VILNIUS GEDIMINAS TECHNICAL UNIVERSITY
INSTITUTE OF MATHEMATICS AND INFORMATICS

Aleksej BAKŠAJEV

STATISTICAL TESTS BASED ON N-DISTANCES

DOCTORAL DISSERTATION

PHYSICAL SCIENCES, MATHEMATICS (01P)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2010

The scientific work was prepared at Institute of Mathematics and Informatics in 2005–2009.

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

Scientific Consultant
Prof Dr Habil Yurij TYURIN (Lomonosov Moscow State University, Physical Sciences, Mathematics – 01P).

VGTU leidyklos TECHNIKA 1715-M mokslo literatūros knyga
http://leidykla.vgtu.lt

ISBN 978-9955-28-535-9

aleksej.bakshaev@gmail.com

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS INSTITUTAS

Aleksej BAKŠAJEV

STATISTINIŲ HIPOTEZIŲ TIKRINIMAS, NAUDOJANT N-METRIKAS

DAKTARO DISERTACIJA

FIZINIAI MOKSLAI, MATEMATIKA (01P)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2010

Disertacija rengta 2004–2009 metais Matematikos ir informatikos institute.

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

Mokslinis konsultantas
prof. habil. dr. Yurij TYURIN (Maskvos valstybinis M. V. Lomonosovo universitetas, fiziniai mokslai, matematika – 01P).

Abstract

The thesis is devoted to the application of a new class of probability metrics, N-distances, introduced by Klebanov (Klebanov, 2005; Zinger et al., 1989), to the problems of verification of the classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence.

First of all a construction of statistics based on N-metrics for testing mentioned hypotheses is proposed. Then the problem of determination of the critical region of the criteria is investigated. The main results of the thesis are connected with the asymptotic behavior of test statistics under the null and alternative hypotheses. In general case the limit null distribution of proposed in the thesis tests statistics is established in terms of the distribution of infinite quadratic form of random normal variables with coefficients dependent on eigenvalues and functions of a certain integral operator. It is proved that under the alternative hypothesis the test statistics are asymptotically normal. In case of parametric hypothesis of goodness of fit particular attention is devoted to normality and exponentiality criteria. For hypothesis of homogeneity a construction of multivariate distribution-free two-sample test is proposed. Testing the hypothesis of uniformity on hypersphere S^{p-1} in more detail $p = 1, 2$ cases are investigated.

In conclusion, a comparison of N-distance tests with some classical criteria is provided. For simple hypothesis of goodness of fit in univariate case as a measure for comparison an Asymptotic Relative Efficiency (ARE) by Bahadur (Bahadur, 1960; Nikitin, 1995) is considered. In parallel to the theoretical results the empirical comparison of the power of the tests is examined by means of Monte Karlo simulations. Besides simple and composite hypotheses of goodness of fit, hypotheses of uniformity on S^1 and S^2, we consider two-sample tests in uni- and multivariate cases. A wide range of alternative hypotheses are investigated.

Reziumė

Disertacinis darbas yra skirtas N-metrikų teorijos (Klebanov, 2005; Zinger et al., 1989) pritaikymui klasikinėms statistinėms suderinamumo, homogeniškumo, simetriškumo bei nepriklausomumo hipotezėms tikrinti.

Darbo pradžioje pasiūlytas minėtų hipotezių testinių statistikų konstravimo būdas, naudojant N-metrikas. Toliau nagrinėjama problema susijusi su suformuotų kriterijų kritinės srities nustatymu. Pagrindiniai darbo rezultatai yra susiję su pasiūlytų kriterijaus statistikų asimptotiniu skirstiniu. Bendru atveju N-metrikos statistikų asimptotinis skirstinys esant nulinei hipotezei sutampa su Gauso atsitiktinių dydžių begalinės kvadratinės formos skirstiniu. Alternatyvos atveju testinių statistikų ribinis skirstinys yra normalusis. Sudėtinės suderinamumo hipotezės atveju išsamiau yra analizuojami normalumo ir ekponentiškumo kriterijai. Daugiamačiu atveju pasiūlyta konstrukcija, nepriklausanti nuo skirstinio homogeniškumo testo. Tikrinant tolygumo hipersferoje S^{p-1} hipotezę detaliau yra nagrinėjami apskritimo $p = 1$ ir sferos $p = 2$ atvejai.

Darbo pabaigoje lyginami pasiūlytos N-metrikos bei kai kurie klasikiniai kriterijai. Neparametrinės suderinamumo hipotezės vienamačiu atveju, kaip palyginimo priemonė, nagrinėjamas Bahaduro asimptotinis santykinis efektyvumas (Bahadur, 1960; Nikitin, 1995). Kartu su teoriniais rezultatais pasiūlytų N-metrikos tipo testų galingumas ištirtas, naudojant Monte-Karlo metodą. Be paprastos ir sudėtinės suderinamumo hipotezių yra analizuojami homogeniškumo testai vienamačiu ir daugiamačiu atvejais. Ištirtas platus alternatyvių hipotezių diapazonas.

Notations

$(\mathfrak{X}, \mathfrak{U})$	measurable space;
$\mathbb{N}$	the set of natural numbers;
$\mathbb{R}^p$	the set of p-dimensional real vectors;
$\mathbb{C}$	the set of complex numbers;
$\boldsymbol{x}$	a real vector $(x_1, \ldots, x_p)$ in $\mathbb{R}^p$;
$[0,1]^p$	the unit square in $\mathbb{R}^p$;
S^{p-1}	the hypersphere in $\mathbb{R}^p$;
$\|\cdot\|$	the Euclidean norm in $\mathbb{R}^p$;
$\langle \cdot, \cdot \rangle$	the scalar product in $\mathbb{R}^p$;
X, Y, Z	random vectors in $\mathbb{R}^p$;
$X_1, ..., X_n$	the sample of independent observations of X;
$F_n(x)$	the empirical distribution function based on the sample $X_1, ..., X_n$;
$\mathbf{E}X$	the mean of X;
$varX$	the variance of X;
$cov(X, Y)$	the covariance between X and Y;
$\xrightarrow{d}$	weak convergence;

$\xrightarrow{P}$	convergence in probability;
$x \wedge y$	$\min(x, y)$ for real numbers x and y;
$x \vee y$	$\max(x, y)$ for real numbers x and y;
$C([0,1]^p)$	the set of continuous functions $x : [0,1]^p \to \mathbb{R}$.

Turinys

Introduction

Scientific problem

In this thesis the problem of verification of classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence is investigated.

Actuality

In the classical statistical analysis of observations in various studies researchers usually begin their investigations by proposing a distribution for their observations. There are several reasons for that:

- The distribution of the sample data may throw a light on the process that generate the data, if a suggested model for the process is correct, the sample data follow a specific distribution.
- Parameters of the distribution may be connected with important parameters in describing the basic model.
- Knowledge of the distribution of the data allows for application of standard statistical testing and estimation procedures.

Sometimes such assumptions about the form of the distribution are made by analyzing the procedure by which the data was obtained or made arbitrarily, often from considerations of convenience in the statistical methods used. In any case there arises a need to check whether the chosen distribution is true.

The researcher may be interested in the question whether the distribution of observed data has a given fixed form (a simple hypothesis) or belongs to a certain family of distributions (composite hypothesis). In case of multivariate observations, in addition to goodness of fit problems, there arises the problem of testing the hypothesis of the independence of the components of the random vector being observed without knowing the precise form of the marginal distributions. Another class of problems is that of comparing two or several samples among themselves. These are the so-called homogeneity tests, designed for testing the hypothesis that the samples obtained are identically distributed.

To solve these problems a large number of goodness of fit, homogeneity and independence procedures have appeared over the years, the choice of which is made depending on the structure of the observations, the hypothesis being tested, the efficiency of the test, etc. Choosing the most efficient test of several ones that are available to the researcher is regarded as one of the basic problems of statistics. However, it is well known that for a variety of problems arising in statistical theory and practice the uniformly most powerful tests are unknown. Therefore creation of new test procedures sensitive to a particular type of hypotheses remains actual and in our days.

Klebanov in (Klebanov, 2005; Zinger et al., 1989) introduced a new class of probability metrics - N-distances, which has many useful properties and therefore could be applied to obtaining new powerful and simply computable statistical tests. The construction of such criteria together with investigation of their properties become a topical problem after Klebanov's works.

Research object

This thesis is devoted to statistical criteria based on N-distances for testing classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence.

Aim and tasks

The main objectives of the thesis are connected with application of N-distance theory, introduced by Klebanov (Klebanov, 2005; Zinger et al., 1989), to testing classical statistical hypotheses of goodness of fit, homogeneity, independence and symmetry. In particularly, we focus on the following tasks:

- Construction of statistics based on N-metrics for testing mentioned hypotheses.
- Establishing the critical region of proposed criteria, obtaining the asymptotic distribution of test statistics under the null and alternative hypotheses.
- Comparison of proposed N-distance tests with some classical criteria using Asymptotic Relative Efficiency (ARE) by Bahadur (Bahadur, 1960; Nikitin, 1995).

In parallel to the theoretical results the empirical comparison of the power of proposed N-distance tests is investigated.

Research methods

Methods of mathematical statistics, general probability theory and stochastic processes are applied. The proofs of the limit behavior of proposed test statistics are based on the theory of U-statistics (Koroljuk and Borovskich, 1994; Lee, 1990) and the properties of the weak convergence of stochastic processes (Bulinskii and Shiryaev, 2005). All the results presented in empirical part of the thesis are produced by the means of Monte Carlo simulations done with the help of R statistical package.

Scientific novelty

Novelty of the results is closely related to the formulated aims and problems. Proposed methods extend, generalize and supplement the results of Klebanov in (Klebanov, 2005), Baringhaus and Franz in (Baringhaus and Franz, 2004) and Szekely and Rizzo in (Szekely and Rizzo, 2005). In particular, proposed criteria and established asymptotic distributions of test statistics in the problems of goodness of fit, uniformity on the hypersphere, independence (in bivariate case) and symmetry (in univariate case) have not been earlier considered in statistical literature.

Practical value of the results

Proposed statistical criteria could be applied to the real data analysis problems connected with verification of the hypotheses about the considered sample of observations.

Defended propositions

- Propositions on the asymptotic distribution of goodness of fit (simple and composite hypotheses) test statistics based on N-distances under the null and alternative hypotheses.
- Construction and asymptotic behavior of the test statistic in the problem of uniformity on the hypersphere S^{p-1}.
- Propositions on the asymptotic distribution of two-sample test statistics based on N-distances under the null and alternative hypotheses; application of bootstrap and permutation procedures to determination of critical region of proposed tests; construction and asymptotic behavior of distribution-free two-sample test.
- Construction and asymptotic null distribution of tests statistic based on N-distances for criteria of symmetry about zero in univariate case and independence in bivariate case.
- Propositions on the computational form of tests statistics based on N-metrics with different strongly negative definite kernels.
- Comparison of proposed N-distance and classical nonparametric goodness of fit tests in univariate case by means of asymptotic relative efficiency by Bahadur.

History of the problem and main results

Goodness of fit tests

In the course of his Mathematical contributions to the theory of evolution, Karl Pearson abandoned the assumption that biological populations are normally distributed. The need to test the fit arose naturally in this context, and in 1900 Pearson invented his chi-squared test. This statistics and others related to it remain

among the most used statistical procedures.

In this section we propose a brief review of the best-known types of goodness-of-fit tests applied to the following hypotheses:

- the distribution of a random variable under observation coincides with a given, completely known distribution G (simple hypothesis),
- the distribution of a random variable under observation belongs to a given parametric family of distributions $\mathbf{\Lambda} = \{G(x, \theta),\ x \in \mathbb{R}^p,\ \theta \in \Theta \subset \mathbb{R}^d\}$ (composite hypothesis).

It is assumed that a sample of independent observations $X_1, ..., X_n$ of random variable X with unknown distribution function $F(x)$ is available for the researcher.

We first consider the case of the simple hypothesis and univariate samples. In case of continuous distributions the most popular tests used to verify the stated hypotheses are based on the empirical distribution functions

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(X_i \leq x\right).$$

These tests are based on finding a measure of the difference between the empirical and theoretical distributions. Goodness of fit tests based on N-distances, introduced in this thesis, also belongs to this class of tests.

The most famous and well-studied statistics of this type is obviously Kolmogorov statistic (Kolmogorov, 1933)

$$D_n = \sqrt{n} \sup_x |G(x) - F_n(x)|$$

and its variations, the one sided statistics of Smirnov (Smirnov, 1944)

$$D_n^+ = \sqrt{n} \sup_x [F_n(x) - G(x)],$$

$$D_n^- = \sqrt{n} \sup_x [G(x) - F_n(x)],$$

as well as Kuiper statistic (Kuiper, 1960)

$$V_n = \sqrt{n} \sup_x [F_n(x) - G(x)] - \sqrt{n} \inf_x [F_n(x) - G(x)].$$

Watson (Watson, 1976) and Darling (Darling, 1983a;b) have introduced the centered versions of Kolmogorov-Smirnov statistic:

$$W_n = \sqrt{n} \sup_x |F_n(x) - G(x) - \int_{-\infty}^{+\infty} (F_n(x) - G(x)) dG(x)|.$$

Another group of statistics is based on the integral distance between G and F_n. The best known among them is Cramer–von Mises statistic (Darling, 1957; Martynov, 1978)

$$\omega_{n,1}^2 = n \int_{-\infty}^{+\infty} (G(x) - F_n(x))^2 dG(x).$$

Anderson and Darling (Anderson and Darling, 1952; 1954) proposed to improve properties of presented statistics by introducing a non-negative weight function $q(x)$

$$D_n = \sqrt{n} \sup_x q(G(x)) |G(x) - F_n(x)|,$$

$$\omega_{n,1}^2 = n \int_{-\infty}^{+\infty} q(G(x)) (G(x) - F_n(x))^2 dG(x).$$

The weight functions are used in these statistics 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.

Another generalization is connected with the consideration of arbitrary positive integer powers k of the empirical process $F_n - G$, i.e.

$$\omega_{n,q}^k = n^{\frac{k}{2}} \int_{-\infty}^{+\infty} q(G(x)) (G(x) - F_n(x))^k dG(x).$$

Obviously, $\omega_{n,q}^k$ are not consistent against all alternatives for odd k, however, for one-sided alternatives these statistics may turn out to be serious competitors to classical criteria.

The most popular weighted integral statistic is Anderson-Darling statistic (Anderson and Darling, 1954)

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{(G(x) - F_n(x))^2}{G(x)(1 - G(x))} dG(x).$$

Another type of statistics, based on the martingale part of the empirical pro-

cess, have been studied by Khmaladze (Khmaladze, 1981) and Aki (Aki, 1986)

$$K_n = \sqrt{n} \sup_x \left| F_n(x) - \int_{-\infty}^{x} \frac{1 - F_n(y)}{1 - G(y)} dG(y) \right|,$$

$$L_n^2 = n \int_{-\infty}^{+\infty} \left| F_n(x) - \int_{-\infty}^{x} \frac{1 - F_n(y)}{1 - G(y)} dG(y) \right|^2 dG(x).$$

All represented statistics can be brought into standard form with the help of transformation $t = G(x)$. When the null hypothesis is valid, the quantities $t_i = G(X_i)$ will have a uniform distribution on $[0, 1]$. After that, for instance, Kolmogorov and Cramer-von Mises statistics will have the form

$$D_n = \sqrt{n} \sup_{0 \le t \le 1} |F_n(t) - t|,$$

$$\omega_{n,1}^2 = n \int_0^1 (F_n(t) - t)^2 dt,$$

where $F_n(t)$ is the empirical distribution function constructed from the quantities t_i, $i = 1, 2, ..., n$.

This transformation $t = G(x)$ shows, that under the null hypothesis the distributions of presented statistics are independent of the form of the hypothesized distribution function G. This property plays an important role and become one of the reasons of popularity of the tests in practice. The statistics D_n^+ and D_n^- have the same limit distribution as $n \to \infty$,

$$P(D_n^{\pm} \le x) \to 1 - e^{-2x^2}, \;\; x > 0.$$

Under the null hypothesis the limit distribution of Kolmogorov statistic D_n has the form

$$P(D_n \le x) \to 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 x^2}.$$

Concerning integral type tests, the asymptotic distribution of Cramer-von Mises $\omega_{n,1}^2$ and Anderson-Darling A_n^2 statistics coincides with the distribution of infinite quadratic form

$$\omega^2 = \sum_{i=1}^{\infty} \frac{\xi_i^2}{(\pi i)^2},$$

$$A^2 = \sum_{i=1}^{\infty} \frac{\xi_i^2}{i(i+1)},$$

where ξ_i, $i = 1, 2, ...$, are independent random variables from the standard normal distribution.

We now consider a generalization of the Kohnogorov-Smirnov and Cramer-von Mises statistics that is applicable to testing the composite parametric hypothesis that the distribution function $F(x)$ of the random variable under observation belongs to the family $\mathbf{\Lambda}$ of distribution functions described at the beginning of this section. Starting form classical work by M. Kac, J. Kiefer, and J. Wolfowitz (Kac et al., 1955), this problem is well known and have generated plenty of attention from researchers both in theoretical and applied statistical literature (Durbin, 1973; Khmaladze, 1977; Lilliefors, 1967; Martynov, 1978; Tyurin, 1970; 1984). All statistics use a preliminary computable estimate $\hat{\theta_n}$ of the unknown parameter θ. Such an estimate can be taken to be, for exmnple, the maximum likelihood estimate. The statistics under consideration are then have the form

$$D_n = \sqrt{n} \sup_x q(G(x, \hat{\theta_n}))|G(x, \hat{\theta_n}) - F_n(x)|,$$

$$\omega_{n,1}^2 = n \int_{-\infty}^{+\infty} q(G(x, \hat{\theta_n}))(G(x, \hat{\theta_n}) - F_n(x))^2 dG(x, \hat{\theta_n}).$$

In contrast to simple hypothesis, considered statistics under the null hypothesis are not distribution-free since their asymptotic distributions depend on G (Durbin, 1973). Worse, they are not even asymptotically parameter-free since these distributions depends in general on the value of unknown parameter θ. However, there exists an important class of composite hypotheses under which the limit distributions of statistics above are independent of the unknown values of the parameters. These include cases when $\mathbf{\Lambda}$ is a location-scale family of distribution functions, that is

$$\mathbf{\Lambda} = \{G(\frac{x - \theta_1}{\theta_2}),\ \theta_1 \in \mathbb{R},\ \theta_2 > 0\}.$$

In the most general case of family Λ the parametric dependence problem of the distribution of statistics in practice can be overcome by utilization of parametric bootstrap methods (Stute et al., 1993; Szucs, 2008).

Another group of tests for parametric composite hypotheses is based on the size of the deviation of order statistics from their mathematical expectations: Shapiro-Wilk, Shapiro-Francia tests (Shapiro and Wilk, 1965; Shapiro and Francia, 1972)

and on the comparison of the sample moments with the theoretical moments, for example, the asymmetry and excess tests for the hypothesis of normality (DAgostino, 1971; DAgostino et al., 1990; Mardia, 1970). This last type of tests is not strictly speaking a type of goodness-of-fit test, since these tests are not consistent against all possible alternatives and are applied only to specific types of families Λ, like Gaussian family. Therefore here we omit the details of these tests and refer the reader to a number of papers proposing numerous tests for normality, as the most widespread problem among composite goodness of fit tests, based on different approaches (Best and Rayner, 1985; Bowman and Shenton, 1975; Locke and Spurrier, 1976; Park, 1999; Prescott, 1976; Spiegelhalter, 1977; Zhang, 1999).

Multivariate goodness of fit tests

Since Pearson criteria, goodness of fit tests have been developed mostly for univariate distributions and, except for the case of multivariate normality (Csorgo, 1986; Epps and Pulley, 1983; Henze, 1994; Koziol, 1983; L.Baringhaus and H.Henze, 1992; 1998; Mardia, 1970; Pettitt, 1979; Szekely and Rizzo, 2005; Zhu et al., 1995), very few references can be found in the literature about multivariate tests of fit (DAgostino and Stephens, 1986). The main difficulty here is that many tests statistics based on the empirical distribution function of the sample have the limit distribution dependent on the data's underlying distribution in a nontrivial way. Thus, to calculate asymptotic significance points of these statistics may be difficult. In principle, the chi-square test can be applied for testing the goodness of fit for arbitrary multivariate distribution. However this procedure also has some weak points, as it is unknown what is the best way to choose the corresponding cell limits.

To extend the two most important classes of univariate goodness of fit tests, the Kolmogorov-Smirnov and Cramer-von Mises statistics, to multivariate case Rosenblatt in (Rosenblatt, 1952b) proposed a transformation of an absolutely continuous p-variate distribution into the uniform distribution on the p-dimensional cube. The main point of suggested transformation is presented in the next theorem

Theorem 1. *Let $X = (X_1, ..., X_p)$ be a random vector with joint density*

$$f(x_1, ..., x_p) = f_1(x_1) f_2(x_2|x_1) ... f_p(x_p|x_1, ..., x_{p-1}).$$

Then vector $Y = (Y_1, ..., Y_p)$ is uniformly distributed on the p-dimensional cube, where

$$Y_1 = F(X_1),$$
$$Y_i = F(X_i|X_1, ..., X_{i-1}), \quad i = 2, .., p.$$

After Rosenblatt's transformation Kolmogorov-Smirnov and omega-square statistics for testing the uniformity on p-dimensional cube will have the form

$$D_n = \sqrt{n} \sup_x |F_n^*(x) - x_1 \cdots x_p|,$$

$$\omega_{n,1}^2 = n \int_{-\infty}^{+\infty} (F_n^*(x) - x_1 \cdots x_p)^2 dx_1 \cdots dx_p,$$

where $F_n^*(x)$ is the empirical distribution function constructed from the transformed sample Y. However, this approach also have some disadvantages. The main one is lack of the uniqueness, mentioned statistics are not invariant because a relabelling of the components of p-dimensional vector would give a different Rosenblatt transformation and, therefore, a different values of statistics. In more detail the asymptotic behavior of represented statistics were studied in (Durbin, 1970; Justel et al., 1997; Krivyakova et al., 1977; Martynov, 1978).

One more approach to multivariate goodness of fit problem, also based on comparisons with to uniform distribution was introduced in (Bickel and Breiman, 1983) . It was proved that the variables

$$U_i = \exp\left[-n \int_{\|x-X_i\|<R_i} f(x)dx\right], \quad i = 1, ..., n,$$

where $f(x)$ is the hypothesized density function, $X_1, ..., X_n$ are n points sampled independently from $f(x)$ and R_i is the distance from X_i to its nearest neighbor, have a univariate distribution that does not depend on $f(x)$ and is approximately uniform. However the computations involved in integrating even a very simple density over p-dimensional spheres are usually hardly feasible.

In practice for arbitrary dimension (usually $p > 3$) it can be rather difficult from the computation point of view to establish the percentiles of the limit distribution of tests statistics. In this case the bootstrap methods could be applied to obtain the critical region of the tests, for example, (Burke, 2000).

Homogeneity tests

One of the classical problems of the theory of nonparametric inference is testing whether two samples come from the same or different populations. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two independent samples with unknown continuous distribution functions $F(x)$ and $G(x)$. A formal statement of the two-sample problem is to test the hypothesis about the equality of these functions: $H_0 : F \equiv G$.

Denote by F_n and G_m the empirical distribution functions based on the initial

samples. For the problem of testing H_0 there exist many statistics based on the difference between F_n and G_m extended from goodness of fit statistics considered above. In univariate case the most prominent of them are analogs of Kolmogorov-Smirnov and Cramer-von Mises type tests proposed in (Lehmann, 1951; Maag and Stephens, 1968; Pettitt, 1976; 1979; Smirnov, 1939; Wald and Wolfowitz, 1940)

$$D_{n,m} = \sqrt{\frac{mn}{m+n}} \sup_x |F_n(x) - G_m(x)|,$$

$$\omega^2_{n,m} = \frac{mn}{m+n} \int_{-\infty}^{+\infty} (F_n(x) - G_m(x))^2 dH_{n,m},$$

where $H_{n,m}$ is an empirical distribution function of the pooled sample.

The limit distribution of the statistics $D_{n,m}$ and $\omega^2_{n,m}$ coincides with their analogs among goodness of fit statistics.

A natural and very popular competitor of the mentioned statistics is the class of linear rank statistics introduced in (Hajek and Sidak, 1967). Denote $N = n+m$, then a simple linear rank statistics has the form

$$S_N = N^{-1} \sum_{i=1}^{m} a_N(R_i/(N+1)),$$

where $R_i, i = 1, .., m$ is the rank of an observation Y_i in the ordered pooled sample and function $a_N(x)$ is constant on all the intervals of the form $[(i-1)/N, i/N)$, $i = 1, ..., N$. It is assumed, moreover, that when $N \to \infty$

$$a_N(x) \xrightarrow{L^2} J(x),$$

where $J(x)$ is a nonconstant function on $[0, 1]$. In case $J(x) = \sqrt{12}(x - 1/2)$ we obtain the famous Wilcoxon rank statistics.

A common feature of all above mentioned procedures is that they only use the information provided by the ranks of observations within the sorted list of pooled sample. Consequently, the respective test statistics are distribution free under H_0.

Classical approaches to the two-sample problem in the univariate case based on comparing empirical distribution functions do not have a natural distribution-free extension to the multivariate case. The situation here is very similar to analogous goodness of fit problem with the limit distribution of test statistics to be dependent on unknown distribution of initial samples. This fact was one of the reasons of applying bootstrap and permutation techniques to verification of homogeneity hypotheses, for example, (Burke, 2000; van der Vaart and Wellner, 1996).

Bikel in (Bickel, 1969), by applying Fisher's permutation principle, constructed a consistent distribution free multivariate extension of the univariate Kolmogorov-Smirnov test by conditioning on the pooled sample. Friedman and Rafsky in (Friedman and Rafsky, 1979) proposed a two-sample test based on the minimal spanning tree of the sample points as a multivariate generalization of the univariate Wald-Wolfowitz runs test.

