Nonlinear Analysis: Modelling and Control, Vol. 20, No. 4, 585–602 ISSN 1392-5113
http://dx.doi.org/10.15388/NA.2015.4.9

Multivariate goodness-of-fit tests based on kernel density estimators

Aleksej Bakshaev, Rimantas Rudzkis

Institute of Mathematics and Informatics, Vilnius University,
Akademijos str. 4, LT-08663 Vilnius, Lithuania
aleksej.bakshaev@gmail.com; rimantas.rudzkis@mii.vu.lt

Received: August 25, 2014 / **Revised:** January 14, 2015 / **Published online:** September 22, 2015

Abstract. The paper is devoted to multivariate goodness-of-fit ests based on kernel density estimators. Both simple and composite null hypotheses are investigated. The test statistic is considered in the form of maximum of the normalized deviation of the estimate from its expected value. The produced comparative Monte Carlo power study shows that the proposed test is a powerful competitor to the existing classical criteria for testing goodness of fit against a specific type of an alternative hypothesis. An analytical way to establish the asymptotic distribution of the test statistic is discussed, using the approximation results for the probabilities of high excursions of differentiable Gaussian random fields.

Keywords: goodness-of-fit, kernel density estimator.

1 Introduction

The goodness of fit problem for testing whether i.i.d. random variables $X^n = (X_1, \dots, X_n)$ are distributed according to the pre-specified distribution F has been well-studied in the literature, and a variety of methods have been proposed. However, the choice of the most efficient test, among the available criteria, is regarded as one of the basic problems in statistics. The topicality of creation of new test procedures is also justified by the absence of the uniformly most powerful tests for a variety of problems arising in practice.

Since Pearson criteria, goodness-of-fit tests have been developed mostly for univariate distributions and, except for the case of normality, much less attention in literature was paid to multivariate tests of fit. In contrast to the classical approaches based on the empirical distribution function $F_n(x)$, e.g. Kolmogorov–Smirnov, Cramer–von Mises, Anderson–Darling criteria, in this paper, we consider the tests based on the kernel density estimator. The idea of using nonparametric kernel density estimators for goodness-of-fit tests goes back to Bickel and Rosenblatt [5, 23]. Since that time the approach has been extended and attracted plenty of attention in statistical papers adjusting the methodology for different practical purpose, see, for example [1, 6, 28] and references therein. A comprehensive review of the goodness of fit density tests is presented in the recent paper by Gonzalez et al. [12]. With some exceptions, see [18], particular attention in the overwhelming majority of mentioned works were devoted to the criteria based on L_p, $p = 1, 2$, distance between the density estimate $\widehat{f}(x) = \widehat{f}(x, X^n)$ and its expected value under the null hypothesis. This is presumably explained by the more competitive performance of integrated distance tests against a wide range of practical alternatives in comparison with sumpremum-type density tests, which are the objective of this paper.

This work is a continuation of our research started in [28] with the main objective to generalize the proposed test procedure to the multivariate case. The test statistic is considered in the form of maximum of the normalized deviation of the density estimate $\widehat{f}(x)$ from its hypothetical expected value. Consideration of the deviations in the uniform metric is justified by the investigation of a specific type of alternative hypothesis, where the null distribution is contaminated with a small tight cluster. Similar "sharp peak" alternatives in univariate case were also considered in [11, 18].

In the multivariate case, the efficient use of kernel estimators requires an appropriate choice of the kernel function $K(\cdot)$ and smoothing matrix W. In contrast to the kernel, the selection problem of the bandwidth matrix is much more critical, since under- or over-smoothing can substantially reduce the precision. In this work, a certain method is proposed to avoid the stated problem. It is suggested to consider the test statistic with different choices of smoothing matrices and make thereby a decision of rejecting the null hypothesis, based on the maximum of statistics values with respect to W.

In practice, the critical region of the test is found by means of Monte Carlo simulations. The problem of analytical approximation of the distribution of the test statistic under the null hypothesis is discussed, using the theory of high excursions of Gaussian (and, in some sense, close to Gaussian) random fields developed in [25, 27].

At the end of the paper, a comparative Monte Carlo power study of the proposed test is presented. The analyzed test is compared with some general classical criteria: Kolmogorov–Smirnov, Cramer–von-Mises, Anderson–Darling and Bickel–Rosenblatt, using the specific type of alternative hypothesis. The behavior of the tests is examined in bivariate, simple and composite hypotheses cases with the standard normal and $\chi^2(2)$ distributions as the null ones. In the case of composite hypothesis with the Gaussian null distribution, the comparative study was also extended with some multivariate normality specific tests, i.e. BHEP, Mardia and Mahalanobis, see [13, 14, 19, 21]. The results of simulations show that the proposed test is a powerful competitor to the existing classical ones.

2 Statement of the problem. Simple hypothesis

Let $X_1, \dots, X_n$ be a sample of independent observations of a random vector X with an unknown probability density function $f(x)$, $x \in \mathbb{R}^d$, $d > 1$. Based on the given sample, it is required to test a simple hypothesis of goodness-of-fit

$$\mathrm{H}_0\colon \quad f(x) = f_0(x)$$

against the complex alternative

$$\mathrm{H}_1\colon \quad f(x) = (1-\epsilon) f_0(x) + \epsilon g(x), \tag{1}$$

where $f_0(x)$ is a pre-specified probability density function, ϵ is a small enough fixed value, i.e. $0 < \epsilon \ll 1$, and $g(x)$ is an arbitrary distribution concentrated on a small d-dimensional interval, e.g. $\sigma_i^g \ll \sigma_i^{f_0}$, $i = 1, \ldots, d$, where σ_i^f is a standard deviation of the ith component of a random vector with the density function f.

The form of the alternative is of a particular interest in some social and economic studies, e.g. determination of small high income clusters of people, in population income distribution. Meaningful applications could be also achieved in a multivariate case, dealing with the problem of detection of tight clusters in multimodal distributions. At the same time, the tightness of the distribution g justifies the usage of a uniform metric as the loss function for $\widehat{f}(x)$. Following the outline of the work [28], we consider a test statistic based on the Parzen–Rosenblatt kernel density estimator of f, which, in the multivariate case, has the form

$$\widehat{f}_W(x) = \frac{1}{n|W|} \sum_{i=1}^{n} K\bigl(W^{-1}(x - X_i)\bigr), \tag{2}$$

where $K(\cdot)$ is the kernel function, W is a smoothing $d \times d$ symmetric and positive definite matrix, and $|W|$ is its determinant. Then the generalization of the univariate case is straightforward and the test statistic is presented in the form

$$\zeta(W) = \max_{x \in I} \bigl|\xi_W(x)\bigr|, \tag{3}$$

where

$$\xi_W(x) = \frac{\widehat{f}_W(x) - m_W(x)}{\sigma_W(x)}, \tag{4}$$

and I is a fixed d-dimensional interval. Here $m_W(x) = \mathbf{E}_0 \widehat{f}_W(x)$ and $\sigma_W^2(x) = \mathrm{Var}(\widehat{f}_W(x))$ denote a mathematical expectation and variance defined in the case of the null hypothesis.

