issue 5, 2013, p. 714-718.

[18] **Sant M., Allemani C., Santaquilani M., Knijn A., Marchesi F., Capocaccia R.** EUROCARE-4. Survival of cancer patients diagnosed in 1995-1999. Results and commentary. European Journal of Cancer, Elsevier, Vol. 45, 2009, p. 931-991.

[19] **Braun R. P., Saurat J. H., French L. E.** Dermoscopy of pigmented lesions: a valuable tool in the diagnosis of melanoma. Swiss Medical Weekly, Vol. 134, Issues 7-8, 2004, p. 83-90.

[20] **Gershenwald J. E., Soong S. J., Balch C. M.** 2010 TNM staging system for cutaneous melanoma and beyond. Annals of Surgical Oncology, Springer, Vol. 17, Issue 6, 2010, p. 1475-1477.

[21] **Scott D. W.** Multivariate Density Estimation: Theory, Practice and Visualization. John Wiley, New York, 1992.

[22] **Fix E., Hodges J. L.** Discriminatory Analysis – Nonparametric Discrimination: Consistency Properties. Report No. 21-49-004, US Air Force School of Aviation Medicine, Random Field, Texas, 1951.

[23] **Silverman B. W., Jones M. C.** E. Fix and J. L. Hodges (1951): an important contribution to nonparametric discriminant analysis and density estimation. International Statistical Review, Vol. 57, Issue 3, 1989, p. 233-247.

[24] **Rosenblatt M.** Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, Vol. 27, 1956, p. 832-837.

[25] **Parzen E.** On the estimation of probability density and mode. The Annals of Mathematical Statistics, Vol. 33, 1962, p. 1065-1076.

[26] **Cencov N. N.** Estimation of unknown density function from observations. SSSR Academy of Sciences, Vol. 147, 1962, p. 45-48.

[27] **Watson G. S., Leadbetter M. R.** On the estimation of the probability density II. The Annals of Mathematical Statistics, Vol. 34, 1963, p. 480-491.

[28] **Loftsgaarden D. O., Quesenberry C. P.** A nonparametric estimate of a multivariate density function. The Annals of Mathematical Statistics, Vol. 36, Issue 3, 1965, p. 1049-1051.

[29] **Schwartz S. C.** Estimation of probability density by an orthogonal series. The Annals of Mathematical Statistics, Vol. 38, Issue 4, 1967, p. 1261-1265.

[30] **Epanechnikov V. A.** Nonparametric estimates of a multivariate probability density. Theoretical Probability Applications, Vol. 14, 1969, p. 153-158.

[31] **Tarter M., Kronmal R.** On multivariate density estimates based on orthogonal expansions. The Annals of Mathematical Statistics, Vol. 41, Issue 2, 1970, p. 718-722.

[32] **Kimeldorf G., Wahba G.** Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications, Vol. 33, 1971, p. 82-95.

[33] **Cacoullos T.** Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics, Vol. 18, Issue 1, 1966, p. 179-189.

[34] **Scott D. W., Tapia R. A., Thompson J. R.** Multivariate Density Estimation by Discrete Maximum Penalized-Likelihood Methods. Graphical Representation of Multivariate Data. Academic Press, New York, 1978.

[35] **Silverman B. W.** Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. The Annals of Statistics, Vol. 6, 1978, p. 177-184.

[36] **Fukunaga K.** Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.

[37] **Gasser T., Müller H. G., Mammitzsch V.** Kernels for nonparametric curve estimation. Journal of the Royal Statistical Society, Vol. 47, 1985, p. 238-252.

[38] **Marron J. S., Nolan D.** Canonical kernels for density estimations. Statistics and Probability Letters, Vol. 7, Issue 3, 1988, p. 195-199.

[39] **Sacks J., Ylvisaker D.** Asymptotically optimum kernels for density estimation at a point. The Annals of Statistics, Vol. 9, 1981, p. 334-346.

[40] **Tapia R. A., Thompson J. R.** Nonparametric Probability Density Estimation. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore and London, 1978.

[41] **Hall P.** Kernel estimation of a distribution function. Communications in Statistics. Theory and Methods, Vol. 14, 1985, p. 605-620.

[42] **Wand M. P., Jones M. C.** Kernel Smoothing. Chapman and Hall, London, 1995.

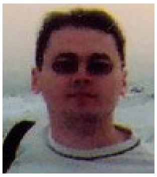
[43] **Hansen B. E.** Lecture Notes on Nonparametrics. University of Wisconsin, 2009, www.ssc.wisc.edu/~bhansen/718/NonParametrics1.pdf

[44] **Jasaitiene D., Valiukeviciene S., Linkeviciute G., Raisutis R., Jasiuniene E., Kazys R.** Principles of high-frequency ultrasonography for investigation of skin pathology. Journal of the European Academy of Dermatology and Venereology, 2011, p. 375-382.

[45] **Kučinskienė V., Samulėnienė D., Gineikienė A., Raišutis R., Kažys R., Valiukevičienė S.** Preoperative assessment of skin tumor thickness and structure using 14-MHz ultrasound. Medicina (B Aires), 2014, p. 150-155.

[46] **Rudzkis R., Bakshaev A.** Goodness of fit tests based on kernel density estimators. Informatica, Vol. 24, Issue 3, 2013, p. 447-460.

Indre Drulyte received Applied Mathematics Master degree at Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences at Kaunas University of Technology in 2012, Lithuania. Now she is the second year Ph.D. student in Prof. K. Baršauskas Ultrasound Research Institute at Kaunas University of Technology, Lithuania. Her current research interests include solutions for clinical decision support in dermatology, data analysis and mathematical statistics.



Tomas Ruzgas received Mathematics Ph.D. degree at Mathematics and Informatics Institute at Vilnius Gediminas Technical University in 2007, Lithuania. Now he is working as an Associate Professor in Department of Applied Mathematics, Faculty of Mathematics and Natural Sciences at Kaunas University of Technology. His current research interests include nonparametric statistics, simulation and data analysis.

Renaldas Raisutis received Measurements Engineering Ph.D. degree at Prof. K. Baršauskas Ultrasound Research Institute at Kaunas University of Technology in 2005, Lithuania. Now he is working as a Chief Researcher in Prof. K. Baršauskas Ultrasound Research Institute at Kaunas University of Technology, Lithuania. His current research interests include fundamental and applied ultrasound, non-destructive testing, measurements, monitoring and quality control, solutions for clinical decision support and diagnostic medicine.

Skaidra Valiukeviciene received Medicine Ph.D. degree at Lithuanian University of Health Sciences in 2002 and in 2008 got a habilitated Ph.D. degree. Now she is working as a dermatovenereologists and as a Professor in Lithuanian University of Health Sciences. Her current research interests include non-invasive ultrasonic and optical equipment for clinical research of melanomas, non-melanomas skin tumours and diagnosis of diabetic foot.