Another class of consistent, asymptotically distribution free tests is based on the nearest neighbors in Euclidean distance metric (Bickel and Breiman, 1983; Henze, 1988) . Let Z be the pooled sample of length $N = n + m$ and define the function $I_i(r)$ as $I_i(r) = 1$, if Z_i and $N_r(Z_i)$ belongs to the same sample and $I_i(r) = 0$ otherwise, where $N_r(Z_i)$ is the rth nearest neighbor to Z_i. Then the hypothesis should be rejected in case of large values of the statistics:

$$T_{N,k} = \sum_{i=1}^{N} \sum_{r=1}^{k} I_i(r).$$

Tests of uniformity on S^{p-1}

Let $X_1, ..., X_n$ be the sample of independent observations of random variable X with continuous distribution function $F(x)$, where $X_i = (x_{i1}, .., x_{ip})$ and $\|X_i\| = 1$, $i = 1, ..., n$.

First, we present some wide spread universal tests for uniformity on S^{p-1} for arbitrary p like Rayleigh (Figueiredo and Gomes, 2003), Ajne (Ajne, 1968; Beran, 1968) and Gine (Gine, 1975) tests. At the end of this subsection we also provide a short review of some popular criteria for testing the uniformity in the special case of $p = 2$ (circle).

Rayleigh test

Let R be the length of the resultant vector defined by

$$R = \| \sum_{i=1}^{n} X_i \|.$$

Denote $\bar{R} = \frac{R}{n}$ the mean resultant length. When X has uniform distribution on S^{p-1}, $\mathbf{E}X = 0$, then it is intuitive to reject the hypothesis of uniformity when the sample vector mean $\frac{1}{n} \sum_{i=1}^{n} X_i$ is far from 0, i.e. for large values of R. It is usual to take the statistic $pn\bar{R}^2$. Under uniformity, the asymptotic distribution of $pn\bar{R}^2$ is $\chi^2(p)$.

Gine test

The Gine statistic is defined by

$$G = \frac{n}{2} - \frac{p-1}{2n}\left(\frac{\Gamma(p-1)/2}{\Gamma(p/2)}\right)^2 \sum_{i<j} \sin\psi_{ij},$$

where ψ_{ij} is the smaller of two angles between X_i and X_j.

We reject uniformity for large values of G. Under the null hypothesis, G is asymptotically distributed as an infinite linear combination of independent χ^2 variables.

Ajne test

The Ajne statistic is defined by

$$A = \frac{n}{4} - \frac{1}{\pi n}\sum_{i<j} \psi_{ij},$$

where ψ_{ij} has the same meaning as in Gine test.

We reject uniformity for large values of A. Under uniformity, A is asymptotically distributed as an infinite linear combination of χ^2 variables.

Some uniformity tests for S^1

We parameterize S^1 with the variable x ranging from 0 to 1, where the zero point and the direction around the circle have been chosen arbitrary.

A modification of Kolmogorov-Smirnov test for S^1 was presented by Kuiper in (Kuiper, 1960)

$$V_n = \sup_{x\in[0,1)} \{F_n(x) - F(x)\} - \inf_{x\in[0,1)} \{F_n(x) - F(x)\},$$

where $F(x)$ and $F_n(x)$ are cumulative and empirical distribution functions respectively.

Watson adapted Cramer-von Mises test for S^1 (Watson, 1961; 1967)

$$U_n^2 = n\int_0^1 \left[F_n(x) - F(x) - \int_0^1 [F_n(y) - F(y)]dF(y)\right]^2 dF(x).$$

The tests proposed by Watson and Kuiper are all distribution-free, consistent against all alternatives and rotation invariant.

Ajne studied a rotation-invariant test for uniformity on S^1 based on semicircles (Ajne, 1968), using the statistics

$$A_n = \frac{1}{n}\int_0^1 (N(x) - \frac{n}{2})^2 dx, \tag{1}$$

where $N(x)$ is the number of data points falling in the semicircle interval $[x, x+\frac{1}{2})$. Ajne also considered the rotation-invariant statistics, which can be considered as an adaptation to S^1 of Kolmogorov-Smirnov statistics

$$N = \sup_{x\in[0,1)} N(x).$$

The main disadvantage of Ajne tests that they are not consistent against all alternatives. Rothman in (Rothman, 1972) developed a more general form of (1) using arcs of arbitrary length, having the property of consistency.

$$A_n^{H(t)} = \int_0^1 \int_0^1 (N(t,x) - nt)^2 dx dH(t),$$

where $H(t)$ is a distribution function and $N(t,x)$ is the number of observations in the arc $(x, x+t]$.

N-distance tests

In this thesis we consider the problems of verification of classical statistical hypotheses of goodness of fit, homogeneity, symmetry and independence. Proposed tests statistics are based on a class of probability metrics N-distances, introduced in (Klebanov, 2005; Zinger et al., 1989). If $(\mathfrak{X}, \mathfrak{U})$ is a measurable space, then the structure of N-distance between two probability measures μ and ν on it has the form

$$\begin{aligned} N(\mu,\nu) \quad := \quad & 2\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\mu(x)d\nu(y) - \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\mu(x)d\mu(y) - \\ & - \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x,y)d\nu(x)d\nu(y), \end{aligned} \tag{2}$$

where $L(x,y)$ is a strongly negative definite kernel (see section 1.1).

Earlier results

Some applications of N-distance theory to testing of statistical hypotheses were first proposed by Klebanov in (Klebanov, 2005). Klebanov gave a construction of multivariate free- of-distribution two-sample test based on the division of initial sample into three equal parts.

The idea is as follows. Suppose that X and Y are two independent random vectors in $\mathbb{R}^p$, and define one dimensional independent random variables U and V by the relation:

$$U = L(X, Y) - L(X, X'),$$
$$V = L(Y', Y'') - L(X'', Y''),$$

where $X =^d X' =^d X''$ and all vectors X, X', X'', Y, Y', Y'' are mutually independent.
Under conditions

$$\mathbf{E}L(X, X') < \infty, \quad \mathbf{E}L(Y, Y') < \infty,$$

Klebanov showed that

$$X =^d Y \Leftrightarrow U =^d V.$$

Thus, instead of testing the homogeneity of multivariate random vectors we can test the homogeneity of one-dimensional variables using a vide range of free of distribution tests. This method leads to essential loss of information, however allows testing the homogeneity hypothesis in high-dimensional cases.

In case of strongly negative definite kernel $L(x, y) = \|x - y\|$, where $\|\cdot\|$ is a Euclidean norm in $\mathbb{R}^p$, the statistics of the form (2) were studied by Szekely and Rizzo in (Szekely and Rizzo, 2005) and Baringhaus and Franz in (Baringhaus and Franz, 2004), where they were applied for testing the hypothesis of multivariate normality and homogeneity respectively. The critical values of Szekely and Rizzo test are obtained by means of Monte Carlo simulations. The asymptotic null distribution of the test statistic in (Baringhaus and Franz, 2004) is derived using the projection method and shown to be the limit of the bootstrap distribution.

In case of $L(x, y) = 1 - \exp(-\|x - y\|^2)$ N-distance statistics are very similar to well-known BHEP tests $B_{n,p}(\beta)$ of multivariate normality defined in the following way (Epps and Pulley, 1983; Henze and Wagner, 1997; Henze and Zirkler, 1990; L.Baringhaus and H.Henze, 1998). Assume that the sample covariance matrix $\hat{\Sigma}$ is nonsingular and $Y_i = \hat{\Sigma}^{-1/2}(X_i - \bar{X})$, $i = 1, ..., n$, is the standardized

sample. The statistic $B_{n,p}(\beta)$ is the weighted integral of the squared difference between the multivariate normal characteristic function and the empirical characteristic function $\Psi_n(t) = \frac{1}{n}\sum_{k=1}^{n} e^{it^T Y_k}$. The test statistic is defined

$$B_{n,p}(\beta) = \int_{R^p} \|\Psi_n(t) - e^{-\frac{\|t\|^2}{2}}\|^2 \varphi_\beta(t) dt,$$

where $\varphi_\beta(t)$ is a weighting function. When the weighting function is

$$\varphi_\beta(t) = (2\pi\beta^2)^{-p/2} e^{-\frac{\|t\|^2}{2\beta^2}}$$

and $\beta = \sqrt{2}$ one can see that $B_{n,p}(\beta)$ absolutely coincides with N-distance statistics with $L(x,y) = 1 - \exp(-\|x-y\|^2)$ (proposition 8).

Main results of the thesis

A natural way for testing the null hypotheses mentioned above is to consider N-metrics between the empirical and hypothesized distributions is case of goodness of fit criterion or between two empirical distributions in case of homogeneity, symmetry and independence tests. The corresponding test statistics proposed in this thesis have the form:

- Goodness of fit test (simple hypothesis)

  $$T_n = -n \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G(x))\, d(F_n(y) - G(y)),$$

  where $F_n(x)$ is the empirical distribution function of initial sample and $G(x)$ is the hypothesized distribution function (section 2.1).

  As a special case of goodness of fit tests we consider tests of uniformity on the hypersphere S^{p-1} (section 3.2) with test statistics of the form

  $$T_n = n\left[\frac{2}{n}\sum_{i=1}^{n} \mathbf{E}L(X_i, Y) - \frac{1}{n^2}\sum_{i,j=1}^{n} L(X_i, X_j) - \mathbf{E}L(Y, Y')\right],$$

  where Y, Y' are independent random variables from the uniform distribution on S^{p-1}.

- Goodness of fit test (composite hypothesis)

$$T_n = -n \int_{R^p} \int_{R^p} L(x,y)\, d(F_n(x) - G(x,\hat{\theta}_n))\, d(F_n(y) - G(y,\hat{\theta}_n)),$$

where $\hat{\theta}_n$ is the estimate of unknown parameter θ under the assumption that X has the distribution from family $\mathbf{\Lambda} = \{G(x,\theta),\ x \in \mathbb{R}^p,\ \theta \in \Theta \subset \mathbb{R}^d\}$ (see section 2.2).

- Two-sample test

$$T_{n,m} = -\frac{nm}{n+m} \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G_m(x))\, d(F_n(y) - G_m(y)),$$

where $F_n(x)$, $G_m(x)$ are empirical distribution functions based on the given samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ (section 3.1).

- Symmetry test (about zero in univariate case)

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x,y)\, d\Delta_n(x)\, d\Delta_n(y),$$

where $\Delta_n(x) = F_n(x) + F_n(-x) - 1$ and $F_n(x)$ is empirical distribution function, constructed from the sample $X_1, \ldots, X_n$ (section 3.3).

- Independence test (bivariate case)

$$T_n = -n \int_{R^4} L(x,y) d\Delta_n(x)\, d\Delta_n(y),$$

where $x, y \in \mathbb{R}^2$, $\Delta_n(x) = F_n(x) - F_{n1}(x_1)F_{n2}(x_2)$,
$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_{i1} < x_1)I(X_{i2} < x_2)$ is a bivariate empirical distribution function and $F_{ni}(x_i)$, $i = 1, 2$ are univariate empirical distribution functions, based on the i-th coordinate of the sample (section 3.3).

We should reject the null hypothesis in case of large values of test statistics. The consistency of proposed tests against all fixed alternatives is ensured by the property of N-distances stated in Theorem 3.

The asymptotic distribution of tests statistics under the null hypothesis is established in Theorems:

- Goodness of fit test (simple hypothesis) – Theorems 7, 9, 10

- Tests of uniformity on S^{p-1} – Theorems 17, 18, 20, 21
- Goodness of fit test (composite hypothesis) - Theorems 11
- Homogeneity test – Theorems 12, 14, 15, 16
- Symmetry test – Theorems 22
- Independence test – Theorems 23

In all the cases it coincides with the distribution of a certain infinite quadratic form

$$\sum_{i=1}^{\infty} \lambda_i \zeta_i^2,$$

where ζ_i, $i = 1, 2, ...$ are independent random variables with standard norm distribution.

Under the alternative hypothesis the normality of some goodness of fit and homogeneity tests statistics is established in Theorems 8, 13.

The results of the power comparison study proposed at the end on this thesis show that N-metrics tests are powerful competitors to existing classical criteria, in the sense that they are consistent against all alternatives and have relatively good power against general alternatives compared with other tests. The possibility in the selection of the kernel for N-distance allows to create the test more sensitive to particular type of alternatives.

Approbation of the thesis

The result of this thesis were presented at the Conferences of Lithuanian Mathematical Society (2008, 2009), The 8th Tartu Conference on Multivariate statistics (Tartu, Estonia, June 26–29, 2007) and 22nd Nordic Conference on Mathematical Statistics (NORDSTAT) (Vilnius, Lithuania, June 16–19, 2008).

Moreover, the results of the thesis were presented at the seminar on Probability Theory and Mathematical Statistics of Institute of Mathematics and Informatics and at the seminar "Nonparametric statistics and time series" of Department of Mathematics and Mechanics of Moscow State University (Moscow, Russia, April, 2007)

Principal publications

The main results of the thesis are published in the following papers:

1. Bakshaev, A. 2008. Nonparametric tests based on N-distances, *Lithuanian Mathematical Journal* 48(4): 368–379. ISSN 0363-1672 (ISI Master Journal List).
2. Bakshaev, A. 2009. Goodness of fit and homogeneity tests on the basis of N-distances, *Journal of Statistical Planning and Inference* 139 (11): 3750–3758. ISSN 0378-3758 (ISI Master Journal List).
3. Bakshaev, A. 2010. N-distance tests for composite hypothesis of goodness of fit, *Lithuanian Mathematical Journal* 50(1): 14–34. ISSN 0363-1672 (ISI Master Journal List).
4. Bakshaev, A. 2010. N-distance tests for uniformity on the hypersphere, accepted for publication in *Nonlinear Analysis, Modelling and Control*. ISSN 1392-5113.

Structure of the thesis

The thesis consists of an introduction, four chapters and the bibliography.

- Chapter 1 provides a review of auxiliary results for further research. The first part is devoted to some general aspects and notions of N-distance theory, used as a basis for construction of proposed in this thesis statistical tests. In the second part we shortly describe methods of calculation of the distribution functions of quadratic forms of normal variables that are needed to compute the limit distribution functions of proposed test statistics.
- Chapter 2 is dedicated to goodness of fit tests. Both simple and composite hypothesis are considered. Particular attention is devoted to establishing the asymptotic distribution of test statistics, based on N-metrics.
- Chapter 3 provides an application of N-distances for testing the hypotheses of homogeneity, uniformity on the hypersphere, independence and symmetry about zero.
- Chapter 4 is devoted to a comparison of proposed N-distance tests with some classical criteria. In the first part as a measure for comparison of criteria Asymptotic Relative Efficiency (ARE) by Bahadur (Bahadur, 1960;

Nikitin, 1995) is considered. In the second part a comparative Monte Carlo power study is proposed. Besides simple and composite hypothesis of goodness of fit, we consider two-sample tests in uni- and multivariate cases.

Acknowledgement

I would like to express profound gratitude to my supervisors Prof. Rimantas Rudzkis and Prof. Yurij Tyurin for their invaluable support, encouragement and useful suggestions throughout this research work. Their support and continuous guidance helped me overcome my doubts and enabled me to complete my work successfully. I am also highly thankful to Prof. Lev Klebanov and Prof. Marijus Radavicius for their valuable consultations and help during my study years.

Further, I would like to thank the staff of the department of Probability Theory and Statistics of Institute of Mathematics and Informatics, Dr. Juozas Machys and my colleagues Dr. Vaidotas Zemlys and Dr. Vaidotas Balys for their help and useful recommendations preparing the manuscript of this thesis.

Finally, I am heartily grateful to my wife Tatjana, parents and my friends for their constant attention and help.

1

Auxiliary results

1.1. N-distances

The statistical tests proposed in this thesis are based on a new class of probability metrics N-distances, introduced by Klebanov (Klebanov, 2005; Zinger et al., 1989). Besides verification of statistical hypotheses, Klebanov suggested some additional applications of these metrics to the problems of recovering measures from a potential, finding new characterizations of probability distributions and investigating their stability, derivation of a new estimation methods. In this section a short review of N-metrics theory will be given.

Let $(\mathfrak{X}, \mathfrak{U})$ be a measurable space and $\mathfrak{B}$ the set of all probability measures μ on it. Suppose that L is real continuous function, and denote by $\mathfrak{B}_L$ the set of all probability measures $\mu \in \mathfrak{B}$ on $(\mathfrak{X}, \mathfrak{U})$ under condition of existence of the integral

$$\int_{\mathfrak{X}} \int_{\mathfrak{X}} L(x, y)\, d\mu(x)\, d\mu(y) < \infty. \tag{1.1}$$

We shall say, that function L is a negative definite kernel if for arbitrary $n \in \mathbb{N}$, any arbitrary points $\forall x_1, \ldots, x_n$ and any complex numbers $c_1, \ldots, c_n$ under condition

$\sum_{i=1}^{n} c_i = 0$ the following inequality holds

$$\sum_{i=1}^{n}\sum_{j=1}^{n} L(x_i, x_j) c_i \bar{c_j} \leq 0.$$

Negative definite kernel L is strictly negative definite if the equality above is true $\forall x_1, \ldots, x_n$ only for $c_1 = \ldots = c_n = 0$.

In the next theorem the structure of N-distances in the set of probability measures $\mathfrak{B}_L$ is introduced.

Theorem 2. *Let L be real continuous function on $\mathfrak{X}^2$ under condition $L(x, y) = L(y, x)\ \forall x, y \in \mathfrak{X}$. The inequality*

$$\begin{aligned} N(\mu, \nu) \quad := \quad & 2\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x, y) d\mu(x) d\nu(y) - \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x, y) d\mu(x) d\mu(y) - \\ & - \int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x, y) d\nu(x) d\nu(y) \geq 0 \end{aligned} \tag{1.2}$$

holds for all $\forall \mu, \nu \in \mathfrak{B}_L$ if and only if L is negative definite kernel.

One of the main notions in the theory of N-distances is that of strong negative definiteness. This additional condition for the kernel $L(x, y)$ will help us to avoid $N(\mu, \nu) = 0$ in case $\mu \neq \nu$ and, as a result, obtain consistent tests against all fixed alternatives.

Let Q be a measure on $(\mathfrak{X}, \mathfrak{U})$, and $h(x)$ be a function integrable with respect to Q and such that $\int_{\mathfrak{X}} h(x)\, dQ(x) = 0$. We shall say that L is strongly negative definite kernel if L is negative definite and equality

$$\int_{\mathfrak{X}}\int_{\mathfrak{X}} L(x, y) h(x) h(y)\, dQ(x)\, dQ(y) = 0$$

implies that $h(x) = 0$ Q-almost everywhere for any measure Q.

Theorem 3. *Let L be a real continuous function on $\mathfrak{X}^2$ satisfying all the conditions of the Theorem 2. The inequality (1.2)*

$$N(\mu, \nu) \geq 0$$

holds for all measures $\mu, \nu \in \mathfrak{B}_L$ with equality in the case $\mu = \nu$ only, if and only if L is a strongly negative definite kernel.

Let us give some examples of strongly negative definite kernels which will be widely used in the following constructions of statistical tests.
Univariate case:

1. $L(x,y) = |x-y|^r$, where $0 < r < 2$. For $r = 2$ $L(x,y)$ is a negative definite kernel, but not strongly negative definite.
2. $L(x,y) = \frac{|x-y|}{1+|x-y|}$.
3. Let $L(x) = \int_{-\infty}^{+\infty}(1 - I(x > a))\, d\sigma(a)$, then $L(x \wedge y)$ is strongly negative definite kernel if and only if $\sigma(x)$ is a suitable strictly monotone distribution function.
4. Let $U(z) = \int_0^{\infty}(1-\cos 2x)\frac{1+x^2}{x^2}d\theta(x)$, where $\theta(x)$ is is real non-decreasing function, $\theta(-0) = 0$. Then $L(x,y) = U(x-y)$ is a strongly negative definite kernel if $supp\theta = [0,\infty)$.

Multivariate case:

1. $L(x,y) = \|x-y\|^r$, where $0 < r < 2$.
2. $L(x,y) = 1 - \exp(-\|x-y\|^2)$.

Some more examples of strongly negative definite kernels can be found in section 4 or in (Klebanov, 2005).

The definition of N-distances as a probability metric comes from the following theorem.

Theorem 4. *(L.B. Klebanov) Let L be a strongly negative definite kernel on $\mathfrak{X}^2$ under condition $L(x,y) = L(y,x)$ and $L(x,x) = 0\ \forall x, y \in \mathfrak{X}$. Then $N(\mu,\nu)^{\frac{1}{2}}$ is a distance on $\mathfrak{B}_L$.*

Another possible expression of $N(\mu,\nu)$ can be given in terms of random variables. Let X and Y be two independent random variables with cumulative distribution functions μ and ν correspondingly. Denote by X' and Y'- independent copies of X and Y. Now we can rewrite $N(\mu,\nu)$ in the form

$$N(\mu,\nu) = 2\mathbf{E}L(X,Y) - \mathbf{E}L(X,X') - \mathbf{E}L(Y,Y'). \tag{1.3}$$

Let us give an example and express the N-distance based on the strongly negative definite kernel $L(x,y) = \|x-y\|^r$, ($x, y \in \mathbb{R}^p$ and $0 < r < 2$) in terms of characteristic functions of μ and ν.

Denote by $f(x)$ and $g(x)$ the characteristic functions of random variables X and Y with probability distributions μ and ν respectively. Using a well-known

Zolotarev's formula

$$\mathbf{E}|X|^r = c_r \int_0^\infty (1 - Re(f(t)))t^{-1-r}dt$$

in univariate case Klebanov got the result

$$N(\mu, \nu) = c_r \int_0^\infty |f(x) - g(x)|^2 x^{-1-r}dx,$$

where c_r - is a constant depending only on r.

In p-variate case the required expression of N-distance can be derived from univariate case using the formula

$$\|x\|^r = \int_{S^{p-1}} |\langle x, \tau\rangle|^r d\rho(\tau),$$

where $\rho(\tau)$ is a measure on the unit sphere S^{p-1} in $\mathbb{R}^p$.

Then

$$\begin{aligned}
&N(\mu, \nu) = \int_{\mathbb{R}^{2p}} \|x - y\|^r d\Delta(x)\, d\Delta(y) = \\
&= \int_{S^{p-1}} d\rho(\tau) \left[\int_{\mathbb{R}^{2p}} |\langle x - y, \tau\rangle|^r d\Delta(x)\, d\Delta(y) \right] = \\
&= c_r \int_{S^{p-1}} d\rho(\tau) \int_{\mathbb{R}^1} \frac{du}{|u|^{r+1}} \int_{\mathbb{R}^{2p}} \left[1 - e^{i\langle x-y,\tau\rangle u}\right] d\Delta(x)\, d\Delta(y) = \\
&= c_r \int_{\mathbb{R}^1} \frac{du}{|u|^{r+1}} \int_{S^{p-1}} |\delta(u\tau)|^2 d\rho(\tau),
\end{aligned}$$

where $\Delta(x) = \mu(x) - \nu(x)$, $\delta(t)$ is the difference between characteristic functions of X and Y, $u\tau = (u\tau_1, ..., u\tau_p)$ and constant c_r depends only on r.

1.2. The distribution functions of quadratic forms

In this thesis the asymptotic distributions of proposed N-distance statistics under the null hypothesis are established in terms of distribution of a certain infinite quadratic form

$$Q = \sum_{i=1}^{\infty} \alpha_i \zeta_i^2, \tag{1.4}$$

where $0 < \alpha_1 \leq \alpha_2 \leq ...$, and ζ_i, $i = 1, 2, ...$ are independent random variables from the standard normal distribution. In this section a short review of methods for computing the distribution function of quadratic forms (1.4) is provided.

In general case let us consider the quadratic form

$$Q^* = X^T A X, \tag{1.5}$$

where A is a symmetric matrix and X is a random vector from the Gaussian distribution with zero mean vector and unit correlation matrix. As A is a real symmetric matrix, then there exists a real orthogonal matrix C, such that $D = C^T A C$ is a diagonal matrix. In this case the distribution function of Q^* coincides with the distribution function of quadratic form

$$\sum_{i=1}^{N} \alpha_i \zeta_i^2, \tag{1.6}$$

where ζ_i, $i = 1, 2, ...$, are independent random variables from the standard normal distribution and α_i are the eigenvalues of matrix A.

Let us further combine all the members with equal coefficients in (1.4) and consider the quadratic form

$$Q_N = \sum_{i=1}^{N} \alpha_i \chi_{s_i}^2, \tag{1.7}$$

where $N \in \mathbb{N}$ and can be infinite, $\alpha_i > 0$, $\forall i = 1, 2, ...$; $\alpha_i \neq \alpha_j$, if $i \neq j$; $s_i \geq 0$ and $\chi_{s_i}^2$ are independent random variables from the χ^2 distribution with s_i degrees of freedom. The characteristic function of Q_N has the form

$$\phi_N(t) = \prod_{k=1}^{N} (1 - 2it\alpha_k)^{-s_k/2}. \tag{1.8}$$

The methods of computing the distribution functions of considered quadratic forms are usually based on various methods of inverting the characteristic functions (1.8).

First results were obtained by Smirnov (1937) and then generalized by Martynov, Sukhatme and Imhoff (Imhof, 1961; Martynov, 1975; Sukhatme, 1972).

First consider the case when all α_i in (1.4) are different. The formula of inversion of characteristic function (1.8) has the form

$$F_N(x) - F_N(0) = \frac{1}{2\pi i} \int_{-\infty i}^{\infty i} \frac{(1 - e^{-ux/2})}{u\sqrt{D_N(2iu)}} du, \tag{1.9}$$

where $D_N(u) = \prod_{k=1}^{N}(1 - \alpha_k u)$. If N is uneven, let us add a zero member to (1.4), so that $\alpha_{N+1} = 0$. After integration Smirnov received the following formula for distribution function of Q (1.4)

$$F_N(x) = 1 + \frac{1}{pi} \sum_{k=1}^{N/2} (-1)^k \int_{1/\alpha_{2k-1}}^{1/\alpha_{2k}} \frac{e^{-ux/2}}{\sqrt{-D_N(u)}} \frac{du}{u}, x \geq 0. \tag{1.10}$$

The Smirnov's formula (1.10) can be generalized to the case $N = \infty$.