In a nonparametric approach, the efficient use of kernel estimators requires an appropriate choice of the kernel $K(\cdot)$ and the bandwidth matrix W. There is a wide range of kernels commonly used in practice, e.g. uniform, Epanechnikov, Gaussian, and others. In contrast to the smoothing matrix, the selection problem of the kernel is much less important. Due to a small loss of efficiency for the kernels listed above, usually the choice is based on the convenience of utilization. Therefore, further in our study, primarily to simplify the calculations, we restrict ourselves to the usage of the Gaussian kernel only.

At the same time, the problem of selecting an optimal bandwidth matrix W is crucial, since it controls the degree of smoothing, applied to the data, and the precision of estimation thereby. Bandwidth selection procedures have attracted much attention of the researchers over the past decades. Among the most popular approaches, plug-in and cross-validation methods could be mentioned [2, 3, 4, 6, 9, 17, 19, 20, 21]. However, in practice, the usage of the fixed bandwidth W for kernel density estimation, even selected optimally, in a certain sense, still does not solve the problem. That is obvious dealing with the stated type of alternatives, where the contaminated cluster is much tighter in comparison with the null distribution. This fact suggests us to consider the test statistic ζ_W with different choices of a smoothing matrix W and make a decision thereby of rejecting the null hypothesis, based on the maximum of ζ_W values with respect to W. As a result, we have the following improved test statistics presented in two alternative standardization forms:

$$M_1 = \max_{W \in J} \frac{\zeta(W) - \mu(W)}{\gamma(W)}, \tag{5}$$

where

$$\mu(W) = \mathbf{E}_0 \zeta(W), \qquad \gamma^2(W) = \mathrm{Var}\big(\zeta(W)\big); \tag{6}$$

$$M_2 = \max_{W \in J} \frac{\zeta(W)}{c_\alpha(W)}. \tag{7}$$

Here $c_\alpha(W)$, $0 < \alpha < 1$, is calculated from

$$\mathbf{P}_0\big(\zeta(W) > c_\alpha(W)\big) = \alpha, \tag{8}$$

where $\mathbf{P}_0$ is a probability distribution corresponding to the null hypothesis and α is selected similarly to tests significance levels, e.g. $\alpha = 0.05$. Here the maximum with respect to W is calculated in a set J, symmetric and positive definite $d \times d$ matrices, defined by a researcher. In practice, under the additional restrictions on the smoothing matrix W, e.g. $W = \mathrm{diag}(h_1, \ldots, h_d)$, the set J could be reduced to a certain multivariate interval, i.e. $(h_1, \ldots, h_d) \in J \subset \mathbb{R}^d$.

Further, if the concrete form of the proposed statistics is not important in the discussion, we will use the common notation M for both analyzed statistics M_i, $i = 1, 2$.

Naturally, we should reject the null hypothesis in the case of large values of the test statistics, i.e, if $M > z_\alpha$, where z_α can be found from the equation

$$\mathbf{P}_0(M > z_\alpha) = \alpha. \tag{9}$$

Here α is a significance level of the test. In practice, the functions $\mu(W)$, $\gamma(W)$ and $c_\alpha(W)$ could be defined instead of formulas (6), (8) by means of further presented formulas (14), (15), obtained using the analytical approximations provided in [25], applied to the random function $\widehat{f}_W(\cdot)$.

3 Analytical approximation of the null distribution of the statistic $\zeta(W)$

In practice, the critical region of the proposed test could be defined by means of Monte Carlo simulations. An alternative approach refers to the establishment of the asymptotic null distribution of the test statistic, which is the objective of this section. Further, we discuss the problem of analytical approximation of the distribution of statistic (3), which could be used as a test statistic with a fixed choice of the smoothing matrix W. The problem is investigated using the theory of high excursions of Gaussian (and, in some sense, close to Gaussian) random fields introduced in [25, 27]. In addition, the obtained results lead to the analytical approximations of functions $\mu(\cdot)$, $\gamma(\cdot)$ and quintile $c_\alpha(\cdot)$ used in the definitions of the statistics (5) and (7), respectively.

First, we are concerned with the asymptotics of the probability

$$P_W(u) = \mathbf{P}_0\bigl\{\zeta(W) < u\bigr\}, \quad n \to \infty, \tag{10}$$

representing the distribution function of $\zeta(W)$. The fact that $\widehat{f}_W(x)$, $x \in \mathbb{R}^d$, is close to the Gaussian random field, in a certain sense, suggests us to apply the mentioned results from the theory of high excursions of Gaussian fields to approximate the probability $P_W(u)$.

It has been shown that if a differentiable (in the mean square sense) Gaussian random field $\{\eta(t), t \in T\}$ with $\mathbf{E}\eta(t) \equiv 0$, $\mathrm{Var}(\eta(t)) \equiv 1$ and continuous trajectories defined on the d-dimensional interval $T \subset \mathbb{R}^d$ satisfies certain smoothness and regularity conditions [25, Thm. 1], then $\mathbf{P}\{-v_1(t) < \eta(t) < v_2(t),\ t \in T\} \cong \mathrm{e}^{-Q}$ as for all $t \in T$, $v_1(t), v_2(t) > \chi$, $\chi \to \infty$, where $v_i(\cdot)$, $i = 1, 2$, are smooth enough functions and Q is a certain constructive functional depending on v_1, v_2, T and the matrix function $R(t) = \mathrm{cov}(\eta'(t), \eta'(t))$. Here $\eta'(t)$ is the gradient of $\eta(t)$. The stated result leads to the following approximation of the probability $P_W(u)$:

$$P_W(u) = \mathbf{P}_0\bigl\{\zeta(W) < u\bigr\} \cong \mathrm{e}^{-2Q(u)} =: \widehat{P}_W(u), \tag{11}$$

where Q depends on u, I and the matrix function $R(x) = \mathrm{cov}(\xi'_W(x), \xi'_W(x))$.