Ibragimov and Rozanov proved that infinite quadratic form (1.4) converges with probability 1 if and only if

$$\sum_{i=1}^{\infty} |\alpha_i| < \infty. \tag{1.11}$$

However, even when the condition (1.11) on α_k is satisfied the series in the right part of (1.10) can be divergent.

Let us formulate some additional conditions on α_k, $k = 1, 2, ...$, to ensure the convergence of series in (1.10). Denote by

$$T_{k1} = \frac{\alpha_{2k}\alpha_{2k+1}}{\alpha_{2k} - \alpha_{2k+1}} ln \left(\frac{\frac{\alpha_{2k-1} - \alpha_{2k+2}}{\alpha_{2k-1}\alpha_{2k+2}}}{\frac{\alpha_{2k} - \alpha_{2k+1}}{\alpha_{2k+1}\alpha_{2k}}} \frac{\alpha_{2k+1}^2}{\alpha_{2k}^2} \right),$$

$$T_{k2} = \frac{\frac{\alpha_{2k-1} - \alpha_{2k+2}}{\alpha_{2k-1}\alpha_{2k+2}}}{\frac{\alpha_{2k} - \alpha_{2k+1}}{\alpha_{2k+1}\alpha_{2k}}} \sum_{m=2k+3}^{\infty} \frac{\alpha_{2k+2} - \alpha_m}{\alpha_{2k+2}\alpha_m},$$

$$T_k = T_{k1} + T_{k2}.$$

Theorem 5. *If the condition (1.11) is fulfilled and the following series*

$$\sum_{k=1}^{\infty} (\frac{1}{\alpha_{2k+1}} - \frac{1}{\alpha_{2k}})$$

is divergent, then the series (1.10) converges when $x > \overline{\sup}_k T_k$.

The next theorem establishes the conditions then (1.10) convergence $\forall x \geq 0$.

Theorem 6. *If the condition 1.11 is fulfilled and the members of the series* $\delta_k = \frac{1}{\alpha_{k+1}} - \frac{1}{\alpha_k}$, $k = 1, 2, ...$ *are growing, starting from a certain* K *and*

$$\lim_k \frac{\delta_{k+1}}{\delta_k} < \infty,$$

then $\overline{\sup}_k T_k = 0$.

In case $\frac{1}{\alpha_k} = (ck^\beta) + o(k^{\beta-2})$, where $1 < \beta < \infty$ and constant $c > 0$, the conditions of Theorem (6) are satisfied and $\overline{\sup}_k T_k = 0$.

In the most general case the distribution function of quadratic form (1.7) was derived by Martynov (Martynov, 1975) and has the form

$$F_N(x) = 1 + \frac{1}{\pi} \arctan \frac{1}{a} - \frac{1}{\pi} \int_0^\infty \frac{e^{-atx/2} \sin(\theta(t,x))}{\varrho(t)} dt, \;\; x \geq 0,$$

where

$$\theta(t,x) = \sum_{k=1}^N \frac{s_k}{2} \omega_k(t) - \frac{tx}{2},$$

$$\varrho(t) = t \prod_{k=1}^N \left((1 - a\alpha_k t)^2 + (\alpha_k t)^2\right)^{s_k/4},$$

$$\omega_k(t) = \arctan \frac{1/\alpha_k - at}{t},$$

where a is a arbitrary parameter. This formula generalizes a known formula of Imhof, which can be obtained from the given formula with $a = 0$. The choice of the value of $a \neq 0$ leads to the appearance of a factor $\exp(-atx/2)$ in the numerator of the integrand, guaranteeing a more rapid decrease in the amplitude of the oscillations of the integrand. This makes the process of numerical integration easier.

2

Goodness of fit test

2.1. Simple hypothesis

Let $X_1, \ldots, X_n$, $X_i = (X_{i,1}, ..., X_{i,p})$ be a p-dimensional sample of independent observations of random variables X with unknown continuous probability distribution functions $F(x)$, $x \in \mathbb{R}^p$.

The nonparametric null hypothesis in the problem of testing goodness of fit is $H_0 : F(x) = G(x)$, where $G(x)$, $x \in \mathbb{R}^p$, is a known continuous cumulative distribution function.

A natural way for testing the null hypothesis is to consider a certain metrics $\rho(\cdot, \cdot)$ between the empirical and hypothesized distributions. In this thesis we propose to use N-distance defined in (1.2) with test statistic of the form

$$T_n = nN_L(\mu_{F_n}, \nu_G), \tag{2.1}$$

where L is a strongly negative definite kernel, μ_{F_n} is the empirical distribution based on the sample $X_1, \ldots, X_n$ and ν_G is the probability distribution with distribution function $G(x)$.

We should reject the null hypothesis in case of large values of test statistic, that

is if $T_n > c_\alpha$, where c_α can be found from the equation

$$P_0(T_n > c_\alpha) = \alpha,$$

here P_0 is probability distribution corresponding to the null hypothesis and α – size of the test.

In case of p-variate sample test statistic T_n will have the form

$$T_n = -n \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G(x))\, d(F_n(y) - G(y)), \tag{2.2}$$

or in terms of mathematical expectations of random variables

$$T_n = n \left[\frac{2}{n} \sum_{i=1}^{n} \mathbf{E} L(X_i, X) - \frac{1}{n^2} \sum_{i,j=1}^{n} L(X_i, X_j) - \mathbf{E} L(X, Y) \right], \tag{2.3}$$

where X, Y are independent random variables with cumulative distribution function $G(x)$, $x \in \mathbb{R}^p$, and $F_n(x)$ is the empirical distribution based on the sample $X_1, \ldots, X_n$.

2.1.1. Asymptotic distribution of test statistic

To determine the critical region of our test $\{T_n > c_\alpha\}$ let us consider the asymptotic distribution of T_n (2.2).
Denote

$$H(x,y) := \mathbf{E} L(x, X) + \mathbf{E} L(X, y) - L(x, y) - \mathbf{E} L(X, X^{'}),$$

where $L(x, y)$ is the strongly negative definite kernel of N-distance (2.1) and X, $X^{'}$ are independent random variables with cumulative distribution function $G(x)$.

Theorem 7. *If $G(x)$ satisfies the condition $\mathbf{E} H^2(X, X^{'}) < \infty$, then under the null hypothesis the asymptotic distribution of T_n coincides with the distribution of infinite quadratic form*

$$Q = \mathbf{E}[L(X, X^{'}) - L(X, X)] + \sum_{j=1}^{\infty} \lambda_j (\zeta_j^2 - 1), \tag{2.4}$$

where ζ_j, $j = 1, 2, ...$, are independent random variables from the standard normal distribution and λ_j are the eigenvalues of integral operator

$$Af(y) = \mathbf{E}H(X, y)f(X), \tag{2.5}$$

Proof. Let us rewrite T_n in the form of V-statistic (Koroljuk and Borovskich, 1994; Lee, 1990) with symmetric kernel $H(x, y)$

$$T_n = \frac{1}{n}\sum_{i=i}^{n}\sum_{j=1}^{n} H(X_i, X_j). \tag{2.6}$$

The proof of the theorem is based on the following lemma establishing the limit distribution of von Mises' functionals of type (2.6) with degenerate kernels (see Theorem 4.3.2 in (Koroljuk and Borovskich, 1994)).

Lemma 1. *Consider a von Mises' functional*

$$V_n = \frac{1}{n^2}\sum_{i,j=1}^{n} W(X_i, X_j),$$

with a symmetric kernel $W(x, y)$ under conditions:

- $\mathbf{E}W(x, X) = 0$ *a.s. (property of degeneracy),*
- $\mathbf{E}W(X, X^{'}) = 0$*,*
- $\mathbf{E}|W(X, X)| < \infty$*,*
- $\mathbf{E}W(X, X^{'})^2 < \infty$*.*

Then

$$nV_n \xrightarrow{d} \mathbf{E}W(X, X) + \sum_{j=1}^{\infty}\lambda_j(\zeta_j^2 - 1),$$

where ζ_j, $j = 1, 2, ...$, are independent random variables from the standard normal distribution and λ_j are the eigenvalues of integral operator

$$Af(y) = \mathbf{E}W(X, y)f(X),$$

In case $\sum_{j=1}^{\infty}|\lambda_j| < \infty$ then

$$\mathbf{E}W(X, X) = \sum_{j=1}^{\infty}\lambda_j.$$

The first three conditions on the kernel of V-statistic (2.6) in lemma follows directly from the definition and properties of N-distances. Denote by X, Y and Z

the independent random variables with probability function $G(x)$, then

- $\mathbf{E}H(x,Z) = \mathbf{E}L(x,X) + \mathbf{E}L(X,Z) - \mathbf{E}L(x,Z) - \mathbf{E}L(X,Y) = 0$ (property of degeneracy),
- $\mathbf{E}H(Y,Z) = N_L(\mu_Y, \nu_Z)) = 0$,

where μ_Y and ν_Z are the probability distributions of random variables Y and Z correspondingly.
Note that if $N_L(\cdot,\cdot)$ is a distance in the space of probability measures under condition (1.1), then

$$H(a,a) = \mathbf{E}L(a,X) + \mathbf{E}L(X,a) - L(a,a) - \mathbf{E}L(X,Y) \geq 0, \quad \forall a \in \mathbb{R}^p.$$

Thus, taking into account (1.1)

- $\mathbf{E}|H(Z,Z)| = \mathbf{E}H(Z,Z) = \mathbf{E}L(X,Z) - \mathbf{E}L(Z,Z) < \infty$.

The forth condition in lemma coincides with the only condition of theorem and this ends the proof of the theorem.

Let us further consider the asymptotic distribution of test statistic T_n (2.2) under alternative hypothesis with fixed alternatives. In this case the probability to reject the null hypothesis with a given size of the test α tends to 1 when $n \to \infty$. Therefore we consider our statistic T_n normalized in a special way.

Let $F(x)$ does not equal identically to $G(x)$ and consider the statistic

$$T_n^* = \frac{T_n}{\sqrt{n}} - a, \tag{2.7}$$

where

$$a = N_L(\mu_F, \nu_G) = \mathbf{E}H(Y,Y^{'}) = 2\mathbf{E}L(X,Y) - \mathbf{E}L(X,X^{'}) - \mathbf{E}L(Y,Y^{'}),$$

where X, $X^{'}$ and Y, $Y^{'}$ are independent random variables with probability distributions X,$X^{'} \sim \mu_F$ and Y,$Y^{'} \sim \nu_G$ and corresponding cumulative distribution functions $F(x)$ and $G(x)$.

Theorem 8. *Denote $H^*(x) := [\mathbf{E}H(x,Y) - a]^2$, if $\mathbf{E}H^*(Y^{'}) > 0$ and $\mathbf{E}H(Y,Y^{'}) < \infty$ then T_n^* asymptotically has the normal distribution with zero mean and variance σ^2:*

$$\sigma^2 = \frac{2}{n(n-1)}[2(n-2)C_1 + C_2], \tag{2.8}$$

where $C_1 = \mathbf{E}H^*(Y^{'})$ and $C_2 = \mathbf{E}[H(Y,Y^{'}) - a]^2$.

Proof. First note that under the alternative hypothesis the kernel of V-statistic (2.6) is nondegenerate. After that the statement of the theorem follows immediately from the Theorem 4.2.5 in (Koroljuk and Borovskich, 1994) establishing the asymptotic normality of von Mises' functionals in this case.

In practice it is usually rather difficult to establish the limit null distribution of test statistic T_n in the form (2.4). The main problem here is connected with calculation of eigenvalues of integral operator (2.5). In sections 2.1.2, 2.1.3 we tried to solve this problem and establish the analytical formulas for the coefficients of quadratic form in some special cases of strongly negative definite kernels $L(x, y)$. First, the proposed methods will be considered in univariate case and then generalized to arbitrary dimension, in particular, the two-dimensional case will be examined.

Remark. An alternative method to the determination of the critical region $\{T_n > c_\alpha\}$ of proposed statistics can be as follows. In case of simple hypothesis when $G(x)$ is a know distribution function the percentiles of the finite sample null distribution of T_n can be established with the help of Monte Carlo simulations:

1. Generate repeatedly i.i.d. random samples $X_1, \ldots, X_n$ from the distribution $G(x)$, $x \in \mathbb{R}^p$.
2. For each sample evaluate T_n using the formula (2.3).
3. Calculate the empirical distribution function $F_n^*(x)$ on the basis of computed T_n values.
4. For a chosen significance level $\alpha > 0$ find c_α from the equation:

$$c_\alpha = \inf\{c_\alpha : F_n^*(c_\alpha) \geq 1 - \alpha\}.$$

2.1.2. Univariate case

The asymptotic distribution of test statistic T_n (2.2) established in (2.4) depends on the initial distribution $G(x)$. To avoid this we propose a special form of strongly negative definite kernel in (2.1) and discuss another approach in determination of the coefficients of quadratic form (2.4).

Consider the statistic T_n in the form

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(G(x), G(y))\, d(F_n(x) - G(x))\, d(F_n(y) - G(y)), \quad (2.9)$$

where $F_n(x)$ is the empirical distribution function based on the sample $X_1, \ldots, X_n$. Note that if $L(x, y)$ is a strongly negative definite kernel, then $L(G(x), G(y))$ also satisfies the conditions of strongly negative definiteness. One can see, that in this case

$$T_n = -n \int_0^1 \int_0^1 L(x, y)\, d(F_n^*(x) - x)\, d(F_n^*(y) - y), \tag{2.10}$$

where $F_n^*(x)$ is the empirical distribution function based on transformed sample $t_1, \ldots, t_n$, $t_i = G(X_i)$, $i = 1, 2, ..., n$. Under the null hypothesis the new sample will have the uniform distribution on $[0, 1]$ and the following theorem will help us to establish the limit distribution of (2.10).

Theorem 9. *Under the null hypothesis statistic T_n will have the same asymptotic distribution as quadratic form*

$$Q = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{a_{kj}}{\pi^2 kj} \zeta_k \zeta_j, \tag{2.11}$$

where ζ_k are independent random variables from the standard normal distribution and

$$a_{kj} = -2 \int_0^1 \int_0^1 L(x, y)\, d\sin(\pi k x)\, d\sin(\pi j y).$$

Proof. Let us consider the random process $Z_n(x) = \sqrt{n}(F_n^* - x)$, where F_n^* is the empirical distribution function based on the sample $t_1, \ldots, t_n$. The natural way to prove the statement of the theorem is to find the weak limit of $Z_n(x)$ and then use the theorem of continuity for the sequence of random processes.

The next lemma implies the continuity of the functional (2.10).

Lemma 2. *Assume that $L(x, y)$ is a strongly negative definite kernel with bounded variation on $[0, 1]^2$ (Towghi, 2002) and $L(x, 0)$, $L(x, 1)$ also have bounded variation on $[0, 1]$ then*

$$B(f(x)) = \int_0^1 \int_0^1 L(x, y)\, df(x)\, df(y) \tag{2.12}$$

is a continuous functional in the space of continuous functions on $[0, 1]$ with uniform metrics.

Proof of lemma 2. The integrals (2.12) for all $f(x) \in C[0, 1]$ are considered after

formal integration by parts with the help of the formula

$$\int_0^1\int_0^1 L(x,y)dF(x,y) = \int_0^1\int_0^1 F(x,y)\,dL(x,y) -$$
$$-\int_0^1 F(1,y)\,dL(1,y) - \int_0^1 F(x,1)\,dL(x,1) +$$
$$+\int_0^1 F(x,0)\,dL(x,0) + \int_0^1 F(0,y)\,dL(0,y) +$$
$$+F(0,0)L(0,0) - F(1,0)L(1,0) - F(0,1)L(0,1) + F(1,1)L(1,1).$$

The continuity and bounded variation of $L(x,y)$, $L(x,0)$ and $L(x,1)$ guarantee us the existence of all the integrals in the right part of equality above in the meaning of Stieltjes integrals (Towghi, 2002).

Assume that $f_n(x)$ are continuous functions on $[0,1]$, such that

$$\sup_{0\le x\le 1}|f_n(x)| \to 0.$$

After integration by parts, taking into account that $L(x,y) = L(y,x)$, $B(f(x))$ (2.12) will have the form

$$B(f(x)) = \int_0^1\int_0^1 f(x)f(y)\,dL(x,y) -$$
$$-2f(1)\int_0^1 f(y)\,dL(1,y) + 2f(0)\int_0^1 f(x)\,dL(x,0) -$$
$$-2f(0)f(1)L(0,1) + f^2(0))L(0,0) + f^2(1)L(1,1).$$

Bounded variation of $L(x,y)$ on $[0,1]^2$ and $L(0,x)$, $L(1,x)$ on $[0,1]$ together with condition that $\sup_{0\le x\le 1}|f_n(x)| \to 0$ implies that $B(f_n(x)) \to 0$, when $n \to \infty$.

Theorem 13.1 in (Billingsley, 1968) establishes the weak limit of $Z_n(z)$ in $C[0,1]$, when $n \to \infty$.

$$Z_n(x) \xrightarrow{d} W_0(x),$$

where $W_0(x)$ is the Brownian bridge. Thus, we have that the asymptotic distribution of T_n will coincide with the distribution of random variable

$$T = -\int_0^1\int_0^1 L(x,y)\,dW_0(x)\,dW_0(y). \tag{2.13}$$

Remark. The trajectories of the empirical random process $Z_n(x)$ are not continuous, we could circumvent the discontinuity problems by adopting a different definition of empirical distribution function. Let $H_n(x)$ be, as a function of x ranging over $[0, 1]$, the distribution function corresponding to the uniform distribution of mass $(n + 1)^{-1}$ over each of the $n + 1$ intervals $[X_{(i+1)}, X_(i)]$, where $X_0 = 0$, $X_{n+1} = 1$ and $X_{(1)}, ..., X_{(n)}$ are the values $X_1, ..., X_n$ ranged in increasing order. The functions $F_n^*(x)$ and $H_n(x)$ are close and the following inequality holds

$$\sup_{0<x<1} |F_n^*(x) - H_n(x)| \leq \frac{1}{n}. \tag{2.14}$$

Now let $Z_n^*(x)$ be an element of $C[0, 1]$ with value

$$Z_n^*(x) = \sqrt{n}(H_n(x) - x).$$

Since each $Z_n^*(t)$ is a random variable, Z_n^* is a random element of $C[0, 1]$. By (2.14), we have

$$\sup_{0<x<1} |Z_n(x) - Z_n^*(x)| \leq \frac{1}{\sqrt{n}}$$

and the asymptotic distribution of $B(Z_n(x))$ will be the same as the asymptotic distribution of $B(Z_n^*(x))$ for any continuous functional B.

Random process $W_0(x)$ with zero mean and correlation function $K(x, y) = \min(x, y) - xy$ with probability 1 can be presented in the form

$$W_0(x) = \sum_{k=1}^{\infty} \zeta_k \phi_k(x), \tag{2.15}$$

where ζ_k are independent random variables from the Gaussian distribution with mean zero and variance λ_k, where $\phi_k(x)$ and λ_k are eigenfunctions and eigenvalues of the linear symmetric integral operator A with the kernel $K(x, y)$

$$A(f(x)) = \int_0^1 K(x, y) f(y)\, dx.$$

In case of $K(x, y) = \min(x, y) - xy$, it is easy to derive, that

$$\lambda_k = (\pi k)^{-2},$$

$$\phi_k(x) = \sqrt{2} \sin(\pi k x).$$

Finally we obtain the representation (2.11) by substituting the expression (2.15)

into (2.13) and complete the proof.

Proposition 3. *Statistics T_n (2.10) with different kernels can be calculated using the formulas:*

- $L(x,y) = \max(x,y)$,

$$T_n = 2\sum_{i=1}^{n} \frac{1+t_i^2}{2} - \frac{1}{n}\sum_{i,j=1}^{n} \max(t_i,t_j) - \frac{2n}{3}.$$

- $L(x,y) = 1 - e^{-(x-y)^2}$,

$$T_n = 2\sum_{i=1}^{n}(1 - \sqrt{\pi}(\Phi(\sqrt{2}(1-t_i))) - \Phi(-\sqrt{2}t_i)) -$$
$$-\frac{1}{n}\sum_{i,j=1}^{n}(1 - e^{-(t_i-t_j)^2}) - Cn,$$

  where $C = 1 - \sqrt{\pi}\left(\int_0^1 \Phi(\sqrt{2}x)dx - \int_{-1}^0 \Phi(\sqrt{2}x)dx\right)$ *and* $\Phi(x)$ *is a distribution function of standard normal distribution.*

- $L(x,y) = \frac{|x-y|}{1+|x-y|}$,

$$T_n = 2\sum_{i=1}^{n}(1 - \ln(1+t_i) - \ln(2-t_i)) -$$
$$-\frac{1}{n}\sum_{i,j=1}^{n}\frac{|t_i-t_j|}{1+|t_i-t_j|} - (3 - 4\ln 2)n.$$

- $L(x,y) = |x-y|^\alpha$, $0 < \alpha < 2$,

$$T_n = \frac{2}{\alpha+1}\sum_{i=1}^{n}(t_i^{\alpha+1} + (1-t_i)^{\alpha+1}) -$$
$$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}|t_i-t_j|^\alpha - \frac{2n}{(\alpha+1)(\alpha+2)}.$$

Proof. All formulas are obtained from representation (2.10) of statistics T_n by calculating corresponding integrals. Let $F_n^*(x)$ be an empirical distribution function based on transformed sample $t_1, \ldots, t_n$, $t_i = G(X_i)$, $i = 1, 2, ..., n$, then

- $L(x,y) = \max(x,y)$,

$$\int_0^1 \int_0^1 \max(x,y) dx dF_n^*(y) = \frac{1}{n} \sum_{i=1}^n \int_0^1 \max(x,t_i) dx = \frac{1}{n} \sum_{i=1}^n \frac{1+t_i^2}{2},$$

$$\int_0^1 \int_0^1 \max(x,y) dx dy = \int_0^1 \frac{1+x^2}{2} dx = \frac{2}{3}.$$

- $L(x,y) = 1 - e^{-(x-y)^2}$,

$$\int_0^1 \int_0^1 (1 - e^{-(x-y)^2}) dx dF_n^*(y) = \frac{1}{n} \sum_{i=1}^n \int_0^1 (1 - e^{-(x-t_i)^2}) dx =$$
$$= \frac{1}{n} \sum_{i=1}^n (1 - \sqrt{\pi}(\Phi(\sqrt{2}(1-t_i)) - \Phi(\sqrt{2}t_i))),$$

$$\int_0^1 \int_0^1 (1 - e^{-(x-y)^2}) dx dy = 1 - \sqrt{\pi}(\int_0^1 \Phi(\sqrt{2}x) dx - \int_{-1}^0 \Phi(\sqrt{2}x) dx),$$

where $\Phi(x)$ is a standard normal distribution function.

- $L(x,y) = \frac{|x-y|}{1+|x-y|}$,

$$\int_0^1 \int_0^1 \frac{|x-y|}{1+|x-y|} dx dF_n^*(y) = \frac{1}{n} \sum_{i=1}^n \int_0^1 \frac{|x-t_i|}{1+|x-t_i|} dx =$$
$$= \frac{1}{n} \sum_{i=1}^n \left(\int_0^{t_i} \frac{x}{1+x} dx + \int_0^{1-t_i} \frac{x}{1+x} dx \right) =$$
$$= \frac{1}{n} \sum_{i=1}^n (1 - \ln(1+t_i) - \ln(2-t_i)),$$

$$\int_0^1 \int_0^1 \frac{|x-y|}{1+|x-y|} dx dy = \int_0^1 (1 - \ln(1+x) - \ln(2-x)) dx = 3 - 4\ln 2.$$

- $L(x,y) = |x-y|^\alpha$, $0 < \alpha < 2$,

$$\int_0^1 \int_0^1 |x-y|^\alpha dx dF_n^*(y) = \frac{1}{n}\sum_{i=1}^n \int_0^1 |x-t_i|^\alpha dx =$$
$$\frac{1}{n}\sum_{i=1}^n \left(\int_0^{t_i} x^\alpha dx + \int_0^{1-t_i} x^\alpha dx \right) = \frac{1}{n}\sum_{i=1}^n \frac{t_i^{\alpha+1} + (1-t_i)^{\alpha+1}}{\alpha+1},$$

$$\int_0^1 \int_0^1 |x-y|^\alpha dx dy = \int_0^1 \frac{x^{\alpha+1} + (1-x)^{\alpha+1}}{\alpha+1} dx = \frac{2}{(\alpha+1)(\alpha+2)}.$$

Let us note, that in case $L(x,y) = |x-y|$ and $L(x,y) = x \vee y$ we obtain very similar statistics. After formal integration by parts of

$$T_n = -n \int_0^1 \int_0^1 L(x,y) d(F_n^*(x) - x) d(F_n^*(y) - y)$$

test statistic T_n can be written as

- $L(x,y) = |x-y|$,
$$T_n = 2n \int_0^1 (F^*(x) - x)^2 dx;$$
- $L(x,y) = x \vee y$,
$$T_n = n \int_0^1 (F^*(x) - x)^2 dx,$$

and coincides with the well-know Cramer-von Mises statistic, where $F_n^*(x)$ is the empirical distribution function of $t_1, \ldots, t_n$, $t_i = G(X_i)$.

The asymptotic distribution of T_n with $L(x,y) = |x-y|$ is the same as the distribution of quadratic form

$$Q = 2\sum_{k=1}^\infty (\pi k)^{-2} \zeta_k^2,$$

where ζ_k are independent random variables from the standard normal distribution.

This follows directly from (2.11), where

$$a_{kj} = -2 \int_0^1 \int_0^1 |x-y| \, d\sin(\pi k x) \, d\sin(\pi j y) = 2\delta_{kj}.$$

Applying Smirnov formula (1.10) we derive the probability distribution function of Q

$$F_Q(x) = 1 + \frac{1}{\pi} \sum_{k=1}^{\infty} (-1)^k \int_{((2k-1)\pi)^2}^{(2k\pi)^2} \frac{\exp -\frac{xu}{4}}{\left(-\frac{\sin\sqrt{u}}{\sqrt{u}}\right)^{\frac{1}{2}}} \frac{du}{u}, \quad x \geq 0.$$

Remark. As it was shown in (2.9), (2.10), taking a special form of strongly negative definite kernel $L(G(x), G(y))$ in (2.1) allows us to obtain a distribution – free test. However, in general case the asymptotic null distribution of statistic T_n also can be established. In case $G(x)$ is a strictly monotone distribution function, the limit distribution of T_n will coincide with the distribution of quadratic form (2.11), where the coefficients are expressed by the formula

$$a_{kj} = -2 \int_0^1 \int_0^1 L(G^{-1}(x), G^{-1}(y)) \, d\sin(\pi k x) \, d\sin(\pi j y).$$

2.1.3. Multivariate case

In multivariate case to avoid the dependence of the distribution of T_n (2.2) on $G(x)$, first transform our sample $X_1, \ldots, X_n$, where $X_i = (X_{i,1}, ..., X_{i,p})$, $i = 1, .., n$, to the sample $t_1, \ldots, t_n$, where $t_i = (t_{i,1}, ..., t_{i,p})$, $i = 1, .., n$, using Rosenblatt transformation

$$t_{i,1} = G(x_{i,1}),$$

$$t_{i,2} = G(x_{i,2}|x_{i,1}),$$

$$\ldots\ldots\ldots\ldots$$

$$t_{i,p} = G(x_{i,p}|x_{i,1}, ..., x_{i,p-1}).$$

Under the null hypothesis the transformed sample will have the uniform distribution on the unit hypercube $C^p = [0,1]^p$ and statistic T_n for testing the uniformity

on C^p has the form

$$T_n = -n \int_{[0,1]^{2p}} L(x,y)\, d(F_n(x) - x_1 \cdots x_p)\, d(F_n(y) - y_1 \cdots y_p), \quad (2.16)$$

where $F_n(x)$, $x = (x_1, ..., x_p)$ is a p-dimensional empirical distribution function based on sample $t_1, \ldots, t_n$

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \mathbf{1}\left(t_{ij} \le x_j\right).$$

The asymptotic distribution of T_n (2.16) is established by the next theorem.

Theorem 10. *Under the null hypothesis the limit distribution of T_n will coincide with the distribution of quadratic form*

$$Q = \sum_{i,j=1}^{\infty} a_{ij} \sqrt{\alpha_i \alpha_j}\, \zeta_i \zeta_j, \quad (2.17)$$

where ζ_i are independent random variables from the standard normal distribution,

$$a_{ij} = -\int_{[0,1]^{2p}} L(x,y)\, d\psi_i(x)\, d\psi_j(y), \quad x, y \in \mathbb{R}^p, \quad (2.18)$$

α_i and $\psi_i(x)$ are eigenvalues and eigenfunctions of the integral operator A

$$Af(x) = \int_{[0,1]^p} K(x,y) f(y) dy \quad (2.19)$$

with the kernel

$$K(x,y) = \prod_{i=1}^{p} \min(x_i, y_i) - \prod_{i=1}^{p} x_i y_i,$$

where $x = (x_1, ..., x_p)$, $y = (y_1, ..., y_p)$.

Proof. The outline of the proof including the statement of Lemma 2 is absolutely the same as for one-dimensional case.

Consider the random process

$$Z_n(x) = \sqrt{n}(F_n(x) - x_1 \cdots x_p), \quad x = (x_1, ..., x_p).$$

The properties of this empirical process were in detail studied by Rosenblatt (Rosenblatt, 1952a), Durbin (Durbin, 1970), Krivyakova, Martynov, Turin (Krivyakova et al., 1977).

$Z_n(x)$, $x \in \mathbb{R}^p$, weakly converges to the Gaussian random field $W_0(x)$, $x \in \mathbb{R}^p$, (a certain analogue of p-dimensional Brownian bridge) with zero mean vector and correlation function

$$K(x,y) = \prod_{i=1}^{p} \min(x_i, y_i) - \prod_{i=1}^{p} x_i y_i. \tag{2.20}$$

Remark. In accordance with univariate case we can circumvent the discontinuity problems of the trajectories of $Z_n(x)$ by smoothing the empirical distribution function $F_n(x)$. One of the possible variants is based on kernel estimation of the cumulative distribution functions in detail discussed in (Liu and Yang, 2008; Yamato, 1973)

$$H_n(x) = \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} K_h(X_i - u)du, \quad \forall x \in \mathbb{R}^p,$$

where $h = (h_1, ..., h_p)$ are positive numbers depending on the sample size n, called bandwidths.

The functions $F_n(x)$ and $H_n(x)$ are close and the following inequality holds

$$\sup_{x \in C^p} |F_n(x) - H_n(x)| \leq \frac{1}{n}. \tag{2.21}$$

Now denote by $Z_n^*(x)$ an element of $C[0,1]^p$

$$Z_n^*(x) := \sqrt{n}(H_n(x) - x).$$

By (2.21), we have

$$\sup_{0<x<1} |Z_n(x) - Z_n^*(x)| \leq \frac{1}{\sqrt{n}}$$

and the asymptotic distribution of $B(Z_n(x))$ will be the same as the asymptotic distribution of $B(Z_n^*(x))$ for any continuous functional B.

Thus, the asymptotic distribution of T_n (2.16) will be the same as the distribution of random variable

$$T = -\int_{[0,1]^{2p}} L(x,y)\, dW_0(x)\, W_0(y), \quad x, y \in \mathbb{R}^p. \tag{2.22}$$

With probability 1 random process $W_0(x)$ with zero mean vector and correlation function (2.20) can be presented in the form

$$W_0(x) = \sum_{k=1}^{\infty} \zeta_k \phi_k(x), \quad x \in \mathbb{R}^p, \tag{2.23}$$

where ζ_k are independent random variables from the Gaussian distribution with mean zero and variance λ_k, where $\phi_k(x)$ and λ_k are eigenfunctions and eigenvalues of the integral operator A

$$A(f(x)) = \int_0^1 K(x,y) f(y)\, dx.$$

The substitution of (2.23) into (2.22) completes the proof of the theorem.

Krivyakova, Martynov, and Tyurin in (Krivyakova et al., 1977) proposed a method for calculating the eigenvalues of covariation operator (2.19), which leads to the following results. First, the eigenvalues are

$$\lambda_{kp} = (\frac{2}{\pi})^{2p} (2k+1)^{-2}, \quad k = 1, 2, ..., \tag{2.24}$$

with multiplicities q_{kp} such that the quantities $q_{kp} - 1$ equal the numbers of distinct representations of $2k+1$ as a product of p ordered factors. All the rest eigenvalues λ have multiplicity one and can be found as the solutions of the equation

$$\sum_{k=1}^{\infty} \frac{(q_{kp}+1)\lambda_{kp}^2}{\lambda_{kp} - \lambda} = \frac{1}{2^p}. \tag{2.25}$$

In some high dimensional cases it can be rather difficult from the computational point of view to establish the distribution of statistic T_n analytically. The main problems in application of the Theorem 10 are connected with calculations of eigenfunctions of the integral operator (2.19). In this case the critical region can be determined using Monte-Carlo simulations and the procedure discussed in the Remark in section 2.1.1.

Further we consider the statistics T_n (2.16) for some special cases of strongly negative definite kernels $L(x,y)$. Some largest coefficients of diagonalized quadratic form (2.17) are established numerically, for simplicity only two-dimensional case is discussed.

In case $p = 2$ the solutions of equation (2.25) were calculated numerically in (Krivyakova et al., 1977). The inverse values to the largest roots are presented

below.

15.8	88.0	203.6	359.7	604.9
843.3	1125.2	1578.2	1929.0	2237.0

After computation of eigenvalues of operator (2.19), let us pass over directly to corresponding eigenfunctions. Denote

$$K_0(x,y) := \min(x_1,y_1)\min(x_2,y_2)$$

and $K_1(x) = x_1x_2$, then the kernel $K(x,y)$ of integral operator (2.19) can be written in the form:

$$K(x,y) = K_0(x,y) - K_1(x)K_1(y).$$

Thus, the integral equation on eigenfunctions $\psi(x)$ and eigenvalue λ of operator (2.19) in bivariate case has the form

$$\int_{C^2} K_0(x,y)\psi(y)dy - K_1(x)\int_{C^2} K_1(y)\psi(y)dy = \lambda\psi(x), \tag{2.26}$$

where C^2 is the unit square.
Note, that eigenfunctions and corresponding eigenvalues of the integral operator with the kernel $K_0(x,y)$ are

$$\upsilon_{ij}(x,y) = 2\sin(\pi(i-\frac{1}{2})x)\sin(\pi(j-\frac{1}{2})y),$$

$$\mu_{ij} = (\pi^2(i-\frac{1}{2})(j-\frac{1}{2}))^{-2}.$$

Denote $z = \int_{C^2} K_1(y)\psi(y)dy$ and let us first look for eigenfunctions in (2.26) under condition $z = 0$. These can be only the eigenfunctions of operator with the kernel K_0. Since

$$\int_{C^2} \upsilon_{ij}(x)K_1(x)dx \neq 0 \quad \forall i,j,$$

then μ_{ii} $\forall i \in \mathbb{N}$ is not a solution of (2.26). If in the set $\{\mu_{ij}\}$ a number λ can be found $q \geq 2$ times, than the intersection of subspace based on corresponding to λ eigenfunctions in (2.26) with hyperplane $\{\psi : \langle K_1, \psi\rangle = 0\}$ is not empty and has the dimension $q-1$. This means that for each eigenvalue λ (2.24) appropriate $q-1$ eigenfunctions can be found as linear combinations of relevant functions

$\upsilon_{ij}(x)$, where $\mu_{ij} = \lambda$. For example,

$$\lambda_{12} = \frac{1}{9}\left[\frac{2}{\pi}\right]^4$$

and

$$\psi_{\lambda_{12}} = \frac{1}{\sqrt{2}}(\upsilon_{12}(x) + \upsilon_{21}(x)) =$$
$$= \frac{2}{\sqrt{2}}[\sin\frac{\pi x_1}{2}\sin\frac{3\pi x_2}{2} - \sin\frac{3\pi x_1}{2}\sin\frac{\pi x_2}{2}].$$

Now consider the case $z \neq 0$, then all the eigenvalues λ in (2.26) are found from the equation (2.25). Taking into account the completeness on the unit square of the family of functions $\{\upsilon_{ij}(x)\}$, $i, j \in \mathbb{N}$ the eigenfunctions $\psi_\lambda(x)$ can be found in form of decomposition to the series

$$\psi_\lambda(x) = \sum_{ij} a_{ij}\upsilon_{ij}(x).$$

In practice however the eigenfunctions $\psi_\lambda(x)$ can be approximated by a finite sequence

$$\psi_\lambda(x) \approx \sum_{i,j=1}^{N} a_{ij}\upsilon_{ij}(x) \tag{2.27}$$

using only the functions $\phi_{ij}(x)$ with the largest eigenvalues μ_{ij}.

Further we propose some numerical results and compute several largest coefficients of the diagonalized quadratic form (2.17)

$$Q = \sum_{i=1}^{\infty} a_i\zeta_i^2, \tag{2.28}$$

where ζ_i are independent random variables from the standard normal distribution.

In our calculations we consider 30 largest eigenvalues of operator (2.19). For all eigenvalues (2.25) we used approximation (2.27) with $N = 50$, after that the coefficients of quadratic form (2.17) were evaluated using the formula (2.18). As a result, the coefficients a_i in (2.28) were computed as eigenvalues of the matrix of quadratic form (2.17).

First consider test statistic T_n (2.16) with the kernel $L(x, y) = 1 - e^{-\|x-y\|^2}$, $x, y \in \mathbb{R}^2$. The inverse values to the largest coefficients a_i of quadratic form (2.28)

are presented below.

9.6	10.5	47.0	141.7	156.1
331.2	487.4	702.8	1042.2	1546.9
4774.8	5897.0	10524.0	15504.0	48801.0

Proposition 4. *In case of bivariate sample $X_1, ..., X_n$ statistic T_n with the kernel $L(x,y) = 1 - e^{-\|x-y\|^2}$ can be calculated using the formula*

$$T_n = n \left[\frac{2}{n}(1 - \pi\Upsilon(X_{i1})\Upsilon(X_{i2})) - \frac{1}{n^2}\sum_{ij=1}^{n} L(X_i, X_j) - C \right], \qquad (2.29)$$

where $C \approx 0.25777$ and $\Upsilon(z) = \Phi(\sqrt{2}(1-z)) - \Phi(-\sqrt{2}z)$.

Proof. Consider statistic T_n in the form (2.3) with $L(x,y) = 1 - e^{-\|x-y\|^2}$, $x, y \in [0,1]^2$. Let $\Phi(z)$, $z \in \mathbb{R}$ be the cumulative distribution function of standard normal distribution and $X\ X^{'}$ - independent random variables with the uniform distribution on $[0,1]^2$.
Each summand in the first sum in (2.3) has the form

$$\mathbf{E}L(X, X_i) = \int_0^1 \int_0^1 (1 - e^{-(x_1 - X_{i1})^2 - (x_2 - X_{i2})^2}) dx_1 dx_2 =$$
$$= 1 - \pi\Upsilon(X_{i1})\Upsilon(X_{i2}),$$

where $\Upsilon(z) = \Phi(\sqrt{2}(1-z)) - \Phi(-\sqrt{2}z)$.
We complete the proof with evaluation of the last summand in (2.3)

$$\mathbf{E}L(X, X^{'}) = \int_0^1 \int_0^1 (1 - \pi\Upsilon(x_1)\Upsilon(x_2)) dx_1 dx_2 =$$
$$= 1 - \pi \left(\int_0^1 \Phi(\sqrt{2}x_1) dx_1 - \int_{-1}^0 \Phi(\sqrt{2}x_1) dx_1 \right)^2 =: C \approx 0.25777.$$

In case $L(x,y) = \max(x_1, y_1) + \max(x_2, y_2) - \max(x_1, y_1)\max(x_2, y_2)$ the inverse values to the largest coefficients a_i of quadratic form (2.28) are presented below.

16.3	55.6	91.5	99.7	123.4
147.6	208.1	287.0	434.9	496.6
663.9	759.7	1343.5	1632.0	1717.3

Proposition 5. *In case of bivariate sample $X_1, ..., X_n$, $X_i = (X_{i1}, X_{i2})$ statistic T_n with the kernel*

$$L(x, y) = \max(x_1, y_1) + \max(x_2, y_2) - \max(x_1, y_1)\max(x_2, y_2)$$

can be calculated using the formula

$$T_n = n\left[\frac{2}{n}\sum_{i=1}^{n}(A_i + B_i - A_iB_i) - \frac{1}{n^2}\sum_{ij=1}^{n} L(X_i, X_j) - \frac{8}{9}\right], \qquad (2.30)$$

where $A_i = \frac{X_{i1}^2+1}{2}$ and $B_i = \frac{X_{i2}^2+1}{2}$.

Proof. Consider statistic T_n in the form (2.3) with $L(x, y) = \max(x_1, y_1) + \max(x_2, y_2) - \max(x_1, y_1)\max(x_2, y_2)$, $x, y \in [0, 1]^2$. Let X and $X^{'}$ be independent random variables with the uniform distribution on $[0, 1]^2$.
Each summand in the first sum in (2.3) has the form

$$\begin{aligned}
&\mathbf{E}L(X, X_i) = \int_0^1\int_0^1 (\max(x_1, X_{i1}) + \max(x_2, X_{i2}) - \\
&- \max(x_1, X_{i1})\max(x_2, X_{i2}))dx_1dx_2 = \\
&= \frac{X_{i1}^2 + 1}{2} + \frac{X_{i2}^2 + 1}{2} - \frac{(X_{i1}^2 + 1)(X_{i2}^2 + 1)}{4} =: S(X_{i1}, X_{i2}).
\end{aligned}$$

Then the last summand in (2.3) equals to

$$\mathbf{E}L(X, X^{'}) = \int_0^1\int_0^1 S(x_1, x_2)dx_1dx_2 = \frac{8}{9}$$

and this ends the proof of the proposition.

Further consider statistic T_n with the kernel $L(x, y) = \|x - y\|$, $x, y \in \mathbb{R}^2$. Let us start from one helpful lemma.

Lemma 6. *Let X be a random variable with the uniform distribution on the rectangle based on the vectors $(a, 0)$ and $(0, b)$, where a, b are fixed positive constants, then*

$$\begin{aligned}
H(a, b) \quad := \quad ab\mathbf{E}\|X\| = \frac{a^3}{6}\left(\frac{\sin\beta}{\cos^2\beta} + \ln|\tan(\frac{\pi}{4} + \frac{\beta}{2})|\right) + \\
+\frac{b^3}{6}\left(\frac{\sin\gamma}{\cos^2\gamma} + \ln|\tan(\frac{\pi}{4} + \frac{\gamma}{2})|\right),
\end{aligned}$$

where $\beta = \arccos \frac{a}{\sqrt{a^2+b^2}}$ *and* $\gamma = \frac{\pi}{2} - \beta$.

Proof. The mathematical expectation of X equals to

$$\mathbf{E}\|X\| = \frac{1}{ab}\int_0^a \int_0^b \sqrt{x^2+y^2}dxdy.$$

After conversion to polar coordinates we get

$$\begin{aligned}
&\mathbf{E}\|X\| = \frac{1}{ab}\left[\int_0^\beta \int_0^{\frac{a}{\cos\alpha}} r^2 dr d\alpha + \int_0^\gamma \int_0^{\frac{a}{\cos\alpha}} r^2 dr d\alpha\right] = \\
&= \frac{1}{ab}\left[\int_0^\beta \frac{a}{3\cos^3\alpha}d\alpha + \int_0^\gamma \frac{b}{3\cos^3\alpha}d\alpha\right] = \\
&= \frac{a^2}{6b}\left(\frac{\sin\beta}{\cos^2\beta} + \ln|\tan(\frac{\pi}{4}+\frac{\beta}{2})|\right) + \\
&+\frac{b^2}{6a}\left(\frac{\sin\gamma}{\cos^2\gamma} + \ln|\tan(\frac{\pi}{4}+\frac{\gamma}{2})|\right),
\end{aligned}$$

where $\beta = \arccos \frac{a}{\sqrt{a^2+b^2}}$ and $\gamma = \frac{\pi}{2} - \beta$.

Proposition 7. *In case of bivariate sample* $X_1, ..., X_n$ *statistic* T_n *(2.16) with* $L(x,y) = \|x-y\|$ *can be calculated using the formula*

$$T_n = n * \left[\frac{2}{n}\sum_{i=1}^n \Lambda(X_i) - \frac{1}{n^2}\sum_{i,j=1}^n \|X_i - X_j\| - C\right], \tag{2.31}$$

where $C \approx 0.5214054$*,*

$$\begin{aligned}
&\Lambda(X_i) = H(1-X_{i1}, 1-X_{i2}) + H(X_{i1}, 1-X_{i2}) + \\
&+H(X_{i1}, X_{i2}) + H(1-X_{i1}, X_{i2}),
\end{aligned}$$

and function $H(\cdot,\cdot)$ *is defined in Lemma 6.*

Proof. Let us consider statistic T_n in the form (2.3) with $L(x,y) = \|x-y\|$, $x, y \in \mathbb{R}^2$,

$$\mathbf{E}\|X - X'\| = \int_{[0,1]^4} \|x-y\|dxdy = C \approx 0.5214054,$$

where X and $X^{'}$ are independent random variables from the uniform distribution on unit square.

$$\mathbf{E}\|X - X_i\| = \int_{[0,1]^2} \|x - X_i\| dx = \int_{-X_{i1}}^{1-X_{i1}} \int_{-X_{i2}}^{1-X_{i2}} \|x\| dx =$$
$$= \int_0^{1-X_{i1}} \int_0^{1-X_{i2}} \|x\| dx + \int_{-X_{i1}}^0 \int_0^{1-X_{i2}} \|x\| dx +$$
$$+ \int_{-X_{i1}}^0 \int_{-X_{i2}}^0 \|x\| dx + \int_0^{1-X_{i1}} \int_{-X_{i2}}^0 \|x\| dx =$$
$$= H(1 - X_{i1}, 1 - X_{i2}) + H(X_{i1}, 1 - X_{i2}) +$$
$$+ H(X_{i1}, X_{i2}) + H(1 - X_{i1}, X_{i2}),$$

where function $H(\cdot, \cdot)$ is defined in the previous Lemma.

In case $L(x, y) = \|x - y\|$ the inverse values to the largest coefficients a_i of quadratic form (2.28) are presented below.

6.6	7.6	9.2	19.3	39.7
54.1	98.5	119.0	126.9	151.5
174.3	274.8	320.5	407.0	611.6

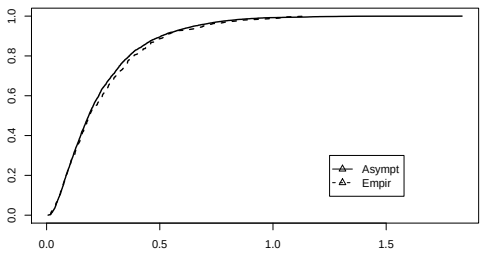

Figure 2.1. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = 1 - e^{-\|x-y\|^2}$, $n = 100$.

A comparison of empirical distribution function of statistics T_n (2.16) and the distribution function of quadratic forms (2.28) is shown in Fig. 2.1–2.3. The

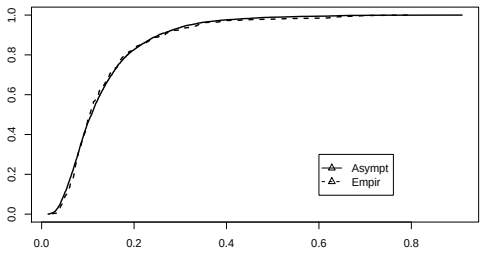

Figure 2.2. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = \max(x_1, y_1) + \max(x_2, y_2) - \max(x_1, y_1)\max(x_2, y_2)$, $n = 100$.

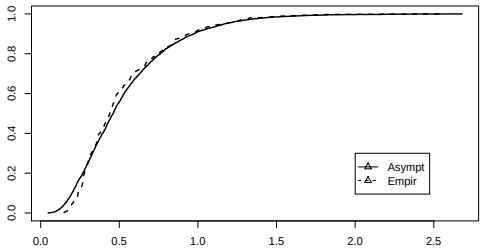

Figure 2.3. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = \|x - y\|$, $n = 100$.

empirical distribution of T_n with considered kernels was calculated by simulation of 400 samples of size 100 form the uniform distribution on the unit square.

2.2. Composite hypothesis

Let $X_1, \ldots, X_n$ be the sample of independent observations of random variable X with continuous probability distribution functions $F(x)$, assumed unknown. The null composite hypothesis in the problem of testing goodness of fit is

$$H_0 : F(x) \in \mathbf{\Lambda} = \{G(x, \theta),\ x \in \mathbb{R}^p,\ \theta \in \Theta \subset \mathbb{R}^d\},$$

where $\mathbf{\Lambda}$ - is a parametric family of distribution functions.

In comparison with simple hypothesis this problem is more interesting from practical point of view, because in reality exact distributions in goodness of fit problems are unknown.

The statistic for testing H_0, based on N-distance with the kernel $L(x, y)$, has the form

$$T_n = -n \int_{R^p} \int_{R^p} L(x, y)\, d(F_n(x) - G(x, \hat{\theta}_n))\, d(F_n(y) - G(y, \hat{\theta}_n)), \quad (2.32)$$

where $\hat{\theta}_n$ is the estimate of unknown parameter θ under the assumption that X has the distribution from family $\mathbf{\Lambda}$, $F_n(x)$ - empirical distribution function.

We should reject H_0 in case of large values of test statistic, that is if $T_n > c_\alpha$. Where c_α can be found from the equation

$$P_0(T_n > c_\alpha) = \alpha,$$

here P_0 is a probability distribution corresponding to the null hypothesis and α - size of the test.

We will follow the outline of presentation determined in the previous section and first consider the asymptotic distribution of statistics T_n (2.32).

2.2.1. Asymptotic distribution of test statistic

In more detail, the problem of asymptotic distribution of T_n will be discussed in univariate case. In the first subsection the general procedure to determine the limit distribution of test statistics is presented. After that some practical results for normality and exponetiality tests are provided.

In case of arbitrary dimension the normality test is only considered. Though, the asymptotic distribution of test statistics is not established, the critical region of proposed criteria is determined using Monte Carlo simulations.

Univariate case

By analogy with simple hypothesis problem in the previous section let us consider statistic T_n with a special form of strongly negative definite kernel

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(G(x, \hat{\theta}_n), G(y, \hat{\theta}_n))\, d\Delta_n(x, \hat{\theta}_n)\, d\Delta_n(y, \hat{\theta}_n), \quad (2.33)$$

where $\Delta_n(x, \hat{\theta}_n) = F_n(x) - G(x, \hat{\theta}_n)\, F_n(x)$ is the empirical distribution function based on the sample $X_1, \ldots, X_n$. Statistic (2.33) can be rewritten in the form

$$T_n = -n \int_0^1 \int_0^1 L(x, y)\, d(F_n^*(x) - x)\, d(F_n^*(y) - y), \quad (2.34)$$

where $F_n^*(x)$ is the empirical distribution function based on transformed sample $t_1, \ldots, t_n$, $t_i = G(X_i, \hat{\theta}_n)$, $i = 1, 2, ..., n$.
Further we assume that an estimate $\hat{\theta}_n$ of the vector-parameter $\theta \in \Theta \subset \mathbb{R}^d$ can be represented in the form

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(X_i, \theta) + R_n, \quad (2.35)$$

where $|R_n| \xrightarrow{P} 0$, $l(x, \theta) = (l_1(x, \theta), ..., l_d(x, \theta))^t$,

$$\mathbf{E} l(x, \theta) = 0,$$

$$\mathbf{E} l(x, \theta) l(x, \theta)^t = B(\theta),$$

where $B(\theta)$ - is a positive-definite matrix. Such property, under some regularity conditions (Durbin, 1973) is satisfied, for example, by the maximum likelihood estimate. In this case

$$l(x, \theta) = I^{-1}(\theta) S(x, \theta),$$

where $S(x, \theta) = (\frac{\partial}{\partial \theta_1} \ln g(x, \theta), ..., \frac{\partial}{\partial \theta_d} \ln g(x, \theta))$, $g(x, \theta) = G'_x(x, \theta)$ and $I(\theta)$ is Fisher information matrix.