To define the functional Q in the general case $x \in \mathbb{R}^d$, let us introduce some additional notation. $N = \{1, \dots, d\}$. Assume that

$$I = \bigl\{x = (x_1, \dots, x_d)^\top\colon\ a_i \leqslant x_i \leqslant b_i,\ i = \overline{1, d}\bigr\}, \quad a_{(\cdot)} < b_{(\cdot)}. \tag{12}$$

For any $z \in \mathbb{R}$ and set $D \subset \mathbb{R}$, let $\delta_z(D) = \mathbf{1}_{\{z \in D\}}$,

$$\lambda_I(\mathrm{d}x) = \prod_{i \in N} \lambda_i(\mathrm{d}x_i) := \lambda_1(\mathrm{d}x_1) \times \dots \times \lambda_d(\mathrm{d}x_d),$$

where $\lambda_i(\mathrm{d}x_i) = \mathrm{d}x_i + \delta_{a_i}(\mathrm{d}x_i) + \delta_{b_i}(\mathrm{d}x_i)$,

$$J = J_x = \{i\colon\ a_i < x_i < b_i,\ i \in N\},$$

$$\mathbb{Y}_{i,x} = \begin{cases} \{0\}, & i \in J, \\ [0, \infty), & x_i = b_i, \\ (-\infty, 0], & x_i = a_i, \end{cases} \qquad \mathbb{Y}_x = \mathbb{Y}_{1,x} \times \dots \times \mathbb{Y}_{d,x},$$

$$\lambda^*_x(\mathrm{d}y) = \prod_{i \in N \setminus J} \mathrm{d}y_i,$$

and $\lambda^*_x(\mathbb{Y}_x) = 1$ if $J = N$.

Then

$$Q(u) = \int\limits_{I} \lambda_I(\mathrm{d}x) \int\limits_{\mathbb{Y}_x} \lambda_x^*(\mathrm{d}y) \int\limits_{u}^{\infty} \phi(z)\phi\bigl(y \mid R(x)\bigr)\bigl|zR(x)\bigr|_J \,\mathrm{d}z. \tag{13}$$

Here, for an arbitrary d-dimensional matrix $B = [B_{i,j}]$, denote $B_J = [B_{i,j}]_{i,j\in J}$ and $B_\emptyset = 1$; by $\phi(\cdot)$ and $\phi(\cdot \mid R)$ we denote the probability density functions of normal distributions $N(0,1)$ and $N(0,R)$, respectively.

Further, from the definition of $\widehat{f}_W(x)$, we obtain by means of easy calculations the following exact expression for the covariance matrix $R(x)$:

$$\begin{aligned} R(x) = {}& \frac{1}{n|W|\sigma_W^2(x)} \int \bigl(W^{-1}K'(y)\bigr)\bigl(W^{-1}K'(y)\bigr)^{\top} f_0(x+Wy)\,\mathrm{d}y \\ &- \frac{m_W'(x)m_W'^{\top}(x)}{n\sigma_W^2(x)} - \frac{(\sigma_W^2(x))'(\sigma_W^2(x))'^{\top}}{4\sigma_W^4(x)}, \end{aligned}$$

$$m_W(x) = \mathbf{E}_0\widehat{f}_W(x) = \int K(y)f_0(x+Wy)\,\mathrm{d}y,$$

$$\sigma_W^2(x) = \mathrm{Var}\bigl(\widehat{f}_W(x)\bigr) = \frac{1}{n|W|}\int K^2(z)f_0(x+Wz)\,\mathrm{d}z - \frac{m_W^2(x)}{n},$$

$$\bigl(\sigma_W^2(x)\bigr)' = \frac{1}{n|W|}\int K^2(y)f_0'(x+Wy)\,\mathrm{d}y - \frac{2m_W(x)m_W'(x)}{n}.$$

Here we assume that, for any functions $f(x)$, $g(x)$ with a multivariate argument $x \in \mathbb{R}^d$, f' is the gradient (vector-column) of f and

$$\int f'(x)g(x)\,\mathrm{d}x := \biggl(\int \frac{\partial f}{\partial x_1} g(x)\,\mathrm{d}x, \ldots, \int \frac{\partial f}{\partial x_d} g(x)\,\mathrm{d}x\biggr)^{\top}.$$

Finally, from (6), we can obtain approximations for the functions $\mu(W)$ and $\gamma(W)$ using the formulas

$$\widehat{\mu}(W) = \int u\,\mathrm{d}\widehat{P}_W(u), \qquad \widehat{\gamma}^2(W) = \int u^2\,\mathrm{d}\widehat{P}_W(u) - \widehat{\mu}^2(W). \tag{14}$$

Next, a special case of $d = 2$ is investigated. We also provide graphical examples of the accuracy of the proposed approximations comparing the empirical distribution function of statistic ζ_W (3) and the asymptotic distribution function (11).

Let $x = (x_1, x_2) \in \mathbb{R}^2$, $R = R(x_1, x_2) = \mathrm{cov}(\xi_W'(x), \xi_W'(x))$ be the covariance matrix of the bivariate random field $\xi_W'(x)$ with elements $R_{i,j} = R_{i,j}(x_1, x_2)$, $i, j = 1, 2$,

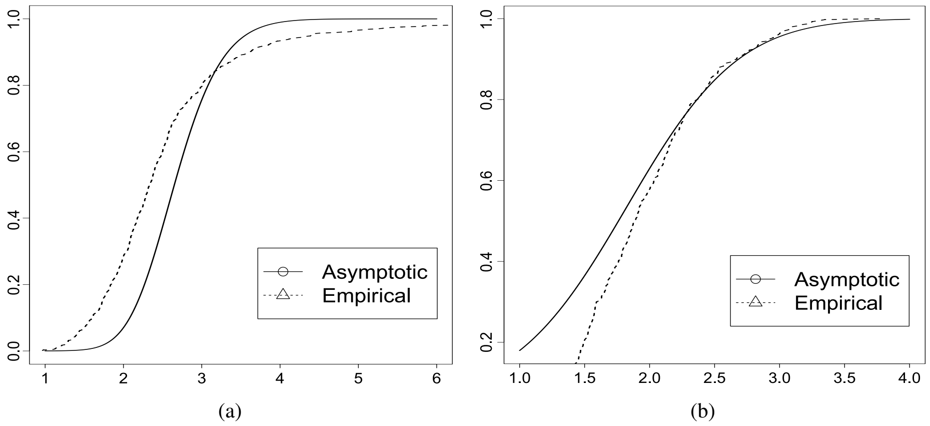

Fig. 1. Empirical and asymptotic distributions of $\zeta(W)$, $n = 2000$, $W = \operatorname{diag}(0.3, 0.3)$: (a) $I = [0,3] \times [0,3]$, (b) $I = [0,1] \times [0,1]$.