The following assumptions will be made in addition to (2.35). Let U denotes the closure of a given neighborhood of θ.

1. $G(x, \theta)$ is continuous in x for all $\theta \in U$.
2. The vector-valued function $q(t, \theta) = \frac{\partial G(x, \theta)}{\partial \theta}$, $t = G(x, \theta)$ exists and is

continuous in (t, θ) for all $\theta \in U$ and all $0 \leq t \leq 1$.

Theorem 11. *Let $\hat{\theta}_n$ be the maximum likelihood estimate of θ. On assumptions stated above, the null asymptotic distribution of statistic T_n (2.34) will coincide with distribution of quadratic form*

$$Q = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{kj} \zeta_k \zeta_j, \tag{2.36}$$

where ζ_k are independent random variables from the standard normal distribution and

$$a_{ij} = -\sqrt{\lambda_i \lambda_j} \int_0^1 \int_0^1 L(x, y) d\psi_i(x) d\psi_j(y), \tag{2.37}$$

where λ_k and $\psi_k(x)$ are eigenvalues and functions of integral operator A

$$Af(x) = \int_0^1 K(x, y) f(y) dy, \tag{2.38}$$

where

$$K(x, y) = \min(x, y) - xy - q(x, \theta) I^{-1}(\theta) q(y, \theta), \tag{2.39}$$

where $q(x, \theta) = G'_\theta(z, \theta)$, $x = G(z, \theta)$, $z \in \mathbb{R}$ and $I(\theta)$ - Fisher information matrix.

Proof. The outline of the proof coincides with the proof of Theorem 9. The weak limit of random process $Z_n(x, \theta) = F_n^*(x) - x$, where $F_n^*(x)$ is the empirical distribution function based on the sample $t_1, \ldots, t_n$, $t_i = G(X_i, \hat{\theta}_n)$, $i = 1, 2, ..., n$, was obtained in Theorem 1 in (Durbin, 1973)

$$Z_n(x, \theta) \xrightarrow{D} \xi(x, \theta), \tag{2.40}$$

$\xi(x, \theta)$ is a Gaussian process with mean zero and correlation function

$$K(x, y) = \min(x, y) - xy - q(x, \theta) I^{-1}(\theta) q(y, \theta),$$

where $q(x, \theta) = G'_\theta(z, \theta)$, $x = G(z, \theta)$ and

$$I(\theta) = \mathbf{E} \left\| \frac{\partial}{\partial \theta_i} \ln g(x, \theta) \frac{\partial}{\partial \theta_j} \ln g(x, \theta) \right\|.$$

After that the statement of the theorem follows immediately from the continuity of functional (2.34), established in Lemma 2.

It is obvious from the form of the correlation function (2.39) that under H_0, T_n is not distribution-free since its asymptotic distribution depends on G. Worse, it is not even asymptotically parameter-free since this distribution depends in general on the value of unknown parameter θ. However, in some cases we can avoid this parametric dependence. These include cases when $\mathbf{\Lambda}$ is a location-scale family of distribution functions, that is

$$\mathbf{\Lambda} = \{G(\frac{x-\theta_1}{\theta_2}),\ \theta_1 \in \mathbb{R},\ \theta_2 > 0\}.$$

The properties of random processes $\xi(x)$ (2.40) in this case were in detail studied in (Martynov, 1978). If both parameters are unknown, the correlation function of $\xi(x)$ has the form

$$K(x,y) = \min(x,y) - xy - \frac{1}{a}K_1(x,y), \tag{2.41}$$

where

$$\begin{aligned} K_1(x,y) &= c_2 w_1(x) w_2(y) + c_1 w_2(x) w_2(y) - \\ &\quad - c_3(w_1(x)w_2(y) + w_2(x)w_1(y)), \end{aligned}$$

$$w_1(x) = g(G^{-1}(x)), \qquad w_2(x) = G^{-1}(x)g(G^{-1}(x)),$$

$$c_1 = \int_{-\infty}^{+\infty} \frac{(g'(x))^2}{g(x)}dx, \qquad c_2 = \int_{-\infty}^{+\infty} x^2 \frac{(g'(x))^2}{g(x)}dx - 1,$$

$$c_3 = \int_{-\infty}^{+\infty} x\frac{(g'(x))^2}{g(x)}dx, \qquad a = c_1c_2 - c_3^2.$$

As an implication from location-scale family for one more family $\mathbf{\Lambda}$ $K(x,y)$ also does not depend on unknown parameters. This is a set of distribution functions with scale and shape parameters of the form

$$\mathbf{\Lambda} = \{G((\frac{x}{\theta_1})^{\theta_2}),\ \theta_1 > 0,\ \theta_2 > 0\},$$

which includes such well-known distributions as Weibull, Log-logistic and others. In this case the kernel $K(x,y)$ will have the same form (2.41), where

$$w_1(x) = G^{-1}(x)g(G^{-1}(x)), \qquad w_2(x) = G^{-1}(x)g(G^{-1}(x))\ln G^{-1}(x),$$

$$c_1 = \int_{-\infty}^{+\infty} \left(1 + \frac{g'(x)}{g(x)}x\right)^2 g(x)dx,$$

$$c_2 = \int_{-\infty}^{+\infty} \left(1 + \frac{g'(x)}{g(x)}x\ln x + \ln x\right)^2 g(x)dx,$$

$$c_3 = \int_{-\infty}^{+\infty} \left(1 + \frac{g'(x)}{g(x)}x\right)\left(1 + \frac{g'(x)}{g(x)}x\ln x + \ln x\right) g(x)dx,$$

$$a = c_1c_2 - c_3^2.$$

The eigenvalues of integral operator (2.38) with the kernel (2.39) consists of two parts (Martynov, 1978). Firstly, these are the set of numbers

$$\{(\pi k)^{-2} : |b_k| = 0, k = 1, 2...\}, \tag{2.42}$$

where $b_k = \sqrt{2}\int_0^1 \beta(x)\sin(\pi kx)dx$, $\beta(x) = I^{-1/2}(\theta)q(x,\theta)$. And secondly - the solutions λ of the following equation

$$det\left(E - \sum_{k=1}^{\infty} \frac{b_k b_k^t}{(\pi k)^{-2} - \lambda}\right) = 0. \tag{2.43}$$

Further we discuss the above mentioned procedure in more detail on the basis of normality and exponentiality tests. Numerical results for the largest coefficients of diagonalized quadratic form (2.36) are proposed.

Remark

In the most general case of family $\mathbf{\Lambda}$ the parametric dependence problem of the distribution of statistic T_n can be overcome by utilization of parametric bootstrap methods (Stute et al., 1993; Szucs, 2008).

Consider statistic T_n based on the empirical process

$$Z_n(x) = \sqrt{n}(F_n(x) - G(x, \hat{\theta}_n))$$

and let T_n^* be the corresponding statistic based on the bootstrapped estimated empirical process

$$Z_n^*(x) = \sqrt{n}(F_n^*(x) - G(x, \hat{\theta}_n^*)),$$

where $F_n^*(x)$ is the empirical distribution function constructed from the sample $X_1^*, ..., X_n^*$ independently generated having distribution function $G(x, \hat{\theta}_n)$ and $\hat{\theta}_n^*$ be an estimator of $\hat{\theta}_n$ based on the generated sample. If T_n has a continuous asymptotic distribution function, then we can test H_0 by the following algorithm

1. Calculate the estimator $\hat{\theta}_n$ based on $X_1, ..., X_n$.
2. Calculate T_n.
3. Generate random values $X_1^*, ..., X_n^*$ having distribution function $G(x, \hat{\theta}_n)$.
4. Calculate the estimator $\hat{\theta}_n^*$ of $\hat{\theta}_n$ based on bootstrap sample.
5. Calculate T_n^*.
6. Repeat steps 3-5 N times, let $T_{n,1}^* \leq ... \leq T_{n,N}^*$ be the order statistic of the resulting N values of T_n^* and let c_α be $(1-\alpha)$ empirical quantile of T_n^*.
7. Reject H_0 if T_n is greater than c_α.

The validity of this procedure under very general conditions on the distribution function $G(x, \hat{\theta})$ and parameter estimation methods is ensured by Theorem 1 in (Szucs, 2008). However in this thesis we omit the details and refer the reader to the mentioned references.

Normality test

Denote by $\Phi(x)$ and $\varphi(x)$ the cumulative distribution and probability density functions of standard normal distribution. Consider the case when both location and scale parameters of hypothesized distribution are unknown.

First, transform initial sample $X_1, \ldots, X_n$ to the sample $t_1, \ldots, t_n$, where $t_i = \Phi(X_i, \bar{X}, \hat{S}^2)$ and $\bar{X}$, $\hat{S}^2$ are maximum likelihood estimates of unknown parameters of mean and variation. And then test the null hypothesis using statistic (2.34).

After described transformation statistic T_n with different kernels can be calculated using the formulas in proposition 3.

The asymptotic distribution of statistic T_n coincides with distribution of quadratic form (2.36), where the kernel (2.39) of the integral operator A (2.38) has the form

$$\begin{aligned} K(x,y) &= \min(x,y) - xy - \varphi(\Phi^{-1}(x))\varphi(\Phi^{-1}(y)) \\ &\quad -\frac{1}{2}\Phi^{-1}(x)\Phi^{-1}(y)\varphi(\Phi^{-1}(x))\varphi(\Phi^{-1}(y)). \end{aligned}$$

To obtain the expressions of the coefficients of quadratic form (2.36) let us pass over directly to calculation of eigenvalues and functions of operator A (2.38). Since $b_k = (b_{k,1}, b_{k,2}) \neq 0$ (2.42), $\forall k = 1, 2...$, where

$$b_k = \sqrt{2}\int_0^1 \beta(x)\sin(\pi k x)dx,$$

$$\beta(x) = \left(\varphi(\Phi^{-1}(x)), \frac{1}{\sqrt{2}}\Phi^{-1}(x)\varphi(\Phi^{-1}(x))\right),$$

the eigenvalues of operator A are only the solutions of equation (2.43). Note, that the coefficients $b_{k,1}$ with even numbers and $b_{k,2}$ with odd numbers are equal to zero. Thus, all eigenvalues λ_k, $k = 1, 2...$, of A are the combination of the solutions of the following equations:

$$\sum_{k=1}^{\infty} \frac{b_{2k-1,1}^2}{((2k-1)\pi)^{-2} - \lambda} = 1, \tag{2.44}$$

$$\sum_{k=1}^{\infty} \frac{b_{2k,2}^2}{(2k\pi)^{-2} - \lambda} = 1. \tag{2.45}$$

The left part of all equations is strictly monotone in each of the corresponding intervals $\left(((2k-1)\pi)^{-2}, ((2k+1)\pi)^{-2}\right)$, $\left((2k\pi)^{-2}, ((2k+2)\pi)^{-2}\right)$ and has in it only one solution. This fact allows us to use simple numerical methods to find the sufficient number of eigenvalues and gives us an estimate of their convergence to zero.

The inverse values to the largest solutions λ_k, $k = 1, 2...$, of equations are presented below.

54.5	186.7	396.6	684.7	1051.2
1496.1	2019.6	2621.7	3302.7	4062.3
74.4	229.1	463.3	776.8	1169.4
1641.1	2192.1	2821.7	3530.7	4319.1

After computation of eigenvalues of operator (2.38), let us pass over directly

to corresponding eigenfunctions. By analogy with bivariate simple hypothesis (see section 2.1.3) the eigenfunctions $\psi_{\lambda_k}(x)$ can be found in the form of decomposition to the series

$$\psi_{\lambda_k}(x) = \sum_{i=1}^{\infty} \alpha_i \sin(\pi i x). \tag{2.46}$$

The choice of the family $\{\sin(\pi i x),\ i = 1, 2...\}$ is justified by two facts: firstly, this family is complete on the interval $[0, 1]$ and secondly, functions $\sqrt{2}\sin(\pi i x)$ are the eigenfunctions of integral operator (2.38) with the kernel $K_0(x, y) = min(x, y) - xy$, which is a part of the kernel $K(x, y)$.

As it was noticed in section 2.1.2 for some of the kernels $L(x, y)$ the following equalities are fulfilled $\forall k, j \in \mathbb{N}$

$$-2\int_0^1\int_0^1 |x-y|\, d\sin(\pi k x)\, d\sin(\pi j y) = 2\delta_{kj}, \tag{2.47}$$

$$-2\int_0^1\int_0^1 \max(x, y)\, d\sin(\pi k x)\, d\sin(\pi j y) = \delta_{kj}. \tag{2.48}$$

These properties help us to avoid the calculation of the coefficients of the series (2.46) and leads to the result

- In case $L(x, y) = \max(x, y)$ the coefficients a_{kj} of quadratic form (2.36) equals to
$$a_{kj} = \sqrt{\lambda_k \lambda_j}\delta_{kj},$$
- In case $L(x, y) = |x - y|$
$$a_{kj} = 2\sqrt{\lambda_k \lambda_j}\delta_{kj},$$

where λ_k are the solutions of the equations (2.44) and (2.45).

For other kernels $L(x, y)$, in practice, the eigenfunctions $\psi_{\lambda_k}(x)$ can be approximated by a finite sequence

$$\psi_{\lambda_k}(x) = \sum_{i}^{N} \alpha_i \sin(\pi i x) \tag{2.49}$$

using only the functions $\sin(\pi i x)$ with $i < K$, $K \in \mathbb{N}$.

Further we propose some numerical results and compute several largest coef-

ficients of the diagonalized quadratic form (2.36)

$$Q = \sum_{i=1}^{\infty} a_i \zeta_i^2, \tag{2.50}$$

where ζ_i are independent random variables from the standard normal distribution.

In our calculations we consider 20 largest eigenvalues of operator (2.38), presented in the table above. For all eigenvalues we used approximation (2.49) with $N = 50$, after that the coefficients of quadratic form (2.36) were evaluated using the formula (2.37). As a result, the coefficients a_i in (2.50) were computed as eigenvalues of obtained matrix of quadratic form (2.36).

First consider test statistic T_n (2.34) with the kernel $L(x, y) = |x - y|^{\frac{3}{2}}$. The inverse values to the largest coefficients a_i of quadratic form (2.50) are presented below.

55.3	100.6	281.2	428.2	795.9
1045.5	1628.9	2029.1	2852.0	3438.8
4507.5	5257.0	6619.5	7600.7	9378.2
10387.0	10682.0	12765.0	13231.0	16319.0

In case $L(x, y) = \frac{|x-y|}{1+|x-y|}$ the inverse values to the largest coefficients a_i of quadratic form (2.50) are presented below.

33.5	42.3	103.6	120.6	149.1
194.5	259.6	264.8	388.1	441.2
588.3	656.3	847.7	921.6	1125.6
1236.5	1450.5	1496.6	1797.8	2320.0

A comparison of empirical distribution function of statistic T_n (2.34) and the distribution function of quadratic forms (2.50) is shown in Fig. 2.4–2.6. The empirical distribution of T_n with considered kernels was calculated by simulation of 600 samples of size 100 form the normal distribution with mean 1 and variance 2.

Exponentiality test

The second example is devoted to the hypothesis of exponentiality with unknown mean parameter, that is

$$G(x, \theta) = 1 - e^{\frac{x}{\theta}}, \theta > 0.$$

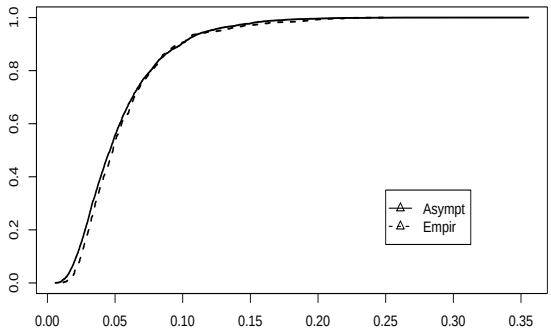

Figure 2.4. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = \max(x, y)$, $n = 100$.

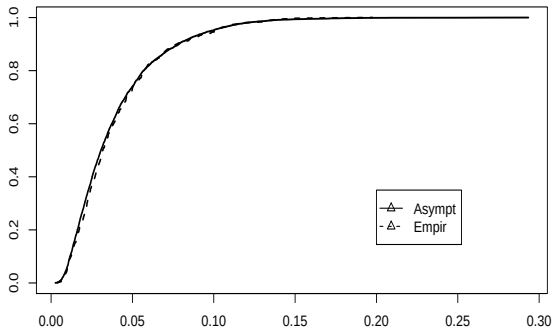

Figure 2.5. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = |x - y|^{\frac{3}{2}}$, $n = 100$.

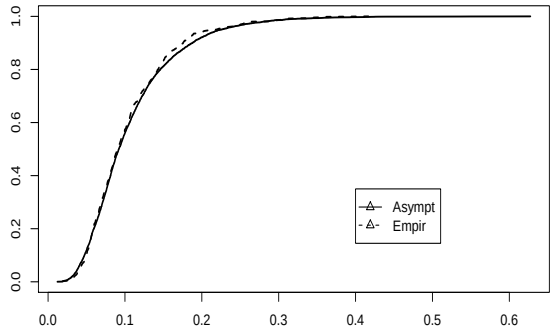

Figure 2.6. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x,y) = \frac{|x-y|}{1+|x-y|}$, $n = 100$.

After transformation of initial sample $X_1, \ldots, X_n$ to the sample $t_1, \ldots, t_n$, where $t_i = G(X_i, \bar{X})$ and $\bar{X}$ is a maximum likelihood estimate of unknown mean, the null hypothesis is tested using statistic (2.34). Proposition 3 provides the formulas for calculation of tests statistic T_n after described transformation.

The asymptotic distribution of statistic T_n coincides with distribution of quadratic form (2.36), where the kernel (2.39) of the integral operator A (2.38) has the form

$$K(x,y) = \min(x,y) - xy - (1-x)\log(1-x)(1-y)\log(1-y). \qquad (2.51)$$

The set of eigenvalues λ_k, $k = 1, 2, ...$, of the integral operator A with the kernel (2.51) coincides with the set of the solutions λ of equation (2.43), which in this case has the form

$$\sum_{k=1}^{\infty} \frac{2c_k^2}{(\pi k)^4} \frac{1}{(\pi k)^{-2} - \lambda} = 1, \qquad (2.52)$$

where $c_k = \int_0^1 \frac{\sin(\pi k x)}{x} dx$, $k = 1, 2, ...$
The left part of the equation (2.52) is monotonically increasing in each of the inter-

vals $\left((k\pi)^{-2}, ((k+1)\pi)^{-2}\right)$ and has only one solution in it. This helps us to find the sufficient number of eigenvalues numerically.

The inverse values to the largest solutions λ_k, $k = 1, 2...$, of equation (2.52) are presented below.

23.8	58.4	122.6	196.5	300.3
413.7	556.9	709.7	892.5	1084.8
1307.0	1538.8	1800.5	2071.8	2373.0
2683.7	3024.4	3374.7	3754.8	4144.6

By analogy with normality hypothesis corresponding eigenfunctions $\psi_{\lambda_k}(x)$ can be found in form of expansion to the series (2.46).

Taking into account the properties (2.47), (2.48) of the kernels $L(x, y) = |x - y|$ and $L(x, y) = \max(x, y)$, these is no need to calculate the coefficients of the series (2.46) and the limit distribution of T_n will coincide with the distribution of quadratic forms

- $L(x, y) = \max(x, y)$,
$$\sum_{k=1}^{\infty} \lambda_k^2 \xi_k^2;$$
- $L(x, y) = |x - y|$,
$$2\sum_{k=1}^{\infty} \lambda_k^2 \xi_k^2,$$

where ξ_k are independent random variables from the standard normal distribution.

For all the other kernels $L(x, y)$ the eigenfunctions $\psi_{\lambda_k}(x)$, in practice, can be approximated by a finite sequence (2.49). Further we propose some numerical results and compute several largest coefficients of the diagonalized quadratic form (2.50).

In our calculations we consider 20 largest solutions of equation (2.52), presented in the table above. For all eigenvalues we used approximation (2.49) with $N = 50$, after that the coefficients of quadratic form (2.36) were evaluated using the formula (2.37). As a result, the coefficients a_i in (2.50) were computed as eigenvalues of obtained matrix of quadratic form (2.36).

First consider test statistic T_n (2.34) with the kernel $L(x, y) = |x - y|^{\frac{3}{2}}$. The inverse values to the largest coefficients a_i of quadratic form (2.50) are presented in the table below.

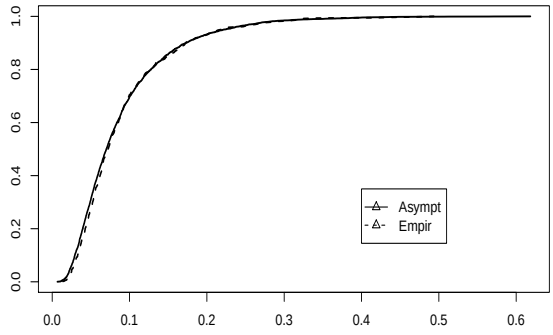

Figure 2.7. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x, y) = \max(x, y)$, $n = 100$.

16.6	73.7	177.5	351.5	594.5
921.8	1324.7	1840.2	2420.0	3153.1
3975.2	4912.9	5956.0	7173.3	8490.9
9985.1	11172.0	12731.0	13441.0	18271.0

In case $L(x, y) = \frac{|x-y|}{1+|x-y|}$ the inverse values to the largest coefficients a_i of quadratic form (2.50) are presented in the table below.

17.8	34.1	70.7	105.0	154.8
190.9	227.7	305.1	386.4	487.6
610.3	703.5	812.9	1020.1	1099.2
1256.1	1615.0	1861.4	2621.6	4560.8

A comparison of empirical distribution function of statistic T_n (2.34) and the distribution function of quadratic forms (2.50) is shown in Fig. 2.7–2.9. The empirical distribution of T_n with considered kernels was calculated by simulation of 600 samples of size 100 form the exponential distribution with mean 2.

Remark

Both for normality and exponentiality tests another procedure for testing the null hypothesis can be suggested. Instead of transformation of initial sample to the

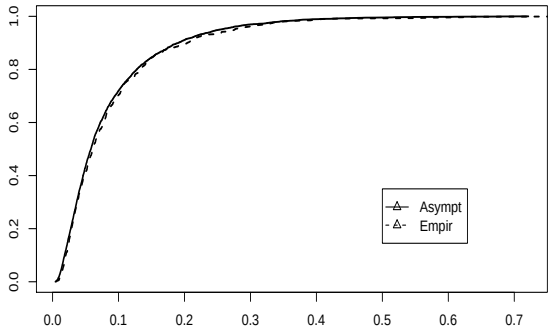

Figure 2.8. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x,y) = |x-y|^{\frac{3}{2}}$, $n = 100$.

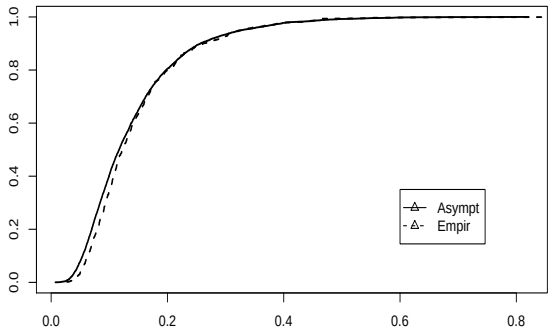

Figure 2.9. Empirical and asymptotic distribution of statistic T_n with the kernel $L(x,y) = \frac{|x-y|}{1+|x-y|}$, $n = 100$.

interval $[0,1]$ we can first standardize our sample $X_1,\dots,X_n$ using the formula below, without loss of generality only normality case is discussed:

$$Y_i = \hat{\sigma}^{-1/2}(X_i - \bar{X}), \quad i = 1, ..., n,$$

where $\bar{X}$ and $\hat{\sigma}$ are maximum likelihood estimates of mean and variance.

The null hypothesis can be tested using statistic

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x,y)\, d(F_n(x) - \Phi(x))\, d(F_n(y) - \Phi(y)), \qquad (2.53)$$

where $F_n(x)$ is the empirical distribution function, based on the sample $Y_1, ..., Y_n$ and $\Phi(x)$ is standard normal distribution function.
One can see, that the joint distribution of $Y_1, ..., Y_n$ does not depend of unknown parameters, therefore the distribution of (2.53) can be modeled with the help of simulations.

2.2.2. Multivariate normality test

In general case of arbitrary dimension to obtain the limit distribution of test statistic (2.32) in the form of distribution of infinite quadratic form becomes a rather complicated task. The main difficulties here are connected with calculation of eigenvalues and functions of a certain integral operator. However, for specific families Λ (see section 2.2), some alternative procedures to determine the critical region of the test can be proposed. In this section we consider normality criterion, as the most widespread problem among multivariate goodness of fit tests.

Let $X_1,\dots,X_n$ be a p-variate sample of independent observations of random variable X with distribution function $F(x)$. Consider the problem of testing the hypothesis

$$H_0 : F(x) \in N_p(a, \Sigma),$$

where a and Σ are mathematical expectation vector and covariance matrix of normal distribution, assumed unknown.

The distribution of test statistic T_n (2.2), applied for testing H_0, is dependent on unknown parameters of hypothesized normal distribution. To avoid this, first standardize the initial sample using the formula:

$$Y_k = \hat{S}^{-1/2}(X_k - \bar{X}),\ k = 1, ..., n, \qquad (2.54)$$

where $\bar{X}$ and $\hat{S}$ are maximum likelihood estimates for a and Σ.

Transformed sample will asymptotically have the p-variate standard normal distribution. And let us reject the hypothesis H_0 in case of large values of statistic

$$T_n = -n \int_{R^{2p}} L(x,y)\, d(F_n(x) - \Phi(x))\, d(F_n(y) - \Phi(y)), \tag{2.55}$$

where $F_n(x)$ is the empirical distribution function, base on the sample $Y_1, \ldots, Y_n$ and $\Phi(x)$ - distribution function of p-variate standard Gaussian distribution.

In bivariate case Pettitt in (Pettitt, 1979) studies the properties of empirical process $\sqrt{n}(F_n(x) - \Phi(x))$ after its transformation to the unit square, that is

$$Z_n(t) = \sqrt{n}(F_n^*(t) - t_1 t_2),$$

where $F_n^*(t)$ is a bivariate empirical distribution function, i.e. $F_n^*(t)$ is the fraction of the Y_i for which the inequalities $\Phi(Y_{i1}) \leq t_1$ and $\Phi(Y_{i2} \leq t_2)$ both hold, here Φ is the standard normal distribution function. The covariance function of the process $Z_n(t)$ is

$$\begin{aligned}
&K(s,t) = \min(t_1, s_1)\min(t_2, s_2) - t_1 t_2 s_1 s_2 - \\
&-\phi(\Phi(t_1))\phi(\Phi(s_1))t_2 s_2 - \phi(\Phi(t_2))\phi(\Phi(s_2))t_1 s_1 - \\
&-\frac{1}{2}(\Phi(t_1)\Phi(s_1)\phi(\Phi(t_1))\phi(\Phi(s_1))t_2 s_2 + \\
&+\Phi(t_2)\Phi(s_2)\phi(\Phi(t_2))\phi(\Phi(s_2))t_1 s_1) - \\
&-\phi(\Phi(t_1))\phi(\Phi(t_2))\phi(\Phi(s_1))\phi(\Phi(s_2)),
\end{aligned}$$

where $\phi(x) = \Phi^{'}(x)$.

However the calculation of the eigenvalues and functions of integral operator with the kernel $K(s,t)$ is rather complicated. To avoid this, let us note, that the joint distribution of $Y_1, \ldots, Y_n$ asymptotically does not depend on unknown parameters a and Σ. This fact allows us to estimate the percentiles of the null distribution of T_n by means of Monte Carlo simulations.

Statistics T_n with different strongly negative definite kernels can be calculated using the following proposition.

Proposition 8. *In case of bivariate sample $Y_1, ..., Y_n$, $Y_i = (Y_{i1}, Y_{i2})$ statistics T_n (2.56) can be calculated using the formulas*

- $L(x,y) = \|x - y\|$,

$$T_n = 2\sum_{i=1}^{n}\Upsilon(Y_i) - \frac{1}{n}\sum_{i,j=1}^{n}\|Y_i - Y_j\| - 2n\Gamma(3/2),$$

where

$$\Upsilon(z) = \sqrt{2}\Gamma(\frac{3}{2}) + \sqrt{\frac{2}{\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!2^k}\frac{\|z\|^{2k+2}}{(2k+1)(2k+2)}\frac{\Gamma(3/2)\Gamma(k+\frac{3}{2})}{\Gamma(k+2)},$$

$z \in \mathbb{R}^2$.

- $L(x,y) = 1 - e^{-\|x-y\|^2}$,

$$T_n = 2\sum_{i=1}^{n}(1 - \frac{1}{3}e^{-\frac{Y_{i1}^2}{3}} * e^{-\frac{Y_{i2}^2}{3}}) - \frac{1}{n}\sum_{i,j=1}^{n}L(Y_i,Y_j) - \frac{4n}{5}.$$

- $L(x,y) = \Phi(x_1 \vee y_1) + \Phi(x_2 \vee y_2) - \Phi(x_1 \vee y_1)\Phi(x_2 \vee y_2)$,

$$T_n = 2\sum_{i=1}^{n}\Upsilon(Y_{i1}) + \Upsilon(Y_{i2}) - \Upsilon(Y_{i1})\Upsilon(Y_{i2}) - \frac{1}{n}\sum_{i,j=1}^{n}L(Y_i,Y_j) - \frac{8n}{9},$$

where $\Upsilon(z) = \frac{1+\Phi^2(y)}{2}$ *and* $\Phi(z)$ *- distribution function of univariate standard normal distribution.*

Proof. Statistic T_n with the kernel $L(x,y) = \|x-y\|$ was obtained by Szekely and Rizzo in (Szekely and Rizzo, 2005). For two other kernels the formulas were derived from representation (2.3) by calculating corresponding mathematical expectations. Let $Y_i = (Y_{i1}, Y_{i2})$, $i = 1, ..., n$, be an element from the standardized sample (2.54) and X, X' two independent random variables from the bivariate standard normal distribution, then

- $L(x,y) = 1 - e^{-\|x-y\|^2}$, $x, y \in \mathbb{R}^2$,

$$\begin{aligned}\mathbf{E}L(X,Y_i) &= \frac{1}{2\pi}\int\int\left(1 - e^{-(x-Y_{i1})^2-(y-Y_{i2})^2}\right)e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = \\ &= 1 - \frac{1}{3}e^{-\frac{Y_{i1}^2}{3}-\frac{Y_{i2}^2}{3}},\end{aligned}$$

$$\mathbf{E}L(X,X') = 1 - \frac{1}{6\pi}\int\int e^{-\frac{x^2}{3}-\frac{y^2}{3}}e^{-\frac{x^2}{2}-\frac{y^2}{2}}dxdy = \frac{4}{5}.$$

- $L(x,y) = \Phi(x_1 \vee y_1) + \Phi(x_2 \vee y_2) - \Phi(x_1 \vee y_1)\Phi(x_2 \vee y_2)$,

where $x = (x_1, x_2)$ and $y = (y_1, y_2)$,
From equality

$$\int \Phi(\max(x,y))d\Phi(x) = \frac{1+\Phi^2(y)}{2}$$

follows that

$$\begin{aligned}\mathbf{E}L(X,Y_i) &= \frac{1+\Phi^2(Y_{i1})}{2} + \frac{1+\Phi^2(Y_{i2})}{2} - \\ &\quad - \frac{1+\Phi^2(Y_{i1})}{2}\frac{1+\Phi^2(Y_{i2})}{2}\end{aligned}$$

As

$$\int \left(\frac{1+\Phi^2(x)}{2}\right) d\Phi(x) = \frac{2}{3},$$

then

$$\mathbf{E}L(X,X') = \frac{8}{9}.$$

Remark. Another possible variant for testing H_0 can be obtained from using instead of standardization (2.54) Mahalanobis transformation:

$$Y_k = (X_k - \bar{X})\hat{S}^{-1/2}(X_k - \bar{X}),\ k = 1, ..., n.$$

Transformed sample will asymptotically have χ^2_p distribution with p degrees of freedom. The null hypothesis should be rejected for large values of statistic

$$T_n = -n \int_{R_+^{2p}} L(x,y)\, d(F_n(x) - G_{\chi^2_p}(x))\, d(F_n(y) - G_{\chi^2_p}), \tag{2.56}$$

where $F_n(x)$ is the empirical distribution function, constructed from the quantities $\{Y_i\}$ and $G_{\chi^2_p}$ is a χ^2_p distribution function with p degrees of freedom.

Koziol in (Koziol, 1983) investigates the properties of empirical process $\sqrt{n}(F_n(x) - G_{\chi^2_p}(x))$, which converges weakly to a Gaussian process with zero mathematical expectation and covariation function

$$K(s,t) = G_{\chi^2_p}(\min(s,t))) - G_{\chi^2_p}(s))G_{\chi^2_p}(t)) - \frac{2st}{p} g_{\chi^2_p}(t))g_{\chi^2_p}(s)),$$

where $g_{\chi^2_p}(t))$ is the density function of the distribution $G_{\chi^2_p}(t))$. The critical region of the test in this case can be also determined using Monte Carlo simulations.

2.3. Conclusions of Chapter 2

1. Based on N-distances, the construction of statistical tests of goodness of fit (simple and composite hypothesis) were proposed.
2. In the general case the limit null distribution of proposed N-metrics statistics coincides with the distribution of infinite quadratic form of Gaussian random variables. Under the alternative hypothesis, considered tests statistics are asymptotically normal.
3. In the general case proposed goodness of fit test statistics are not distribution-free.
4. For normality and nonparametric hypotheses of goodness of fit in high dimensional cases, when it is difficult from computational point of view to determine the limit null distribution of N-distance statistic analytically, the critical region of the test can be established by means of Monte Karlo simulations.

3

Nonparametric tests based on N-distances

3.1. Homogeneity test

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two samples of independent observations of random variables X and Y with unknown continuous distribution functions $F(x)$ and $G(x)$. The null hypothesis in the problem of testing homogeneity is $H_0 : F(x) = G(x)$.

The statistic for testing H_0 on the basis of N-distance between the empirical distributions constructed from corresponding samples has the form

$$T_{n,m} = -\frac{nm}{n+m} \int_{\mathbb{R}^{2p}} L(x,y)\, d(F_n(x) - G_m(x))\, d(F_n(y) - G_m(y)), \quad (3.1)$$

where $F_n(x)$, $G_m(x)$ - are empirical distribution functions based on the samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

In practice statistic (3.1) can be computed using the formula

$$T_{n,m} = \frac{2}{n+m}\sum_{i,j} L(X_i, Y_j) - \frac{m}{n(n+m)}\sum_{i,j} L(X_i, X_j) - \\ -\frac{n}{m(n+m)}\sum_{i,j} L(Y_i, Y_j). \tag{3.2}$$

We should reject the null hypothesis in case of large values of our test statistic. Following the outline of presentation determined in the previous sections, first consider the asymptotic distribution of statistic $T_{n,m}$ (3.1) in the most general case.

3.1.1. Asymptotic distribution of test statistic

Denote

$$H(x_1, y_1, x_2, y_2) = L(x_1, y_2) + L(x_2, y_1) - L(x_1, x_2) - L(y_1, y_2). \tag{3.3}$$

One can see that test statistic $T_{n,m}$ (3.2) can be rewritten in the form of two-sample U-statistic (Koroljuk and Borovskich, 1994)

$$T_{n,m} = \frac{nm}{n+m}\int H(x_1, y_1, x_2, y_2) dF_n(x_1) dG_m(y_1) dF_n(x_2) dG_m(y_2), \tag{3.4}$$

where $L(x, y)$ is the strongly negative definite kernel of N-distance and $x_1, x_2, y_1, y_2 \in \mathbb{R}^p$.
Note, that $H(x_1, y_1, x_2, y_2)$ satisfies the conditions $\forall x_1, x_2, y_1, y_2$:

- $H(x_1, y_1, x_2, y_2) = H(x_2, y_2, x_1, y_1)$ symmetry on $(x_1, y_1) \leftrightarrow (x_2, y_2)$,
- $H(x_1, y_1, x_2, y_2) = -H(y_1, x_1, x_2, y_2)$ anti-symmetry on $x_1 \leftrightarrow y_1$.

Under the null hypothesis, when X and Y has the same distribution function $F(x)$, the kernel of U-statistic (3.4) satisfies the property of degeneracy, that is

$$\mathbf{E}H(X, Y, x_2, y_2) = \mathbf{E}L(X, y_2) + \mathbf{E}L(x_2, Y) - \\ -\mathbf{E}L(X, x_2) - \mathbf{E}L(Y, y_2) = 0.$$

Let X' and Y' be independent copies of random variables X and Y respec-

tively. Assume that $\mathbf{E}H^2(X, Y, X', Y') < \infty$. Then, according to the spectral theorem there exist the orthogonal sequence of functions $\psi_1, \psi_2, ...$ in L_2, $\mathbf{E}\psi_j(X, Y) = 0$, $j \geq 1$, and the sequence of numbers $\lambda_1, \lambda_2, ...$ in $\mathbb{R}$,

$$\sum_{j=1}^{\infty} \lambda_j^2 = \mathbf{E}H^2(X, Y, X', Y') < \infty,$$

such that $\lim_{s\to\infty} \|H - H^s\|_{L_2}^2 = 0$ for

$$H^s(x_1, y_1, x_2, y_2) = \sum_{j=1}^{s} \lambda_j \psi_j(x_1, y_1)\psi_j(x_2, y_2).$$

Theorem 12. *Let* $\min(n, m) \to \infty$*, under* H_0 *and assumptions stated above the limit distribution of* $T_{n,m}$ *coincides with the distribution of random variable*

$$T = \sum_{j=1}^{\infty} \lambda_j \sigma_j^2 \zeta_j^2, \tag{3.5}$$

where

$$\sigma_j^2 = \int_{\mathbb{R}^p} (\mathbf{E}\psi_{\mathbf{j}}(\mathbf{x_1}, \mathbf{Y}))^2 dF(x_1), \quad j = 1, 2, ...,$$

and ζ_j*,* $j = 1, 2, ...$*, are independent standard normal random variables.*

Proof. From Theorem 5.6.1 in (Koroljuk and Borovskich, 1994) and properties of weak convergence of random processes it follows that the weak limit of $T_{n,m}$ coincides with the weak limit of random variable

$$U = \sum_{i=1}^{\infty} \lambda_i (a_i W_{1i}(1) + b_i W_{2i}(1))^2,$$

where $W_{1i}(\cdot)$, $W_{2i}(\cdot)$, $i = 1, 2, ...$, are independent Wiener processes on $[0, 1]$ and

$$a_i^2 = \beta \int_{\mathbb{R}^p} (\mathbf{E}\psi_i(x, Y))^2 dF(x),$$

$$b_i^2 = \alpha \int_{\mathbb{R}^p} (\mathbf{E}\psi_i(X, y))^2 dF(x),$$

where $0 < \alpha, \beta < 1$ and $m/(n + m) \to \alpha$, $n/(n + m) \to \beta$.
By virtue of the property of the anti-symmetry of $H(x_1, y_1, x_2, y_2)$,

$\psi_i(x_1, y_1)$ is also anti-symmetric. Thus

$$a_i^2 + b_i^2 = \sigma_i^2 = \int_{\mathbb{R}^p} (\mathbf{E}\psi_{\mathbf{i}}(\mathbf{x}, \mathbf{Y}))^2 dF(x)$$

and

$$[a_i W_{1i}(1) + b_i W_{2i}(1)] \sim N(0, \sigma_i^2).$$

Let us further consider the asymptotic distribution of test statistic $T_{n,m}$ (3.2) under alternative hypothesis. In this case the probability to reject the null hypothesis with a given size of the test α tends to 1 when $n, m \to \infty$. Therefore we consider our statistic $T_{n,m}$ normalized in a special way.

Let $F(x)$ does not equal identically to $G(x)$ and denote

$$H(x_1, y_1, x_2, y_2) := \frac{1}{2} H_0(x_1, y_1, x_2, y_2) - L(x_1, x_2) - L(y_1, y_2), \tag{3.6}$$

where

$$H_0(x_1, y_1, x_2, y_2) = L(x_1, y_2) + L(x_2, y_1) + L(x_1, y_1) + L(x_2, y_2)$$

Note, that $H(x_1, y_1, x_2, y_2)$ satisfies the property of symmetry by $x_1 \leftrightarrow x_2$ and $y_1 \leftrightarrow y_2$, so statistic $T_{n,m}$ can be represented in the form V-statistic

$$T_{n,m} = \int H(x_1, y_1, x_2, y_2) dF_n(x_1) dG_m(y_1) dF_n(x_2) dG_m(y_2). \tag{3.7}$$

Let X, $X^{'}$ and Y,$Y^{'}$ be independent random variables with probability distribution functions $F(x)$ and $G(x)$ respectively. Denote

$$a := \int H(x_1, y_1, x_2, y_2) dF(x_1) dG(y_1) dF(x_2) dG(y_2) \tag{3.8}$$

and define the functions

$$g_1(x) = \mathbf{E}(H(X, Y, X^{'}, Y^{'})|X = x) - a,$$

$$g_2(x) = \mathbf{E}(H(X, Y, X^{'}, Y^{'})|Y = x) - a.$$

Assume that $\sigma_1^2 = \mathbf{E}g_1^2(X)$ and $\sigma_2^2 = \mathbf{E}g_1^2(Y)$.

Theorem 13. *If* $\mathbf{E}H^2 < \infty$ *and* $\sigma_1^2 \neq 0$, $\sigma_2^2 \neq 0$, *then*

$$(4\sigma)^{-1}(T_{n,m} - a) \xrightarrow{d} \zeta,$$

as $min(n, m) \to \infty$, $\frac{n}{m} \to const. \neq 0$, *where* $\sigma^2 = \mathbf{E}(T_{n,m} - a)^2 = \frac{2}{n}\sigma_1^2 + \frac{2}{m}\sigma_2^2$ *and* ζ *is a random variable form the standard normal distribution.*

Proof. Under alternative hypothesis the kernel $H(x_1, y_1, x_2, y_2)$ of V-statistic (3.7) is nondegenerate.

Let us rewrite $T_{n,m}$ in the form

$$\begin{aligned} n^2m^2T_{n,m} &= 4\sum_{i<j,k<l} H(X_i, Y_k, X_j, Y_l) + 2\sum_{i<j,k} H(X_i, Y_k, X_j, Y_k) + \\ &+2\sum_{i,k<l} H(X_i, Y_k, X_i, Y_l) + \sum_{i,k} H(X_i, Y_k, X_i, Y_k) = \\ &= \xi_1 + \xi_2 + \xi_3 + \xi_4. \end{aligned}$$

Note, that for $i = 2, 3, 4$,

$$\frac{1}{n^2m^2}\sigma^{-1}\xi_i \xrightarrow{P} 0.$$

Thus the limit distribution of $T_{n,m}$ is defined by the asymptotic behavior of ξ_1.

Random variable ξ_1 corresponds to a two-sample U-statistic with the kernel $H(\cdot)$. The asymptotic normality of ξ_1 immediately follows from Theorem 4.5.1 in (Koroljuk and Borovskich, 1994) establishing the limit distribution of multi-sample U-statistics.

Under the null hypothesis the limit distribution of test statistic $T_{n,m}$ (3.2) depends on the common distribution F of random variables X and Y, which is unknown. Although one can estimate the null distribution in the case of a completely specified F, it is of much greater practical interest to develop a test procedure for the case where the only information available about the sampled populations is contained in the observed samples. In the next subsections we propose some ways for solving this problem. We suggest a bit different approaches in uni- and multivariate cases, so further they are considered separately.

3.1.2. Univariate case

Consider test statistic $T_{n,m}$ in univariate case

$$T_{n,m} = -\frac{nm}{n+m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x,y)\, d\Delta_{n,m}(x)\, d\Delta_{n,m}(y), \tag{3.9}$$

where $\Delta_{n,m}(x) = F_n(x) - G_m(x)$ and $F_n(x)$, $G_m(x)$ - are empirical distribution functions constructed from the samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$.

To avoid the dependence of the distribution of $T_{n,m}$ on the distribution of variables X and Y let us first transform initial samples to the samples $t_1, \ldots, t_n$ and $s_1, \ldots, s_m$, using the formulas

$$t_i = H_{n,m}(X_i), \quad i = 1, ..., n,$$

$$s_j = H_{n,m}(Y_j), \quad j = 1, ..., m,$$

where $H_{n,m}(x)$ is the empirical distribution function based on combined sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Under the null hypothesis the transformed samples will asymptotically have the uniform distribution on $[0, 1]$ and the statistic $T_{n,m}$ for testing the homogeneity of $t_1, \ldots, t_n$ and $s_1, \ldots, s_m$ will have for following form

$$T_{n,m} = -\frac{nm}{n+m} \int_0^1 \int_0^1 L(s,t)\, d\Delta^*_{n,m}(x)\, d\Delta^*_{n,m}(y), \tag{3.10}$$

where $\Delta^*_{n,m}(x) = F^*_n(x) - G^*_m(x)$ and $F^*_n(t)$ and $G^*_m(t)$ are the empirical distribution functions based on transformed samples $t_1, \ldots, t_n$ and $s_1, \ldots, s_m$ respectively.

The asymptotic distribution of $T_{n,m}$ can be found in the same way as it was done for goodness of fit tests (see Theorem 9) and brings to the following result.

Theorem 14. *Under the null hypothesis statistic $T_{n,m}$ will have the same asymptotic distribution as quadratic form*

$$T = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{a_{kj}}{\pi^2 kj} \zeta_k \zeta_j, \tag{3.11}$$

where ζ_k are independent random variables from the standard normal distribution and

$$a_{kj} = -2 \int_0^1 \int_0^1 L(s,t)\, d\sin(\pi k s)\, d\sin(\pi j t).$$

Proof. Denote

$$\xi_1(t) := \sqrt{n}(F^*_n(t) - t),$$

$$\xi_2(t) := \sqrt{n}(G_m^*(t) - t).$$

Empirical processes $\xi_1(t)$ and $\xi_2(t)$ converges jointly in distribution to independent Brownian bridges $W_1(t)$ and $W_2(t)$ (Billingsley, 1968; van der Vaart and Wellner, 1996). Note that,

$$\xi_{n,m}(t) := \sqrt{\frac{nm}{n+m}}(F_n^*(t) - G_m^*(t)) = \sqrt{\frac{m}{n+m}}\xi_1(t) - \sqrt{\frac{n}{n+m}}\xi_2(t).$$

If $n, m \to \infty$ such that $\frac{m}{n+m} \to \alpha \neq 0$ $\xi_{n,m}(t)$ weakly converges to random process $\sqrt{1-\alpha}W_1(t) - \sqrt{\alpha}W_2(t)$, which possesses the same distribution as Brownian bridge $W(t)$, $t \in [0, 1]$.

After establishing the weak limit of $\xi_{n,m}(t)$ the statement of the theorem follows directly form Lemma 2 and the proof of Theorem 9.

3.1.3. Multivariate case

An attempt to establish the asymptotic distribution of $T_{n,m}$ (3.1) in multivariate case

$$T_{n,m} = -\frac{nm}{n+m}\int_{\mathbf{R^{2p}}} L(x,y)\, d(F_n(x) - G_m(x))\, d(F_n(y) - G_m(y)), \quad (3.12)$$

leads to the same problem of its dependence on unknown distribution functions of X and Y.

Theorem 15. *Under the null hypothesis statistic $T_{n,m}$ converge in distribution to*

$$T = -\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} L(x,y) dW_F(x) dW_F(y), \quad (3.13)$$

where $W_F(x)$ is a Brownian bridge process corresponding to the distribution $F(x)$, that is a zero-mean Gaussian process with covariance function

$$K(x,y) = F(\min(x,y)) - F(x)F(y),$$

where $\min(x,y) = (\min(x_1,y_1), ..., \min(x_p,y_p)$.

Proof. The outline of the proof practically coincides with the proof of Theorem 2.2 in (Baringhaus and Franz, 2004).

To avoid complicated calculations, without loss of generality, consider the case $p = 1$.

Denote

$$\xi_{n,m}(x) := \sqrt{\frac{nm}{n+m}}(F_n(x) - G_m(x)).$$

Statistic $T_{n,m}$ can be rewritten in the form

$$T_{n,m} = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} L(x,y)d\xi_{n,m}(x)d\xi_{n,m}(y).$$

Under the null hypothesis, as $\min(n,m) \to \infty$, $\xi_{n,m}$ weakly converges to Brownian bridge process $W_F(x)$ corresponding to the distribution $F(x)$ (van der Vaart and Wellner, 1996) . If $L(x,y)$ satisfies the conditions of Lemma 2 on each finite square $[-C,C]^2$, the continuous mapping theorem applies to get that for each real $C > 0$

$$T_{n,m,C} = -\int_{-C}^{C}\int_{-C}^{C} L(x,y)d\xi_{n,m}(x)d\xi_{n,m}(y)$$

converge in distribution to

$$T_C = -\int_{-C}^{C}\int_{-C}^{C} L(x,y)dW_F(x)dW_F(y),$$

where the integral is considered as a Stieltjes integral after formal integration by parts using the formula in Lemma 2

$$\begin{aligned}
&-\int_{-C}^{C}\int_{-C}^{C} L(x,y)dW_F(x)dW_F(y) = \\
&= -\int_{-C}^{C}\int_{-C}^{C} W_F(x)W_F(y)\,dL(x,y) - \\
&-\int_{-C}^{C} W_F(C)W_F(y)\,dL(C,y) - \\
&-\int_{-C}^{C} W_F(x)W_F(C)\,dL(x,C) + \\
&+\int_{-C}^{C} W_F(x)W_F(-C)\,dL(x,-C) + \\
&+\int_{-C}^{C} W_F(-C)W_F(y)\,dL(-C,y)+
\end{aligned}$$

$$
\begin{aligned}
&+W_F(-C)W_F(-C)L(-C,-C) - \\
&-W_F(C)W_F(-C)L(C,-C) - \\
&-W_F(-C)W_F(C)L(-C,C) + \\
&+W_F(C)W_F(C)L(C,C).
\end{aligned}
$$

Clearly, T is the almost sure limit of T_C, as $C \to \infty$. Taking into account, that

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} L(x,y)\,dF(x)\,dF(y) < \infty$$

and

$$\mathbf{E}W_F(x)W_F(y) = F(\min(x,y)) - F(x)F(y) \to 0, \quad \min(x,y) \to \infty,$$

the mathematical expectation of T equals to

$$\mathbf{E}T = -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} F(\min(x,y)) - F(x)F(y)\,dL(x,y)$$

and is almost surely finite. Therefore, for each $\varepsilon > 0$ and $\delta > 0$, we can choose $C > 0$ such that

$$-\int_{\mathbb{R}\setminus[-C,C]^2} F(\min(x,y)) - F(x)F(y)\,dL(x,y) < \varepsilon\delta.$$

Since

$$\mathbf{E}\xi_{n,m}(x)\xi_{n,m}(y) = F(\min(x,y)) - F(x)F(y),$$

using Markov's inequality we obtain

$$
\begin{aligned}
\limsup_{n,m\to\infty} \quad & P(|T_{n,m} - T_{n,m,C}| > \varepsilon) \leq \\
& \leq -\frac{1}{\varepsilon}\int_{\mathbb{R}\setminus[-C,C]^2} F(\min(x,y)) - F(x)F(y)\,dL(x,y) < \delta.
\end{aligned}
$$

Applying Theorem 3.2 in (Billingsley, 1968) the statement of the theorem follows.