and $I = I_1 \times I_2 = [a_1, b_1] \times [a_2, b_2]$. Afterwards, the functional Q can be written in the form

$$\begin{aligned} Q(u) &= \frac{1}{2\pi}\bigl(1 - \varPhi(u) + u\phi(u)\bigr) \int\limits_I |R|^{1/2}\,\mathrm{d}x_1\,\mathrm{d}x_2 \\ &\quad + \frac{\phi(u)}{2\sqrt{2\pi}} \int\limits_{I_1} \bigl(R_{1,1}^{1/2}(y, a_2) + R_{1,1}^{1/2}(x_1, b_2)\bigr)\,\mathrm{d}x_1 \\ &\quad + \frac{\phi(u)}{2\sqrt{2\pi}} \int\limits_{I_2} \bigl(R_{2,2}^{1/2}(a_1, x_2) + R_{2,2}^{1/2}(b_1, x_2)\bigr)\,\mathrm{d}x_2, \end{aligned}$$

For a graphical assessment of precision of the proposed approximations, we consider the case, where f is a standard normal distribution. In the simulations, the smoothing matrix W for the kernel density estimator $\widehat{f}_W(x)$ is assumed to be diagonal with the elements (h_1, h_2), where $h_1 = h_2 \in [0.2, 1]$. The empirical distribution of ζ_W was obtained by generating 500 samples of sizes 1000–5000 from the distribution f. Different variants of intervals I in the definition of the statistic ζ_W were investigated.

The experimental results show that sufficiently good accuracy of approximation for moderate and small significance levels ($\alpha < 0.1$) and all the considered bandwidth matrices could be achieved in the intervals of order $\sigma \times \sigma$, $\sigma^2 = \operatorname{Var} X$, $X \sim f$, and sample sizes $n \geqslant 2000$, which is natural for a nonparametric approach. Some simulation results for the sample size $n = 2000$ are presented in Fig. 1. The precision of approximations strongly depends on the smoothing matrix, the size of the chosen interval I, probability $\mathbf{P}\{X \in I\}$, and the sample size n. A similar precision level, obtained for $W = \operatorname{diag}(0.3, 0.3)$, $I = [0,1] \times [0,1]$ and $n = 2000$, for the case $I = [1,2] \times [1,2]$ could be achieved only when $n = 5000$. There is a strong general tendency of a growing precision of approximation with a decreasing size of the interval I observed for all bandwidth matrices considered. At the same time, consideration of relatively smaller intervals I in the tests, based on ζ_W, could be practically meaningful as the researcher might be interested in the potential discrepancies from the hypothesized distribution only in a certain pre-specified interval, obtained from a priori information.

The fact, that the proposed asymptotic distributions provide quite a good approximation to the null distribution of the statistic ζ_W only for moderate and small significance levels, is a theoretical justification for the usage of statistic M_2 (7) with quintile standardization, where $c_\alpha(W)$ could be obtained from

$$\widehat{P}_W\bigl(c_\alpha(W)\bigr) = 1 - \alpha. \tag{15}$$

4 Composite hypothesis

In addition to the simple hypothesis of goodness of fit, where all the parameters of the null distribution are specified, let us further consider the composite hypothesis variant. In this case, on the basis of the sample $X_1, \dots, X_n$, we wish to test whether the unknown density function f lies in a certain parametric family, i.e.

$$\mathrm{H}_0\colon \quad f = f_0(\cdot, \theta),$$

against an alternative

$$\mathrm{H}_1\colon \quad f(x) = (1-\epsilon) f_0(x, \theta) + \epsilon g(x),$$

where $\theta \in \Theta \subset \mathbb{R}^p$ is an unknown vector parameter and distribution $g(x)$ satisfies the conditions in (1).

The modifications of proposed statistics (3), (5) and (7) for testing composite hypothesis H_0 are quite straightforward. Since in this case expectation $m_W(x)$ and variance $\sigma_W^2(x)$ in the definition of empirical process $\xi_W(x)$ depend on the unknown parameter θ one has to replace it by $\widehat{\theta}_n$, where $\widehat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of $\theta_0 \in \Theta$, the true parameter value under H_0. In our simulation study we use maximum likelihood estimator for θ_0. Thus, the process (4) in this case has the form

$$\xi_W(x, \widehat{\theta}_n) = \frac{\widehat{f}_W(x) - m_W(x, \widehat{\theta}_n)}{\sigma_W(x, \widehat{\theta}_n)}. \tag{16}$$

Here $m_W(x, \widehat{\theta}_n) = \mathbf{E}_{\widehat{\theta}_n} \widehat{f}_W(x)$ and $\sigma_W^2(x, \widehat{\theta}_n) = \mathrm{Var}_{\widehat{\theta}_n}(\widehat{f}_W(x))$. Finally, the composite hypothesis H_0 could be verified by using statistics (5) and (7) with an estimate $\widehat{\theta}_n$ substituted for the unknown parameter θ in (6), (8) and process $\xi_W(x)$ replaced by $\xi_W(x, \widehat{\theta}_n)$ in (3).

Statistic $\xi_W(x, \widehat{\theta}_n)$ may be regarded as an analogue of a modified Kolmogorov–Smirnov statistic based on the empirical cumulative distribution function $F_n(x)$ and used for testing whether an unknown distribution function lies in a certain parametric family of distribution functions. Recall that in the case of the simple hypothesis, the critical region of the test based on (3) is defined by $\zeta_h > z_\alpha$, where z_α is a cutoff point obtained from the proved approximation (11), i.e. $\widehat{P}_h(z_\alpha) = 1 - \alpha$ and α is the significance level of the test. Unlike Kologorov–Smirnov statistic with estimated parameter, where the limit distribution under the null hypothesis depends on the distribution of $\widehat{\theta}_n$, under some additional assumptions, e.g. partial derivatives of f_0 with respect to θ_i, $i = 1, \dots, p$, are bounded for θ in a neighborhood of θ_0 for all x (for more details, see [5,18,24]), the same cutoff points may be used for testing the specified composite hypothesis using statistic $\xi_W(x, \widehat{\theta}_n)$. This is due to the fact that the normalizing factor $\sqrt{n}$ in the Kolmogorov–Smirnov test statistic is of the same order as the square of the rate of consistency of the parameter estimator, while in the density test case this factor is $\sqrt{n|W|}$ which tends to infinity slower. Here we assume the smoothing matrix $W = W_n$ tends to zero as $n \to \infty$ and $n^{-1} = o(W)$, where o notation is applied elementwise. Therefore, the parameter estimator's influence can be neglected in the problem presented here.

5 Simulation study

After the introduction of a certain test procedure, it is practically important to establish whether the proposed test is powerful enough in comparison with the well-known criteria. This section is aimed at the comparative Monte Carlo power study, where analyzed tests (5)–(7) are compared with some classical criteria: Anderson–Darling (AD), Cramer–von Mises (CM), Kolmogorov–Smirnov (KS) and Bickel–Rosenblatt (BR). A brief description of all considered tests is presented below. Recall that the power of a statistical test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true.