The way of transformation of initial samples to $[0,1]^p$, discussed in univariate case, can be rather complicated due to discrete nature of empirical distribution functions. To get the critical values in practice we suggest to use a permutation or a bootstrap approaches. These procedures amount to sampling without and with replacement, respectively, from the pooled data $(Z_{N,1}, ..., Z_{N,N}) =$

$(X_1, ..., X_n, Y_1, ..., Y_m)$, where $N = n + m$.

Denote by $H_{n,m}(x)$, $x \in \mathbb{R}^p$ the empirical distribution function constructed from the combined sample $Z_{N,1}, ..., Z_{N,N}$.

Permutation approach

Sampling without replacement from a pooled data $Z_{N,1}, ..., Z_{N,N}$ can be presented in terms of random permutation. Let $r = (r_1, ..., r_N)$ be a random vector with uniform distribution of the set of all permutations of
$\{1, 2, ..., N\}$ independent from X and Y. The two-sample permutation empirical distribution functions are

$$F_{n,N}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(Z_{N,r_i} \leq x\right), \tag{3.14}$$

$$G_{m,N}(x) = \frac{1}{m} \sum_{i=n+1}^{N} \mathbf{1}\left(Z_{N,r_i} \leq x\right), \tag{3.15}$$

where $x = (x_1, ..., x_p) \in \mathbb{R}^p$, $Z_{N,i} = (Z_{N,i}^{(1)}, ..., Z_{N,i}^{(p)})$ and

$$\mathbf{1}\left(Z_{N,i} \leq x\right) := \prod_{j=1}^{p} \mathbf{1}\left(Z_{N,i}^{(j)} \leq x_j\right).$$

Consider permutation empirical process

$$\xi_{n,m}^{(per)}(x) = \sqrt{\frac{nm}{n+m}}(F_{n,N}(x) - G_{m,N}(x)). \tag{3.16}$$

The proof that permutation approach works can be done in nearly the same way as in Theorem 15.

Assume that $\frac{m}{n+m} \to \gamma \in (0, 1)$, applying Theorems 3.7.1–2 in (van der Vaart and Wellner, 1996) we get that permutation process $\xi_{n,m}^{(per)}(x)$ converges in distribution to Brownian bridge process $W_H(x)$ corresponding to the distribution $H(x) = \gamma F(x) + (1-\gamma)G(x)$ given almost every sequence $X_1, X_2, ..., Y_1, Y_2, ...$ Thus, under the null hypothesis permutation statistic,

$$T_{n,m}^{(per)} := -\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} L(x, y) d\xi_{n,m}^{(per)}(x) d\xi_{n,m}^{(per)}(y)$$

converges in distribution to random variable T (3.13).

The distribution of random variable T coincides with the distribution of quadratic form of independent standard normal random variables (see Theorem 12). This implies that the distribution function of T is absolutely continuous and strictly increasing in $[0, +\infty)$.

Defining the upper α-quantile of $T_{n,m}^{(per)}$ by

$$c_{n,m}^{(per)} = \inf\{x : P(T_{n,m}^{(per)} > x) \leq \alpha\}$$

if follows that $c_{n,m}^{(per)}$ tends to the unique upper α-quantile of T almost surely as $n, m \to \infty$. Thus, the test rejecting the hypothesis if $T_{n,m} > c_{n,m}^{(per)}$ is asymptotically a test of level α for any $F = G$.

In practice the distribution of $T_{n,m}$ can be obtained by dividing the pooled data into two samples of sizes n and m by all possible C_{n+m}^{n} ways. Each time the value of statistic $T_{n,m}$ are calculated using the formula (3.1) and $c_{n,m}^{(per)}$ is established as an upper α-quantile of obtained empirical distribution.

Bootstrap approach

Instead of sampling without replacement from the pooled sample $Z_{N,1}, ..., Z_{N,N}$, we can sample with replacement. This leads to two-sample bootstrap empirical distribution functions

$$F_{n,N}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left(\hat{Z}_{N,i} \leq x\right), \tag{3.17}$$

$$G_{m,N}(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left(\hat{Z}_{N,n+i} \leq x\right), \tag{3.18}$$

where $\hat{Z}_{N,1}, ..., \hat{Z}_{N,N}$ is a sample randomly taken one by one from pooled sample $Z_{N,1}, ..., Z_{N,N}$ with replacement.

Consider bootstrap empirical process

$$\xi_{n,m}^{(bp)}(x) = \sqrt{\frac{nm}{n+m}}(F_{n,N}(x) - G_{m,N}(x)). \tag{3.19}$$

Using Theorems 3.7.6–7 in (van der Vaart and Wellner, 1996) the assertion on the limiting distribution of permutation process $\xi_{n,m}^{(per)}$ can be carried over to $\xi_{n,m}^{(bp)}$. That

implies that bootstrap statistic

$$T_{n,m}^{(bs)} := -\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} L(x,y)d\xi_{n,m}^{(bs)}(x)d\xi_{n,m}^{(bs)}(y)$$

weakly converges to random variable T (3.13). Consequently, the upper α-quantiles

$$c_{n,m}^{(bs)} = \inf\{x : P(T_{n,m}^{(bs)} > x) \leq \alpha\}$$

of the distribution $T_{n,m}^{(bs)}$ can be used as critical values for our test $T_{n,m} > c_{n,m}^{(bs)}$. The critical values set by bootstrap method possess exactly the same behavior as in permutation approach as $n, m \to \infty$. Thus, the test rejecting the hypothesis if $T_{n,m} > c_{n,m}^{(bs)}$ is asymptotically a test of level α. Because of computational difficulties, usually $c_{n,m}^{(bs)}$ will be approximated by the empirical upper α-quantile of independent observations of $T_{n,m}^{(bs)}$ obtained by Monte-Carlo simulations.

3.1.4. Distribution-free two-sample test

In this section a construction of multivariate distribution-free homogeneity test is proposed. An approach is based on the representation (1.3) of N-distance in terms of mathematical expectations of random variables X and Y

$$N(\mu_X, \nu_Y) = 2\mathbf{E}L(X,Y) - \mathbf{E}L(X,X') - \mathbf{E}L(Y,Y'), \tag{3.20}$$

where μ_X, ν_Y are probability distributions of independent random variables $X, X^{'}$ and $Y, Y^{'}$ respectively.

Let us randomly split each of two samples $X \sim X_1, \ldots, X_n$ and $Y \sim Y_1, \ldots, Y_m$ into two equal parts and consider each of the parts as a separate independent sample X, X' and Y, Y'. Then on the basis of these new samples calculate the corresponding sample quantities in the expression (3.20) of N-metrics. The test statistic for the hypothesis of homogeneity of two samples X and Y will have the form

$$T_{n,m} = \frac{2}{mn}\sum_{i,j} L(X_i, Y_j) - \frac{1}{n^2}\sum_{i,j} L(X_i, X_j{}') - \frac{1}{m^2}\sum_{i,j} L(Y_i, Y_j') \tag{3.21}$$

with the asymptotic behavior established by the theorem

Theorem 16. *Under the null hypothesis*

$$\frac{T_{n,m}}{\sqrt{\mathbf{D}T_{n,m}}} \xrightarrow{d} \zeta,$$

as $\min(n,m) \to \infty$, $\frac{n}{m} \to const. \neq 0$, *where* ζ *is a standard normal random variable.*

Proof. Let us first calculate the variance of statistic $T_{n,m}$ under the null hypothesis. If $X \stackrel{d}{=} Y$, then

$$\mathbf{E}T_{n,m} = 2\mathbf{E}L(X,Y) - \mathbf{E}L(X,X') - \mathbf{E}L(Y,Y') = 0.$$

After some simple calculations we have

$$\begin{aligned}\mathbf{Var}T_{n,m} &= A(n,m)\mathbf{E}L(X_1,Y_1) + B(n,m)\mathbf{E}(L(X_1,Y_1)L(X_1,Y_2)) + \\ &+C(n,m)\mathbf{E}(L(X_1,Y_1)L(X_1,Y_1)),\end{aligned} \tag{3.22}$$

where

$$A(n,m) = \frac{4}{nm} + \frac{1}{n^2} + \frac{1}{m^2} - \frac{10}{n} - \frac{10}{m},$$

$$B(n,m) = \frac{2m+2n-8}{mn},$$

$$C(n,m) = \frac{1}{n^2} + \frac{1}{m^2} + \frac{4}{mn},$$

Thus, in case $n, m \to \infty$ and $\frac{n}{m} \to \rho \neq 0$

$$\mathbf{Var}T_{n,m} = O(\frac{1}{n} + \frac{1}{m}).$$

To prove the asymptotic normality of $T_{n,m}$ let us show that all the cumulants Γ_k of the orders $k \geq 3$ of random variables $\frac{T_{n,m}}{\sqrt{\mathbf{D}T_{n,m}}}$ asymptotically converge to zero. The proof is based on a well known equality of the cumulant of the sum of random variables $\xi_1 + \ldots + \xi_n$

$$\Gamma_k(\xi_1 + \ldots + \xi_n) = \sum_{i_1,\ldots,j_k} \Gamma_k(\xi_{i_1}, \ldots, \xi_{i_k})$$

and the property for mixed cumulant $\Gamma_k(\xi_{i_1}, \ldots, \xi_{i_k})$ to be zero if the group of variables $\xi_{i_1}, \ldots, \xi_{i_k}$ can be divided into two independent parts.

After the application of the above mentioned properties and some easy calculations we have

$$\Gamma_k\left(\frac{T_{n,m}}{\sqrt{\mathbf{D}T_{n,m}}}\right) = c_{n,m}\sum_{i_1,\dots,i_k}\Gamma_k(\xi_{i_1},\dots,\xi_{i_k}), \tag{3.23}$$

where ξ_{i_j} denotes the random variables from the set

$$\{L(X_i,Y_j), L(X_i,X'_j), L(Y_i,Y'_j)\}, \quad i=1,\dots,n, \quad j=1,\dots,m,$$

$c_{n,m}$ depends only on n and m, and the following equality holds

$$c_{n,m} = O(n^\alpha m^\beta), \quad \alpha+\beta = -\frac{3}{2}k.$$

Since samples X, X', Y, Y' are independent, the number of nonzero mixed cumulants in the right part of (3.23) has the order $O(n^\gamma m^\nu)$, where $\gamma+\nu = k+1$. Thus, all the cumulants with $k \geq 3$ converge to zero and $\frac{T_{n,m}}{\sqrt{\mathbf{D}T_{n,m}}} \sim N(0,1)$ asymptotically, when $n, m \to \infty$.

The expression in (3.22) can be used for numerical estimation of dispersion of statistic $T_{n,m}$ with replacement of corresponding mathematical expectations by their sample estimates.

Of course, this method leads to essential loss of information, but is correct from the theoretical point of view, as it leads to a distribution-free criterion and allows testing the homogeneity hypothesis in high-dimensional cases.

3.2. Tests of uniformity on the hypersphere

In this section we propose an application of N-distance theory for testing the hypothesis of uniformity of spherical data. The proposed procedures have a number of advantages, consistency against all alternatives, computational simplicity and ease of application even in high-dimensional cases. Particular attention is devoted to $p = 2$ (circular data) and $p = 3$ (spherical data).

Consider the sample $X_1, ..., X_n$ of independent observations of random variable X, where $X_i \in \mathbb{R}^p$ and $\|X_i\| = 1$, $i = 1, ..., n$. Let us test the hypothesis H_0 that X has a uniform distribution on S^{p-1}.

The statistic for testing H_0 based on N-distance with the kernel $L(x, y)$ has the form

$$T_n = n\left[\frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\mathbf{Y}}L(X_i,Y) - \frac{1}{n^2}\sum_{i,j=1}^{n}L(X_i,X_j) - \mathbf{E}L(Y,Y')\right], \quad (3.24)$$

where X, Y, Y' are independent random variables from the uniform distribution on S^{p-1} and $\mathbf{E}_{\mathbf{Y}}L(X_i,Y) = \int L(X_i,y)dF_Y d(y)$ is a mathematical expectation calculated by Y with fixed X_i, $i = 1, ..., n$.

We should reject the null hypothesis in case of large values of our test statistic.

For our further research let us consider a strongly negative definite kernel $L(x,y) = L(\|x-y\|)$, where $\|\cdot\|$ is the Euclidean norm. In other words, $L(x,y)$ depends on the length of the chord between two points on hypersphere. As an example of such kernels we propose the following ones

$$L(x,y) = \|x-y\|^{\alpha}, \quad 0 < \alpha < 2,$$

$$L(x,y) = \frac{\|x-y\|}{1+\|x-y\|},$$

$$L(x,y) = log(1+\|x-y\|^2).$$

Note, that considered kernels are rotation-invariant. This property implies that the mathematical expectation of the length of the chord between two independent uniformly distributed random variables Y and Y' on S^{p-1} is equal to the mean length of the chord between a fixed point and a uniformly distributed random variable Y on S^{p-1}. Thus, we can rewrite (3.24) in the form

$$T_n = n\left[\mathbf{E}L(\|Y-Y'\|) - \frac{1}{n^2}\sum_{i,j=1}^{n}L(\|X_i-X_j\|)\right]. \quad (3.25)$$

In practice statistic T_n with the kernel $L(x,y) = \|x-y\|^{\alpha}$, $0 < \alpha < 2$ can be calculated using the proposition.

Proposition 9. *In cases of $p = 2, 3$ statistics T_n will have the form*

$$T_n = \frac{(2R)^{\alpha}\Gamma(\frac{\alpha+1}{2})\Gamma(\frac{1}{2})}{\pi\Gamma(\frac{\alpha+2}{2})}n - \frac{1}{n}\sum_{i,j=1}^{n}\|X_i-X_j\|^{\alpha} \quad (p=2),$$

$$T_n = (2R)^\alpha \frac{2n}{\alpha+2} - \frac{1}{n}\sum_{i,j=1}^{n} \|X_i - X_j\|^\alpha \quad (p=3),$$

where R is the radius of hypersphere and $\alpha \in (0,2)$.

Proof. The stated above formulas follows directly from (3.25) and the property

$$\mathbf{E}\|Y - Y'\|^\alpha = \mathbf{E}\|Y - a\|^\alpha,$$

where Y, Y' are independent random variables from the uniform distribution on S^{p-1} and a is a fixed point on S^{p-1}.

In two-dimensional case, let us calculate the mathematical expectation of the length of the chord between fixed point $a = (0,R)$ and a uniformly distributed random variable Y

$$\begin{aligned}
&\mathbf{E}\|a-Y\|^\alpha = \frac{1}{2\pi R}\int_0^{2\pi} R(R^2\cos^2\phi + (R\sin^2\phi - R)^2)^{\frac{\alpha}{2}}d\phi = \\
&= \frac{2^{\frac{\alpha}{2}-1}R^\alpha}{\pi}\int_0^{2\pi}(1-\cos\phi)^{\frac{\alpha}{2}}d\phi = \frac{2^{\alpha+1}R^\alpha}{\pi}\int_0^{\frac{\pi}{2}}\sin^\alpha\phi d\phi = \\
&= \frac{(2R)^\alpha\Gamma(\frac{\alpha+1}{2})\Gamma(\frac{1}{2})}{\pi\Gamma(\frac{\alpha+2}{2})}.
\end{aligned}$$

In case $p=3$ let us fix point $a=(0,0,R)$ and calculate the average length of the chord

$$\begin{aligned}
&\mathbf{E}\|a-Y\|^\alpha = \frac{1}{4\pi R^2}\int_{-\pi}^{\pi}\int_0^{\pi} R^2\sin\theta(R^2(\sin^2\theta\cos^2\phi + \\
&+\sin^2\theta\sin^2\phi + (\cos\theta-1)^2))^{\frac{\alpha}{2}}d\theta d\phi = \\
&= \frac{2^{\frac{\alpha}{2}}R^\alpha}{4\pi}\int_{-\pi}^{\pi}\int_0^{\pi}(1-\cos\theta)^{\frac{\alpha}{2}}\sin\theta d\theta d\phi = \\
&= 2^{\alpha+1}R^\alpha\int_0^{\frac{\pi}{2}}\sin^{\alpha+1}\theta d\sin\theta = (2R)^\alpha\frac{2}{\alpha+2}.
\end{aligned}$$

In case of $L(x,y) = \|x-y\|$, test statistic (3.25) is very similar to Ajne's

statistic A, where instead of chord is taken the length of the smaller arc.

$$A = \frac{n}{4} - \frac{1}{\pi n} \sum_{i,j=1}^{n} \psi_{ij},$$

where ψ_{ij} is the smaller of two angles between X_i and X_j.

One can see, that Ajne's test is not consistent against all alternatives, as an example consider the distribution on the circle concentrated in two diametrically opposite points with equal probabilities. Taking instead of arc the length of the chord lead to a consistency of the N-distance test against all fixed alternatives.

$$\frac{T_n}{n} \xrightarrow{P} N(X,Y), \quad n \to \infty,$$

where $N(X,Y)$ is N-distance (1.3) between probability distributions of random variables X and Y. If $X \neq^d Y$, then $N(X,Y) > 0$ and $T_n \to \infty$, as $n \to \infty$.

Further we consider the asymptotic distribution of statistic T_n (3.24) under the null hypothesis. Particular attention is devoted to circular and spherical data (p=2,3). In these cases the asymptotic behavior of proposed tests under the null hypothesis is established using two approaches. First is based on an adaptation of methods of uniformity tests described in section 2.1.1, and second using Gine theory based on Sobolev norms (Gine, 1975; Hermans and Rasson, 1985). For arbitrary dimension it is rather difficult from the computational point of view to establish the distribution of test statistic T_n analytically, in this case the critical region of our criteria can be determined with the help of simulations of independent samples from the uniform distribution on S^{p-1}.

3.2.1. Asymptotic distribution of test statistic

Uniformity on the circle S^1

For our further research, without loss of generality, we consider the circle S^1 with unit length, that is $R = \frac{1}{2\pi}$. Let us transform our circle, and therefore our initial sample $X_1, ..., X_n$, $X_i = (X_{i1}, X_{i2})$, $X_{i1}^2 + X_{i2}^2 = R^2$ to the interval $[0, 1)$ by making a cut in arbitrary point of the circle

$$x \leftrightarrow x^*, \quad x \in S^1,\ x^* \in [0, 1).$$

It is easy to see, that if X has a uniform distribution on S^1, after described transformation we will get the random variable X^* with uniform distribution on

$[0,1)$.

Let $L(x,y)$ be a strongly negative definite kernel in $\mathbb{R}^2$, then function $H(x^*,y^*)$ on $[0,1)$

$$H(x^*,y^*) := L(x,y) \tag{3.26}$$

is a strongly negative definite kernel on $[0,1)$. In this case N-distance statistic T_n^*, based on $H(x^*,y^*)$, for testing the uniformity on $[0,1)$ has the form

$$T_n^* = -n\int_0^1\int_0^1 H(x^*,y^*)d(F_n(x^*)-x^*)d(F(y^*)-y),$$

where $F_n(x^*)$ is the empirical distribution function, based on the sample $X_1^*,...,X_n^*$, $X_i^* \in [0,1)$, $i=1,...,n$.

Due to (3.26) the following equality holds

$$T_n = T_n^*, \tag{3.27}$$

where T_n is defined by (3.24).

Thus, instead of testing the initial hypothesis on S^1 using T_n, we can test the uniformity on $[0,1)$ for X^* on the basis of statistic T_n^* with the same asymptotic distribution. The limit distribution of T_n^* is established in Theorem 9 and leads to the result.

Theorem 17. *Under the null hypothesis statistic T_n will have the same asymptotic distribution as quadratic form:*

$$T = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\frac{a_{kj}}{\pi^2 kj}\zeta_k\zeta_j, \tag{3.28}$$

where ζ_k are independent random variables from the standard normal distribution and

$$a_{kj} = -2\int_0^1\int_0^1 H(x^*,y^*)\,d\sin(\pi k x^*)\,d\sin(\pi j y^*).$$

It is easy to see, that in case $L(x,y)$ is a rotation-invariant function on the circle the considered transformation of S^1 to $[0,1)$ does not depend on the choice of point of cut.

Proposition 10. *If strongly negative definite kernel $L(x,y) = \|x-y\|^\alpha$, where*

$0 < \alpha < 2$, $x, y \in S^2$, *then*

$$H(x^*, y^*) = \left[\frac{\sin \pi d}{\pi}\right]^{\alpha},$$

where

$$d = \min(|x^* - y^*|, 1 - |x^* - y^*|), \quad x^*, y^* \in [0, 1).$$

Proof. Kernel $L(x, y)$ equals to the length of the chord in the circle between two points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in α power. After proposed transformation, the length of the smaller arc between x and y equals to $d = \min(|x^* - y^*|, 1 - |x^* - y^*|)$. And the length of the chord in the circle with $R = \frac{1}{2\pi}$ based on the angle $2\pi d$ equals to $\frac{\sin \pi d}{\pi}$.

Uniformity on the sphere S^2

In case of a sphere we also try to substitute the initial hypothesis of uniformity on S^2 by testing the uniformity on the unit square. Consider sphere S^2 with unit surface area, that is $R^2 = \frac{1}{4\pi}$.

Note, that if $X^* = (X_1^*, X_2^*)$ has the uniform distribution on $[0, 1)^2$ then random variable $X = (X_1, X_2, X_3)$:

$$X_1 = R\cos\theta_1, \;\; X_2 = R\sin\theta_1\cos\theta_0, \;\; X_3 = R\sin\theta_1\sin\theta_0, \tag{3.29}$$

where

$$\theta_0 = 2\pi X_1^*, \quad \theta_1 = \arccos(1 - 2X_2^*)$$

has the uniform distribution on S^2.

Consider the strongly negative definite kernel $H(x^*, y^*)$ on $[0, 1)^2$ defined by

$$H(x^*, y^*) := L(x, y), \tag{3.30}$$

where $L(x, y)$ is a strongly negative definite kernel in $\mathbb{R}^3$, $x^*, y^* \in [0, 1)^2$, $x, y \in S^3$ and the correspondence between x and x^* follows from (3.29).

N-distance statistic, based on $H(x^*, y^*)$, for testing the uniformity on $[0, 1)^2$

$$T_n^* = -n \int_{[0,1)^2} \int_{[0,1)^2} H(x^*, y^*) d(F_n(x^*) - x_1^* x_2^*) d(F(y) - y_1^* y_2^*),$$

where $F_n(x^*)$, $x^* \in \mathbb{R}^2$ is the empirical distribution function based on the transformed sample X^*.

The equations (3.29) and (3.30) implies that:

$$T_n = T_n^*. \tag{3.31}$$

Thus, the asymptotic distribution of T_n coincides with the limit distribution of T_n^*, established in Theorem 10.

Theorem 18. *Under the null hypothesis statistic T_n will have the same asymptotic distribution as quadratic form*

$$T = \sum_{i,j,k,l=1}^{\infty} a_{ijkl}\sqrt{\alpha_{ij}\alpha_{kl}}\zeta_{ij}\zeta_{kl}, \tag{3.32}$$

where ζ_{ij} - independent random variables from the standard normal distribution,

$$a_{ijkl} = -\int_{[0,1]^4} H(x,y)\, d\psi_{ij}(x)\, d\psi_{kl}(y), \quad x, y \in \mathbb{R}^2,$$

α_{ij} and $\psi_{ij}(x,y)$ are eigenvalues and eigenfunctions of the integral operator A

$$Af(x) = \int_{[0,1]} K(x,y)f(y)dy \tag{3.33}$$

with the kernel

$$K(x,y) = \prod_{i=1}^{2} \min(x_i, y_i) - \prod_{i=1}^{2} x_i y_i.$$

Note, that if $L(x,y)$ is a rotation-invariant function on the sphere then the values of statistic T_n and T_n^* does not depend on the choice of coordinate system on S^2.

The main difficulties in application of the Theorem 18 are connected with calculations of eigenfunctions of the integral operator A. One of the possible solutions of these problems is in detail discussed in section 2.1.3 and is based on numerical approximation of eigenfunctions by a finite sequence (2.27). Another approach is considered in the next subsection, where the asymptotic distribution of proposed statistics for some strongly negative definite kernels is established with the help of Gine theory based on Sobolev tests.

Alternative approach to limit distribution of T_n

In this section we propose an application of Gine theory of Sobolev invariant tests for uniformity on compact Riemannian manifolds to establishing the null limit distribution of some N-distance statistics on the circle and sphere. We start from a brief review of Sobolev tests, for more details see (Gine, 1975; Jupp, 2005).

Let M be a compact Riemannian manifold. The Riemannian metric determines the uniform probability measure μ on M. The intuitive idea of the Sobolev tests of uniformity is to map the manifold M into the Hilbert space $L_2(M, \mu)$ of square-integrable functions on M by a function $t : M \to L_2(M, \mu)$ such that, if X is uniformly distributed, then the mean of $t(X)$ is 0.

The standard way of constructing such mappings t is based on the eigenfunctions of the Laplacian operator on M. For $k \geq 1$, let E_k denote the space of eigenfunctions corresponding to the kth eigenvalue, and put $d(k) = dim E_k$. Then there is a map t_k of M into E_k given by

$$t_k(x) = \sum_{i=1}^{d(k)} f_i(x) f_i,$$

where $f_i, 1 \leq i \leq d(k)$, is any orthonormal basis of E_k. If $a_1, a_2, ...$ is a sequence of real numbers such that

$$\sum_{i=1}^{\infty} a_k^2 d(k) < \infty,$$

then

$$x \mapsto t(x) = \sum_{i=1}^{\infty} a_k t_k(x)$$

defines a mapping t of M into $L_2(M, \mu)$. The resulting Sobolev statistic evaluated on observations $X_1, ..., X_N$ on M is

$$S_n(\{a_k\}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \langle t(X_i), t(X_j) \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on $L_2(M, \mu)$.

The asymptotic null distribution of statistic S_n is established by the following theorem (see Theorem 3.4 in (Gine, 1975)).