In our empirical analysis, we consider a simple hypotheses of goodness of fit with Gaussian and $\chi^2(2)$ null distributions and composite hypothesis of normality in bivariate case, i.e.

$$\mathrm{H}_0: \quad f(x) = f_0(x) \quad \text{or} \quad f(x) = f_0(x, \theta),$$

where

- $f_0 \sim N(0, \Sigma)$ and Σ is assumed to be a unit matrix;
- $f_0 \sim (Y_1, Y_2)$, Y_i, $i = 1, 2$, are independent univariate random variables with the χ^2 distribution with two degrees of freedom;

against an alternative

$$\mathrm{H}_1: \quad f(x) = (1 - \epsilon) f_0(x) + \epsilon g(x),$$

where $g \sim N(m, \sigma^2 \Sigma)$, $\sigma \ll 1$ and ϵ is small.

In our comparative study, we examine the behavior of (5) and (7) tests, where functions $\mu(W)$, $\gamma^2(W)$ and $c_\alpha(W)$, $\alpha = 0.05$, are calculated using the proved approximations (14)–(15). In addition, the following versions of M_1 and M_2 test statistics are investigated:

$$M_1^* = \max_{W \in J} \frac{\zeta(W) - \widehat{\mu}^*(W)}{\widehat{\gamma}^*(W)}, \tag{17}$$

$$M_2^* = \max_{W \in J} \frac{\zeta(W)}{\widehat{c}_\alpha^*(W)}, \tag{18}$$

where the functions $\widehat{\mu}^*(W)$, $\widehat{\gamma}^*(W)$ and $\widehat{c}_\alpha^*(W)$, $\alpha = 0.05$, are obtained by Monte Carlo simulations.

The main objective of the study is to compare the sensitivity of the proposed tests to the given type of alternative with the general criteria listed above. However, in the composite hypothesis case, where the null distribution f is Gaussian, we have added several multivariate normality criteria to our analysis, i.e. BHEP [13, 14, 21], Mardia [19] and Mahalanobis criteria. In this case, statistics (5), (7), (17), (18) are applied with the estimated mean vector and covariance matrix of the Gaussian distribution using the maximum likelihood estimator and procedure described in Section 4.

5.1 Classical tests used in the comparison study

In this subsection, we provide the specifications of the mentioned best-known multivariate goodness of fit tests further applied in our simulation comparative study. It is assumed that a sample of independent observations $X_1, \ldots, X_n$ of random variable X with unknown distribution function $F(x)$, $x \in \mathbb{R}^d$, is available for the researcher. For simplicity, we present the test statistics for the case of simple hypothesis of goodness of fit with a given, completely known distribution F_0 as a null one. The modifications of statistics for the case of composite hypothesis is done in the same way described in Section 4.

In the case of continuous distributions, the most popular general tests used to verify the stated hypothesis are based on the empirical distribution functions $F_n(x)$. The most famous and well-studied statistic of this type is obviously Kolmogorov–Smirnov statistic [17, 31]

$$D_n = \sqrt{n} \sup_x \bigl|F_0(x) - F_n(x)\bigr|.$$

Another group of statistics is based on the integral distance between F_0 and F_n. The best known among them is Cramer–von Mises statistic (see [4, 20])

$$\omega_n^2 = n \int \bigl(F_0(x) - F_n(x)\bigr)^2 \, \mathrm{d}F_0(x).$$

Anderson and Darling in [2,3] proposed to improve the properties of the presented statistic by introducing a non-negative weight function, used in this statistic in order to vary the contribution of the deviations of the empirical distribution function from the theoretical distribution function in different ranges of its argument. One of the most wide spread variant of ω_n^2 statistic with weight has the form

$$A_n^2 = n \int \frac{(F_0(x) - F_n(x))^2}{F_0(x)(1 - F_0(x))} \, \mathrm{d}F_0(x).$$

In contrast to univariate case, the probability distribution functions of statistics D_n, ω_n^2 and A_n^2 in multivariate case are not distribution-free. However, this problem could be overcome by means of Rosenblatt transformation (see [22]), which transforms an absolutely continuous d-variate distribution into the uniform distribution on the d-dimensional cube. Let sample $Y_1, \dots, Y_n$ be the result of the Rosenblatt transformation of the sample $X_1, \dots, X_n$. Thus, intial goodness of fit problem could be reduced to testing the unifomity on the hipercube based on $Y_1, \dots, Y_n$.

In the general case $d \geqslant 2$, all the statistics above are characterized by considerable computational difficulties, therefore, in our simulation study ($d = 2$), we use some convenient for calculation formulas of the form

$$\widehat{D}_n = \sqrt{n} \sup_{y \in \{Y_1, \dots, Y_n\}} \bigl| y_1 y_2 - F_n^*(y) \bigr|,$$

$$\widehat{\omega}_n^2 = \sum_{i=1}^{n} \left(\frac{r_i}{n} - Y_{1,i} Y_{2,i} \right)^2, \qquad \widehat{A}_n^2 = \sum_{i=1}^{n} \frac{(\frac{r_i}{n} - Y_{1,i} Y_{2,i})^2}{(1 - Y_{1,i} Y_{2,i}) Y_{1,i} Y_{2,i}},$$

where $Y_i = (Y_{1,i}, Y_{2,i})$, $i = 1, \dots, n$, $r_i = nF_n^*(Y_i) + 1$ and $F_n^*(\cdot)$ is the empirical distribution function of the tranformed sample. All the presented formulas reflect statistics, which are asimptotically equavalent to the original ones (for more details, see [16, 20]).

In addition to proposed statistics (5)–(7), the class of tests based on kernel density estimators is also presented in the study by Bickel–Rosenblatt test [5, 23] reflecting the L_2 distance between the density estimate $\widehat{f}(x)$ and its expected value under the null hypothesis

$$T_n = \int \frac{(\widehat{f}_W(x) - \mathbf{E}_0 \widehat{f}_W(x))^2}{f_0(x)} \, \mathrm{d}x,$$

where $W = W(n)$ is a smoothing parameter, optimally chosen in the sense of minimization of the asymptotic mean integrated square error (AMISE). In bivariate case, i.e. $X_i = (X_{1,i}, X_{2,i})$, $i = 1, \dots, n$, Gaussian null hypothesis and under the condition $W = \operatorname{diag}(h_1, h_2)$, test statistic could be calculated using the formula

$$T_n = \frac{1}{n^2 |W|^2} \sum_{i,j=1}^{n} L_1(X_{1,i}, X_{1,j}, h_1) L_1(X_{2,i}, X_{2,j}, h_2) - 1,$$

where

$$L_1(x_1, x_2, h) = \frac{h}{\sqrt{2 - h^2}} \mathrm{e}^{-1/(2h^2)} \left(x_1^2 + x_2^2 - \frac{(x_1 + x_2)^2}{2 - h^2} \right).$$

In the case of considered $\chi^2(2)$, null distribution test statistic has the form

$$\begin{aligned} T_n = {} & \frac{1}{n^2 |W|^2} \sum_{i,j=1}^{n} L_2(X_{1,i}, X_{1,j}, h_1) L_2(X_{2,i}, X_{2,j}, h_2) \\ & - \frac{2}{n} \sum_{i=1}^{n} \left(1 - \Phi\left(-\frac{X_{1,i}}{h_1} \right) \right) \left(1 - \Phi\left(-\frac{X_{2,i}}{h_2} \right) \right) + 1, \end{aligned}$$

where

$$L_2(x_1, x_2, h) \\ = \frac{h}{\sqrt{\pi}}\left[1 - \varPhi\left(-\frac{x_1 + x_2 + \frac{h^2}{2}}{\sqrt{2}h}\right)\right]\mathrm{e}^{-1/(2h^2)}\left(x_1^2 + x_2^2 - \frac{(x_1 + x_2 + \frac{h^2}{2})^2}{2}\right).$$

In the case of composite hypothesis with Gaussian null distribution, the comparative study is extended with some popular normality tests. The first one is based on the comparison of the sample moments with the theoretical ones. In univariate case, standardized third and fourth moments are often used to indicate distribution skewness and kurtosis. For a given sample with sample mean $\bar{X}$ and sample covariance matrix $\widehat{S}$, Mardia in [19] defined the d-variate skewness and kurtosis statistics as

$$b_1 = \frac{1}{n^2}\sum_{i,j=1}^{n}\left[(X_i - \bar{X})^\top \widehat{S}^{-1}(X_j - \bar{X})\right]^3$$

and

$$b_2 = \frac{1}{n}\sum_{i=1}^{n}\left[(X_i - \bar{X})^\top \widehat{S}^{-1}(X_i - \bar{X})\right]^2,$$

respectively. Under the multinormality, b_1 and b_2 are affine invariant, asymptotically independent and the limiting distribution of $n(b_1/6)$ and $\sqrt{n}(b_2 - d(d+2))/\sqrt{8d(d+2)}$ are a chi-squared distribution with $d(d+1)(d+2)/6$ degrees of freedom and $N(0,1)$ distribution respectively. Finally, the null hypothesis should be rejected in the case of the large values of statistic $n(b_1/6) + n(b_2 - d(d+2))^2)/(8d(d+2))$.

Another well-known normality criteria, called BHEP (see [13, 14, 21]), is based on the empirical characteristic function $\varPsi_n(t) = (1/n)\sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}t^T Y_k}$, where $Y_k = \widehat{S}^{-1/2} \times (X_k - \bar{X})$, $k = 1, \dots, n$, is the standardized sample. The test statistic $B_n(\beta)$ is the weighted integral of the squared difference between the multivariate normal characteristic function and the empirical characteristic function, i.e.

$$B_n(\beta) = n\int\left(\varPsi_n(t) - \mathrm{e}^{-\|t\|^2/2}\right)^2 \varphi_\beta(t)\,\mathrm{d}t,$$

where $\varphi_\beta(t) = (2\pi\beta^2)^{-d/2}\mathrm{e}^{-\|t\|^2/(2\beta^2)}$ is a weight function with optimal parameter β value in bivariate case equals to $(1/\sqrt{2})(5n/4)^{1/6}$. In our simulation study, statistic $B_n(\beta)$ is calculated using the formula

$$B_n(\beta) = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathrm{e}^{-(\beta^2/2)\|X_i - X_j\|^2} \\ - \frac{2}{(1+\beta^2)n}\sum_{i=1}^{n}\mathrm{e}^{-\beta^2\|X_i\|^2/(2(1+\beta^2))} + \frac{1}{1+2\beta^2}.$$

The last normality criteria in our study is based on Mahalanobis transformation of initial sample

$$Y_i = (X_i - \bar{X})^\top \widehat{S}^{-1/2}(X_i - \bar{X}), \quad i = 1, \dots, n.$$

Transformed univariate sample has chi-squared limiting distribution with d degrees of freedom. After that the null hypothesis should be rejected in the case of large values of one-sample Kolmogorov–Smirnov statistic.

5.2 Simulation design

In the study, the smoothing matrix W is considered to be diagonal with the elements (h_1, h_2). Recall that the kernel $K(\cdot)$ in the density estimator (2) is assumed to be Gaussian.

In all the cases, we investigate the behavior of the above mentioned tests for sample sizes $n = 1000, 2000$ and the significance level $\alpha = 0.05$. The maxima with respect to x and W in the considered test statistics are calculated using the following intervals:

- $J = [0.2, 1] \times [0.2, 1]$;
- $I = [0, 3] \times [0, 3]$ (normality case);
  $I = [0, 10] \times [0, 10]$ (χ^2 case).

The interval J considered for the selection of the smoothing parameters h_i, $i = 1, 2$, in the kernel density estimator includes a wide range of choices of the bandwidth W, obtained by applying the most common optimality criteria.

The critical regions of both classical and investigated tests were established using the finite sample null distribution of corresposning test statistics, including proposed (5),(7),(17),(18), obtained by the Monte Carlo method. The power of the tests was estimated by simulating 1000 samples of alternative distributions $(1-\epsilon)f_0(x)+\epsilon g(x)$, where $g \in N(m, \sigma^2\Sigma)$, $m = (m_1, m_2)$, $\epsilon \sim Unif(0.01, 0.05)$. The following characteristics of the mixing distribution g were investigated:

- Σ is a unit matrix;
- $m_1 = m_2 \sim \mathrm{Unif}(0, 3)$ (normality case);
  $m_1 = m_2 \sim \mathrm{Unif}(0, 10)$ (χ^2 case);
- $\sigma \sim \mathrm{Unif}(0.1, 0.5)\ /\ \mathrm{Unif}(0.1, 0.3)\ /\ \mathrm{Unif}(0.1, 0.2)$.

The simulations of alternative samples were produced in two steps. First m, σ and ϵ values were generated, then the obtained values were used for simulating the sample from the distribution $(1-\epsilon)f_0(x) + \epsilon g(x)$.

Considered several variants of tightness of the distribution cluster g and mixing probabilities ϵ in (1), give us a wide range of deviations from the null hypothesis and allow us to test the sensitivity of criteria to each of them.

5.3 Summary of the simulation results

The empirical results summarized in Tables 1–3 illustrate that the criteria proposed are powerful competitors not only to the classical general tests, but also to specific normality criteria in the goodness-of-fit problem against a complex alternative, when a hypothesized distribution is contaminated with a small tight cluster.

Table 1. Percentage of the rejected simple hypothesis of normality, $m \sim \mathrm{Unif}(0, 3)$: (a) $\sigma \sim \mathrm{Unif}(0.1, 0.5)$, (b) $\sigma \sim \mathrm{Unif}(0.1, 0.3)$, (c) $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

Tests	$n = 1000$			$n = 2000$		
	(a)	(b)	(c)	(a)	(b)	(c)
KS	23	23	24	48	50	53
CM	20	20	20	40	41	41
AD	17	18	19	37	38	39
BR	31	32	32	42	43	44
M_1	52	60	68	70	79	87
M_1^*	65	71	71	77	82	84
M_2	53	62	68	69	79	86
M_2^*	66	73	74	77	85	87

Table 2. Percentage of the rejected simple hypothesis of χ^2, $m \sim \mathrm{Unif}(0, 10)$: (a) $\sigma \sim \mathrm{Unif}(0.1, 0.5)$, (b) $\sigma \sim \mathrm{Unif}(0.1, 0.3)$, (c) $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

Tests	$n = 1000$			$n = 2000$		
	(a)	(b)	(c)	(a)	(b)	(c)
KS	26	27	27	45	47	49
CM	23	23	23	39	40	41
AD	24	24	24	38	38	38
BR	42	42	43	49	50	50
M_1	53	59	66	72	80	81
M_1^*	65	70	70	85	86	87
M_2	54	59	66	74	80	81
M_2^*	66	70	71	85	87	88

Table 3. Percentage of the rejected composite hypothesis of normality, $m \sim \mathrm{Unif}(0, 3)$: (a) $\sigma \sim \mathrm{Unif}(0.1, 0.5)$, (b) $\sigma \sim \mathrm{Unif}(0.1, 0.3)$, (c) $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

Tests	$n = 1000$			$n = 2000$		
	(a)	(b)	(c)	(a)	(b)	(c)
KS	35	38	39	57	67	73
CM	45	49	52	59	68	71
AD	46	49	52	57	65	68
BR	33	35	36	50	51	53
BHEP	53	60	64	70	81	84
Mardia	47	47	47	52	54	54
Mahalanobis	27	28	28	46	49	53
M_1	43	56	64	61	82	89
M_1^*	55	63	65	69	83	87
M_2	44	57	64	61	82	89
M_2^*	57	68	69	69	86	90

Recall that we are using the same notation M for all the analyzed statistics M_i, $i = 1, 2$, in case the specific form of the statistics is not important for us.

The general behavior of all the proposed tests M is characterized by an obvious tendency. We observe the increasing power of the test, if the tightness of the distribution cluster g becomes smaller. That is clearly explained by the usage of the uniform metric as the loss function for $\widehat{f}(x)$. Produced additional simulations also show a significant increase in the comparative power of tests, while shifting the cluster g towards the tails of the null distribution. This fact could be justified by the application of the standardization factor $\operatorname{Var}(\widehat{f}_W(x))$ in the process $\xi_W(x)$ (4), which assigns more weight to discrepancies in the tails of the distribution.

All the proposed M tests are considerably more powerful than the target general criteria (KS, CM, AD and BR) against the investigated alternatives in all examined cases of the null distributions. Furthermore, in the composite Gaussian hypothesis case and a tight enough distribution cluster g, the performance of tests M is comparable or better than the Mardia normality test, based on multivariate measures of skewness and kurtosis, and the BHEP criterion, which reflects the L_2 weighted distance between the empirical and actual characteristic functions of Gaussian distributions. For the χ^2 alternative, reasonably competitive performance to the proposed criteria is shown only by Kolmogrov–Smirnov and Bickel–Rosenblatt tests in the case of large sample sizes. In general, the Bickel–Rosenblatt criterion, as a integrated distance test also based on the kernel density estimator, is considered to be the main competitor among the general tests in our study. However its comparison with the proposed maximum-type tests shows the significant superiority of the latter ones against the considered alternative, which was also noticed in the univariate case in [18]. Graphically the behavior of the best M_2 and M_2^* tests for sample sizes $n = 1000, 2000$ in all investigated cases of the null distributions is summarized in Figs. 2–4.

A comparison of the performance of M_i and M_i^*, $i = 1, 2$, tests shows a moderate superiority of the latter ones. It is explained mostly by the fact that the precision of the approximations (presented in Section 3) to the null distribution of the statistic ζ_W is high only for moderate and small significance levels and considerably small intervals I. This fact also justifies the behavior of the M_2 test with analytically established $c_\alpha(W)$ values, the results of which only slightly exceed that of M_1.

The simulation study on the accuracy of established approximations to the null distribution of ζ_W, see Fig. 1, suggests the following rule for the option between M_i and

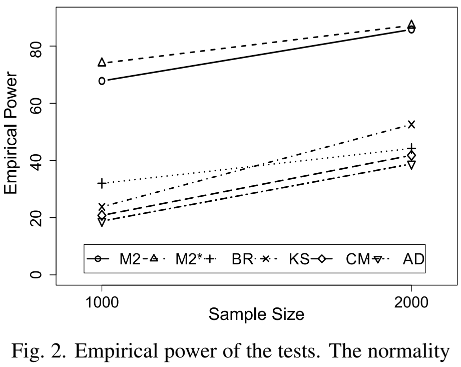

Fig. 2. Empirical power of the tests. The normality case, simple hypothesis, $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

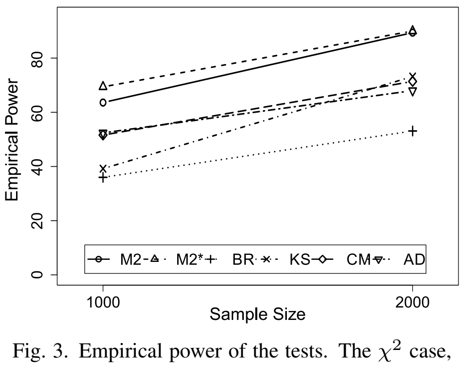

Fig. 3. Empirical power of the tests. The χ^2 case, simple hypothesis, $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

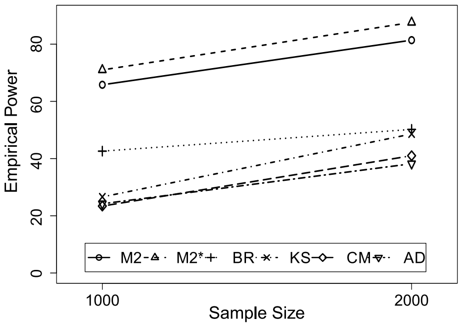

Fig. 4. Empirical power of the tests. The normality case, composite hypothesis, $\sigma \sim \mathrm{Unif}(0.1, 0.2)$.

M_i^*, $i = 1, 2$, statistics. If, based on a priori information, the multivariate interval I for detecting the contamination cluster g could be sufficiently minimized, we suggest using M_i, $i = 1, 2$, statistics, where the functions $\mu(W)$, $\gamma^2(W)$ and $c_\alpha(W)$, in their definitions are obtained from approximations (14)–(15). This certainly makes the application of the test more convenient. Otherwise, it is recommended to use M_i^*, $i = 1, 2$, tests with the mentioned functions, derived by means of Monte Carlo simulations.

6 Conclusion

In this paper, we have proposed supremum-type statistics for comparing multivariate distributions based on kernel density estimators. All considered tests are practical to apply for moderate dimension and arbitrary sample sizes. Produced Monte Carlo power study illustrates that the relative performance of proposed tests is impressive in comparison with classical criteria based on empirical distribution function, e.g. Kolmogorov–Smirnov, Anderson-Darling, integrated distance Bickel–Rosenblatt test also based on kernel density estimator and some specific normality tests, e.g. BHEP, Mardia, in testing goodness of fit, sensitive to contamination of the null distribution with small tight clusters. Furthermore it is advisable to use analytical approximations for the parameters of proposed test statistics obtained by means of the theory of large excursions of Gaussian fields.

References

1. I.A. Ahmad, P.B. Cerrito, Goodness of fit tests based on the L_2-norm of multivariate probability density functions, *J. Nonparametric Stat.*, **2**:169–181, 1993.

2. T. Anderson, D. Darling, Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes, *Ann. Math. Stat.*, **23**(2):193–212, 1952.

3. T. Anderson, D. Darling, A test of goodness of fit, *J. Am. Stat. Assoc.*, **49**(268):765–769, 1954.

4. T. Anderson, D. Darling, The Kolmogorov–Smirnov, Cramér-von Mises test, *Ann. Math. Stat.*, **28**(3):823–838, 1957.

5. P.J. Bickel, M. Rosenblatt, On some global measures of the deviations of density function estimates, *Ann. Stat.*, **1**(6):1071–1095, 1973.

6. R. Cao, G. Lugosi, Goodness-of-fit tests based on the kernel density estimate, *Scand. J. Stat.*, **32**:599–616, 2005.

7. J.E. Chacon, T. Duong, Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices, *Test*, **19**(2):375–398, 2010.

8. T. Duong, M.L. Hazelton, Plug-in bandwidth matrices for bivariate kernel density estimation, *J. Nonparametric Stat.*, **15**(1):17–30, 2003.

9. T. Duong, M.L. Hazelton, Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation, *J. Multivariate Anal.*, **93**:417–433, 2005.

10. T. Duong, M.L. Hazelton, Cross-validation bandwidth matrices for multivariate kernel density estimation, *Scand. J. Stat.*, **32**:485–506, 2005.

11. B. Ghosh, W.-M. Huang, The power and optimal kernel of the Bickel–Rosenblatt test for goodness of fit, *Ann. Stat.*, **19**(2), 999–1009, 1991.

12. W. Gonzalez-Manteiga, R.M. Crujeiras, An updated review of goodness-of-fit tests for regression models. Invited paper with discussion, *Test*, **22**:361–411, 2013.

13. N. Henze, L. Baringhaus, A consistent test for multivariate normality based on the empirical characteristic function, *Metrika*, **35**:339–348, 1998.

14. N. Henze, B. Zirkler, A class of invariant and consistent tests for multivariate normality, *J. Multivariate Anal.*, **19**:3595–3617, 1990.

15. I. Horova, J. Zelinka, Contribution to the bandwidth choice for kernel density estimates, *Comput. Stat.*, **22**:31–47, 2007.

16. A. Justel, D. Pena, R. Zamar, A multivariate Kolmogorov–Smirnov test of goodness of fit, *Stat. Probab. Lett.*, **35**:251–259, 1997.

17. A.N. Kolmogorov, On the empirical determination of a distribution law, *G. Ist. Ital. Attuari*, **4**:1–11, 1933 (in Italian).

18. H. Liero, H. Lauter, V.D. Konakov, Nonparametric versus parametric goodness of fit, *Statistics*, **31**:115–149, 1998.

19. K.V. Mardia, Measures of multivariate skewness and kurtosis with applications, *Biometrika*, **57**:519–530, 1970.

20. G. Martynov, *Omega-Square Criteria*, Nauka, Moscow, 1978.

21. L.B. Pulley, T.W. Epps, A test for normality based on the empirical characteristic function, *Biometrika*, **70**:723–726, 1983.

22. M. Rosenblatt, Remarks on multivariate transformation, *Ann. Math. Stat.*, **23**:470–472, 1952.

23. M. Rosenblatt, On the maximal deviation of k-dimensional density estimates, *Ann. Probab.*, **4**(6):1009–1015, 1976.

24. M. Rosenblatt, *Stochastic Curve Estimation*, NSF-CBMS Reg. Conf. Ser. Probab. Stat., Vol. 3, IMS, Hayward CA, 1991.

25. R. Rudzkis, Probabilities of large excursions of empirical processes and fields, *Sov. Math., Dokl.*, **45**(1):226–228, 1992.

26. R. Rudzkis, On the distribution of supremum-type functionals of nonparametric estimates of probability and spectral densities, *Theory Probab. Appl.*, **37**(2):236–249, 1992.

27. R. Rudzkis, A. Bakshaev, Probabilities of high excursions of Gaussian fields, *Lith. Math. J.*, **52**(2):196–213, 2012.

28. R. Rudzkis, A. Bakshaev, Goodness of fit tests based on kernel density estimators, *Informatica*, **24**(3):447–460, 2013.

29. S.R. Sain, K.A. Baggerly, D.W. Scott, Cross-validation of multivariate densities, *J. Am. Stat. Assoc.*, **89**:807–817, 1994.

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

31. N.V. Smirnov, Approximate laws of distribution of random variables from empirical data, *Usp. Mat. Nauk*, **10**:179–206, 1944.

32. M.P. Wand, M.C. Jones, Multivariate plug-in bandwidth selection, *Comput. Stat.*, **9**:97–116, 1994.

33. M.P. Wand, M.C. Jones, *Kernel Smoothing*, Chapman & Hall, London, 1995.