Theorem 19. *Let $X_1, .., X_n$ be a sequence of independent random variables with uniform distribution on M, then*

$$S_n(\{a_k\}) \xrightarrow{d} \sum_{k=1}^{\infty} a_k^2 \chi_k,$$

where $\{\chi_k\}_{k=1}^{\infty}$ is a sequence of independent random variables, such that, for each k, χ_k has a chi-square distribution with $d(k)$ degrees of freedom.

Further consider N-distance and Sobolev tests for two special cases of the circle and the sphere.

Let M be the circle $x_1^2 + x_2^2 = 1$ in $\mathbb{R}^2$. Gine showed that in this case Sobolev tests $S_n(\{a_k\})$ has the form

$$S_n(\{a_k\}) = 2n^{-1} \sum_{k=1}^{\infty} a_k^2 \sum_{i,j=1}^{n} \cos k(X_i - X_j) \tag{3.34}$$

with the limit null distribution established by Theorem 19, where χ_k are independent random variables with chi-square distribution with $d(k) = 2$ degrees of freedom.

Consider statistic T_n on M with strongly negative definite kernel $L(x, y) = \|x - y\|$, $x, y \in \mathbb{R}^2$. From proposition 9 we have

$$T_n = \frac{4n}{\pi} - \frac{1}{n} \sum_{i,j=1}^{n} \|X_i - X_j\| = \frac{4n}{\pi} - \frac{2}{n} \sum_{i,j=1}^{n} \sin \frac{X_i - X_j}{2}, \tag{3.35}$$

where $X_i - X_j$ and $\|X_i - X_j\|$ denotes the length of the arc and chord between X_i and X_j respectively.

Under the null hypothesis the limit distribution of T_n is established by the theorem

Theorem 20. *If $X_1, ..., X_n$ is a sample of independent observations from the uniform distribution on the circle with unit radius, then*

$$\frac{\pi}{4} T_n \xrightarrow{d} \sum_{k=1}^{\infty} a_k^2 \chi_k^2, \tag{3.36}$$

where χ_k^2 are independent random variables with chi-square distribution with two

degrees of freedom and

$$a_k^2 = \frac{1}{2\pi}\int_0^{2\pi}(1-\frac{\pi}{2}\sin\frac{x}{2})\cos kxdx.$$

Proof. Let us express statistic T_n (3.35) in the form

$$T_n = \frac{4}{\pi}n^{-1}\sum_{i,j=1}^{n} h(X_i - X_j),$$

where $h(x) = 1 - \frac{\pi}{2}\sin\frac{x}{2}$.
Function $h(x)$ can be represented in the form of a series by complete orthonormal sequence of functions $\{\sqrt{2}\cos kx\}$ on $[0, 2\pi]$

$$h(x) = \sqrt{2}\sum_{k=1}^{\infty}\alpha_k\cos kx,$$

where $\alpha_k = \frac{\sqrt{2}}{2\pi}\int_0^{2\pi}(1-\frac{\pi}{2}\sin\frac{x}{2})\cos kxdx$. Note, that $\alpha_k > 0$, $\forall k = 1, 2, ...$, really after some simple calculations we have

$$\int_0^{2\pi}(1-\frac{\pi}{2}\sin\frac{x}{2})\cos kxdx = 4\int_0^{\pi}\sin x\sin^2 kxdx - 4,$$

$$\int_0^{\pi}\sin x\sin^2 kxdx = -k^2\int_0^{\pi k}\sin(\frac{1}{k}-2)xdx -$$
$$-\frac{k^2}{2k+1}\int_0^{\pi k}\sin\frac{x}{k}dx = \frac{4k^3}{(2k-1)(2k+1)} > 1 \ \ \forall k = 1, 2, ...$$

Thus statistic T_n can be rewritten in the form of Sobolev statistic (3.34)

$$\frac{4}{\pi}T_n = 2n^{-1}\sum_{k=1}^{\infty}a_k^2\sum_{i,j=1}^{n}\cos k(X_i - X_j),$$

where $\sqrt{2}a_k^2 = \alpha_k$. After that the statement of the theorem follows directly from Theorem 19.

A comparison of empirical distribution function of statistic $\frac{4}{\pi}T_n$ and the dis-

tribution function of random variable (3.36) is shown in Fig. 3.1. The empirical distribution of $\frac{4}{\pi}T_n$ was calculated by simulation of 600 samples of size 100 form the uniform distribution on the circle S^1.

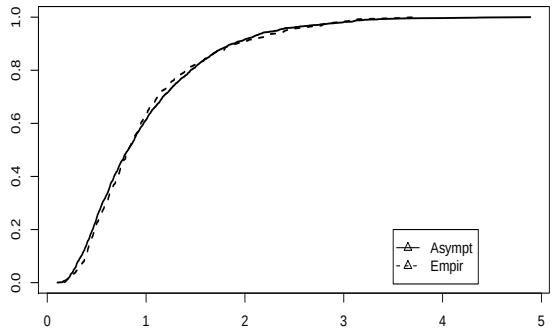

Figure 3.1. Empirical and asymptotic distribution of statistic $\frac{4}{\pi}T_n$ with the kernel $L(x,y) = \|x-y\|$, $x, y \in S^1$, $n = 100$.

We now pass over to N-distance and Sobolev tests on the sphere. If $M = S^2$ is the unit sphere $x_1^2 + x_2^2 + x_3^2 = 1$ on $\mathbb{R}^3$, then $d\mu = (4\pi)^{-1}\sin\theta d\theta d\phi$, where μ is the uniform distribution on S^2 and (θ, ϕ) are usual spherical coordinates. The general expression of Sobolev test statistics on the sphere has the form

$$S_n(\{a_k\}) = n^{-1}\sum_{k=1}^{\infty}(2k+1)a_k^2\sum_{i,j=1}^{n}P_k(\cos\widehat{X_i, X_j}), \tag{3.37}$$

where $\widehat{X_i, X_j}$ is an smaller angle between X_i and X_j, P_k are Legendre polynomials

$$P_k(x) = (k!2^k)^{-1}(d^k/dx^k)(x^2-1)^k.$$

Under the null hypothesis the limit distribution of $S_n(\{a_k\})$ coincides with the distribution of random variable

$$\sum_{k=1}^{\infty}a_k^2\chi_{2k+1}^2, \tag{3.38}$$

where χ^2_{2k+1} are independent random variables with chi-square distribution with $2k+1$ degrees of freedom.

Consider statistic T_n on S^2 with strongly negative definite kernel $L(x,y) = \|x-y\|$, $x, y \in \mathbb{R}^3$. From proposition 9 we have

$$T_n = \frac{4n}{3} - \frac{1}{n}\sum_{i,j=1}^{n} \|X_i - X_j\| = \frac{4n}{3} - \frac{2}{n}\sum_{i,j=1}^{n} \sin\frac{\widehat{X_i, X_j}}{2}, \tag{3.39}$$

where $\widehat{X_i, X_j}$ and $\|X_i - X_j\|$ denotes the smaller angle and chord between X_i and X_j respectively.

The asymptotic distribution of T_n is established by the next theorem.

Theorem 21. *If $X_1, ..., X_n$ is a sample of independent observations from the uniform distribution on S^2, then*

$$\frac{3}{4}T_n \xrightarrow{d} \sum_{k=1}^{\infty} a_k^2 \chi^2_{2k+1}, \tag{3.40}$$

where χ^2_{2k+1} are independent random variables with chi-square distribution with $2k+1$ degrees of freedom and

$$a_k^2 = \frac{1}{2}\int_0^{\pi} (1 - \frac{3}{2}\sin\frac{x}{2}) \sin x P_k(\cos x) dx, \tag{3.41}$$

where $P_k(x)$ are Legendre polynomials.

Proof. The proof of the theorem can be done in nearly the same way as that of Theorem 20. Let us first rewrite statistic T_n in the form

$$T_n = \frac{4}{3} n^{-1} \sum_{i,j=1}^{n} h(\widehat{X_i, X_j}),$$

where $h(x) = 1 - \frac{3}{2}\sin\frac{x}{2}$. And then decompose $h(x)$ to the series by orthonormal sequence of functions $\{\sqrt{2k+1}P_k(\cos x)\}$ for $x \in [0, \pi]$

$$h(x) = \sum_{k=1}^{\infty} \sqrt{2k+1}\alpha_k P_k(\cos x),$$

where

$$\alpha_k = \frac{\sqrt{2k+1}}{4\pi} \int_0^{2\pi} \int_0^{\pi} (1 - \frac{3}{2} \sin \frac{\theta}{2}) \sin \theta P_k(\cos \theta) d\theta d\phi.$$

As a result statistic T_n can be expressed in the form of Sobolev statistic (3.37)

$$\frac{4}{3} T_n = n^{-1} \sum_{k=1}^{\infty} (2k+1) a_k^2 \sum_{i,j=1}^{n} P_k(\cos \widehat{X_i, X_j}),$$

where $\sqrt{2k+1} a_k^2 = \alpha_k$. Applying Theorem 19 the assertion of the theorem follows.

The inverse values to the largest coefficients a_k^2 (3.41) are presented below.

5	35	105	231	429
715	1105	1615	2261	3059
4025	5175	6525	8091	9889

A comparison of empirical distribution function of statistic $\frac{3}{4} T_n$ and the distribution function of random variable (3.40) is shown in Fig. 3.2. The empirical distribution of $\frac{3}{4} T_n$ was calculated by simulation of 600 samples of size 100 form the uniform distribution on the sphere S^2.

3.3. Symmetry and independence tests

In this section we consider an application of N-distance statistics for testing the hypothesis of symmetry about zero in univariate case and independence in bivariate case. Under the null hypothesis the asymptotic distribution of proposed statistics is established and coincides with the distribution of infinite quadratic form of standard normal random variables. In essence the method of obtaining the limit distribution of test statistics is much the same considered in sections 2.1.2, 2.2.1, therefore here in the proofs of theorems we avoid some details and refer the reader to the mentioned sections.

3.3.1. Symmetry test

Let $X_1, \ldots, X_n$ be a sample of independent observations of random variable X with continuous distribution function $F(x)$, $x \in \mathbb{R}$, assumed unknown. Let us

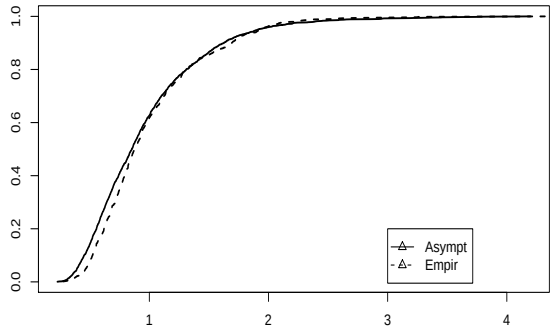

Figure 3.2. Empirical and asymptotic distribution of statistic $\frac{3}{4}T_n$ with the kernel $L(x,y) = \|x-y\|$, $x, y \in S^2$, $n = 100$.

test the hypothesis, that X has a symmetric distribution with respect to $x = 0$, that is $H_0 : X =^d -X$ or in terms of distribution functions $H_0 : F(x) = 1 - F(-x)$.

As a statistic, based on N-metrics, for testing this hypothesis we propose

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x,y)\, d\Delta_n(x)\, d\Delta_n(y), \tag{3.42}$$

where $\Delta_n(x) = F_n(x) + F_n(-x) - 1$, $F_n(x)$ - is empirical distribution function, constructed from the sample $X_1, \ldots, X_n$.

The asymptotic distribution of test statistic T_n depends on unknown distribution function $F(x)$. To avoid this let us transform our sample $X_1, \ldots, X_n$ to the sample $t_1, \ldots, t_n$, where $t_i = 1 - F_n(-X_i)$. Under the null hypothesis the transformed sample will asymptotically have the uniform distribution on $[0, 1]$ and the statistic T_n for testing the uniformity of $t_1, \ldots, t_n$ will have the form

$$T_n = -n \int_0^1 \int_0^1 L(t,s)\, d(F_n^*(t) - t)\, d(F_n^*(s) - s), \tag{3.43}$$

where $F_n^*(t)$ is the empirical distribution function, based on the sample $t_1, \ldots, t_n$.

In practice statistics (3.43) for different strongly negative definite kernels $L(t,s)$ can be calculated using the formulas in proposition 3.

Let us further consider the asymptotic distribution of T_n, which helps us to determine the critical region of our test. In accordance with symmetry test discussed in (Martynov, 1978), the limit distribution of T_n coincide with the distribution of random variable

$$\xi = -\int_0^1 \int_0^1 L(t,s)\, d(\zeta(t)+\zeta(1-t))\, d(\zeta(s)+\zeta(1-s)), \tag{3.44}$$

where $\zeta(t)$ is a gaussian random process with zero mean and correlation function $K(x,y) = \min(x,y) - xy$.

Taking into account the symmetry of $\zeta(t)+\zeta(1-t)$ about $t = \frac{1}{2}$ and symmetry by arguments of $L(x,y)$, the random variable ξ can be expressed in the form

$$\xi = -\int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} H(t,s)\, d(\zeta(t)+\zeta(1-t))\, d(\zeta(s)+\zeta(1-s)), \tag{3.45}$$

where $H(t,s) = L(t,s) + L(1-t,1-s) - L(t,1-s) - L(1-t,s)$.
Note, that when $\max(t,s) < \frac{1}{2}$, the correlation function of the random process $\zeta(t)+\zeta(1-t)$ is equal to $K(t,s) = 2\min(t,s)$. The eigenvalues λ_k and functions $\psi_{\lambda_k}(t)$ of integral operator A with the kernel $K(t,s)$

$$Af(t) = \int_0^1 K(t,s) f(s) ds$$

are equal to

$$\lambda_k = 2(\pi(k-\frac{1}{2}))^{-2},$$

$$\psi_{\lambda_k}(t) = \sqrt{2}\sin(\pi(k-\frac{1}{2})t).$$

Consequently, the process $\zeta(t)+\zeta(1-t)$ with probability 1 can be presented in the form:

$$\zeta(t)+\zeta(1-t) = \sum_{k=1}^{\infty} \zeta_k \psi_{\lambda_k}(t), \tag{3.46}$$

where ζ_k are independent random variables from the Gaussian distribution with mean zero and variance $2(\pi(k-\frac{1}{2}))^{-2}$.

As a result, the asymptotic distribution of T_n is summarized in the following theorem:

Theorem 22. *Under the null hypothesis statistic T_n will have the same asymptotic*

distribution as quadratic form

$$T = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} a_{kj} \sqrt{\lambda_k \lambda_j} \zeta_k \zeta_j, \tag{3.47}$$

where ζ_k are independent random variables from the standard normal distribution, $\lambda_k = 2(\pi(k - \frac{1}{2}))^{-2}$ and

$$a_{kj} = -2 \int_0^{\frac{1}{2}} \int_0^{\frac{1}{2}} H(t,s)\, d\sin(\pi(k - \frac{1}{2})t)\, d\sin(\pi(j - \frac{1}{2})s),$$

where

$$H(t,s) = L(t,s) + L(1-t, 1-s) - L(t, 1-s) - L(1-t, s).$$

3.3.2. Independence test

Let $X_1, \ldots, X_n$, $X_i = (X_{i1}, X_{i2})$ be the sample of independent observations of random vector X with unknown continuous distribution function $F(x)$, $x \in \mathbb{R}^2$. Consider the hypothesis of independence of the coordinates of X, which can be expressed as follows

$$H_0 : F(x_1, x_2) = F_1(x_1) F_2(x_2),$$

where $F_i(x_i), i = 1, 2$ - continuous univariate distribution functions.

N-distance statistics T_n for testing H_0 in this case have the form

$$T_n = -n \int_{R^4} L(x,y) d\Delta_n(x)\, d\Delta_n(y),$$

where $x, y \in \mathbb{R}^2$, $\Delta_n(x) = F_n(x) - F_{n1}(x_1) F_{n2}(x_2)$,
$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_{i1} < x_1) I(X_{i2} < x_2)$ is a bivariate empirical distribution function and $F_{ni}(x_i), i = 1, 2$ are univariate empirical distribution functions, based on the i-th coordinate of the sample.

Following the procedure used for symmetry tests, first transform our sample $X_1, \ldots, X_n$ to the sample $Y_1, \ldots, Y_n$, using the formula

$$Y_i = (Y_{i1}, Y_{i2}) = (F_{n1}(X_{i1}), F_{n2}(X_{i2})).$$

Under H_0 the transformed sample will asymptotically have the uniform distribution

on the unit square and the statistic for testing uniformity of $Y_1, \dots, Y_n$ will have the form

$$T_n = -n \int_{[0,1]^4} L(t,s) d(F_n^*(t) - t_1 t_2) d(F_n^*(s) - s_1 s_2), \tag{3.48}$$

where $t = (t_1, t_2)$, $s = (s_1, s_2)$ and $F_n^*(t)$ is empirical distribution function constructed from transformed sample $Y_1, \dots, Y_n$.

In practice statistics (3.48) for different strongly negative definite kernels $L(t,s)$ can be calculated using the formulas in propositions 4–7.

Pass over to determination of the null asymptotic distribution of T_n. If $n \to \infty$, the distribution of T_n coincides with the distribution of random variable

$$\xi = -\int_{[0,1]^4} L(s,t) d(\zeta(s)) d(\zeta(t)),$$

where $\zeta(t), t \in \mathbb{R}^2$ is a gaussian random process with mean zero and correlation function

$$K(t,s) = (\min(t_1, s_1) - t_1 s_1)(\min(t_2, s_2) - t_2 s_2).$$

After calculation of the eigenvalues and functions of corresponding integral operator with the kernel $K(t,s)$, the random process $\zeta(t), t \in \mathbb{R}^2$, with probability 1 can be presented in the form

$$\zeta(t) = \sum_{ij=1}^{\infty} \zeta_{ij} \varphi_{ij}(t),$$

where $\varphi_{ij}(t) = 2 \sin(\pi i t_1) \sin(\pi j t_2)$ and ζ_{ij} are independent random variables from the normal distribution with mean zero and variance $(\pi^2 ij)^{-2}$.

Finally, the following theorem determines the asymptotic distribution of T_n.

Theorem 23. *Under the null hypothesis statistic T_n will have the same asymptotic distribution as quadratic form*

$$T = \sum_{ijkl=1}^{\infty} a_{ijkl} \sqrt{\lambda_{ij} \lambda_{kl}} \zeta_{ij} \zeta_{kl}, \tag{3.49}$$

$$a_{ijkl} = -\int_{[0,1]^4} L(t,s)\, d\varphi_{ij}(t)\, d\varphi_{kl}(s),$$

where ζ_{ij} are independent random variables from the standard normal distribution, $\lambda_{ij} = (\pi^2 ij)^{-2}$ and $\varphi_{ij}(t) = 2 \sin(\pi i t_1) \sin(\pi j t_2)$.

3.4. Conclusions of Chapter 3

1. Based on N-distances, the construction of statistical tests of uniformity of the hypersphere, homogeneity, symmetry and independence were proposed.
2. In the general case the limit null distribution of proposed N-metrics statistics coincides with the distribution of infinite quadratic form of Gaussian random variables. Under the alternative hypothesis, proposed tests statistics are asymptotically normal.
3. In the general case proposed N-metrics statistics are not distribution-free. In case of homogeneity hypothesis to avoid this problem bootstrap and permutation approaches are suggested to be used.
4. A construction of multivariate distribution-free two sample test, based on N-distances is proposed.

4

Power comparison

In this section we compare proposed N-distance tests with some classical criteria. In the first part as a measure for comparison of criteria we consider Asymptotic Relative Efficiency (ARE) by Bahadur (Bahadur, 1960; Nikitin, 1995). For simplicity, we deal solely with nonparametric goodness of fit tests in univariate case. In the second part a comparative Monte Carlo power study is proposed. Besides simple and composite hypothesis of goodness of fit, we consider two-sample tests in uni- and multivariate cases. A wide range of alternative hypotheses are investigated.

4.1. Asymptotic relative efficiency of criteria

The problem of comparing of nonparametric tests on the basis of some quantitative characteristic that will make it possible to order these tests and recommend the proper test one should use in a given problem is extremely important. The asymptotic efficiency is just the most known and useful characteristic of such kind (Nikitin, 1995).

Let U_n and V_n be two sequences of statistics based on a given sample of size n and assigned for testing the null hypothesis H_0 against the alternative H_1. Assume that H_1 is characterized by a certain parameter θ and for $\theta = \theta_0$ turns into H_0. Denote by $N_U(\alpha, \beta, \theta)$ the sample size necessary for the sequence U_n in order to

attain the power β under the level α and the alternative value of parameter θ. The relative efficiency of the sequences U_n with respect to the sequence V_n is specified as the quantity

$$e_{U,V}(\alpha,\beta,\theta) = \frac{N_V(\alpha,\beta,\theta)}{N_U(\alpha,\beta,\theta)}. \tag{4.1}$$

The value of $e_{U,V}(\alpha,\beta,\theta)$ greater than 1 says that for given α, β and θ the sequence of V_n is preferable to U_n since for this sequence of statistics we need less observations at the given level of α and the alternative θ to reach the power β. Unfortunately it is extremaly difficult to explicitly calculate $N_U(\alpha,\beta,\theta)$ even for rather simple sequences of statistics U_n. There are several ways to avoid this problem, one of which was proposed by Bahadur in (Bahadur, 1960).

The Bahadur approach prescribes to fix the power of criterion β and compare the speed of decreasing of their levels α once the sample size n increases. That is if for fixed $\beta \in (0,1)$ and θ there exist a limit

$$e_{U,V}(\beta,\theta) = \lim_{\alpha\to 0} e_{U,V}(\alpha,\beta,\theta),$$

then it is called relative asymptotic efficiency of the sequence U_n with respect to V_n by Bahadur.

Denote for any θ and t and any sequence of statistics U_n

$$F_n(t,\theta) = P_\theta(\omega : U_n(\omega) < t)$$

and

$$G_n(t) = F_n(t,\theta_0),$$

where $\theta = \theta_0$ corresponds to null hypothesis. The value of

$$L_n(\omega) = 1 - G_n(U_n(\omega)) \tag{4.2}$$

is called the attained level or P-value. In case $\theta = \theta_0$ and $F_n(t,\theta_0)$ is a continuous function, L_n is uniformly distributed on [0, 1]. Under the alternative hypothesis, when $\theta \neq \theta_0$ there exists convergence by P_θ-probability

$$\lim_{n\to\infty} n^{-1} \ln L_n = -\frac{1}{2} c_U(\theta), \tag{4.3}$$

where $c_U(\theta)$ is a nonrandom positive function of parameter θ, that is called Bahadur exact slope of the sequence U_n. According to Bahadur if (4.3) is valid for the

sequence of statistics U_n with $c_U(\theta) > 0$, then

$$N_U(\alpha, \beta, \theta) \sim \frac{2 \ln 1/\alpha}{c_U(\theta)}, \quad \alpha \to 0.$$

Therefore the calculation of ARE by Bahadur would be reduced to the ratio of exact slopes of the sequences U_n and V_n. One of the simplest methods for calculating Bahadur exact slopes can be derived from the following theorem (see Theorem 1.2.2 in (Nikitin, 1995)).

Theorem 24. *Let the sequence of statistics U_n satisfy the conditions:*

- $\frac{U_n}{\sqrt{n}} \to b(\theta)$ *in P_θ-probability,*
- $\lim_{n\to\infty} n^{-1} \ln[1 - G_n(t)] = -f(t)$ *for all t from the open interval I, where $f(t)$ is continuous on I and $\{b(\theta)\} \subset I$.*

Then (4.3) holds and

$$c_U(\theta) = 2f(b(\theta)).$$

In some cases it is rather difficult to establish the function $f(t)$, because of the complexity of calculation of probability of large-deviations. To avoid this problem it was proposed by Bahadur that the exact distribution of statistics U_n in (4.2) be replaced by its limiting distribution. Suppose that for all $t \in \mathbb{R}$ there exists continuous distribution function F such that

$$F_n(t, \theta_0) \to F(t), \quad n \to \infty,$$

then (4.2) can be represented in the form

$$L_n^*(\omega) = 1 - F(U_n(\omega)).$$

In case there exist a limit in P_θ-probability

$$\lim_{n\to\infty} n^{-1} \ln L_n^* = -\frac{1}{2} c_U^*(\theta) > 0 \tag{4.4}$$

function $c_U^*(\theta)$ is called the approximate slope of the sequence U_n and the ratio of approximate slopes of two sequences of statistics U_n and V_n is called their approximate Bahadur ARE. The method for calculating the approximate slopes is much the same described in Theorem 24. If in P_θ-probability

$$\frac{U_n}{\sqrt{n}} \to b(\theta)$$

and for some constant a, $0 < a < \infty$, limiting function $F(t)$ satisfies the condition

$$\ln(1 - F(t)) \sim -\frac{1}{2}at^2, \quad t \to \infty,$$

then (4.4) holds and

$$c_U^* = ab^2(\theta). \tag{4.5}$$

The approximate and exact slopes are often locally equivalent as $\theta \to \theta_0$, so the approximate ARE gives the notion of the local exact ARE.

Consider now N-distance tests for simple hypothesis of goodness of fit with a strongly negative definite kernel $L(x, y)$ (see section 2.1.2)

$$T_n = -n \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x, y)\, d(F_n(x) - G(x))\, d(F_n(y) - G(y)),$$

where $F_n(x)$ is the empirical distribution function based on a given sample $X_1, \ldots, X_n$ of observations of random variable X with continuous distribution function $F(x)$ and $G(x)$ is continuous distribution function corresponding to the null hypothesis. Let us first standardize our sequence of statistics $\{[T_n]^{1/2}\}$ so, that Theorem 24 becomes applicable. After that the expression for function $b(\theta)$ can be obtained with the help of Glivenko-Cantelli theorem

$$b(\theta) = \left[- \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x, y)\, d(F(x) - G(x))\, d(F(y) - G(y)) \right]^{1/2}.$$

By using the same arguments as for Cramer-von Mises type statistics, see e.g. (Koziol, 1986; Nikitin, 1995) one can see, that $\{[T_n]^{1/2}\}$ is sequence with the approximate slope

$$c_{T_n}^*(F, G) = \frac{b^2(\theta)}{\lambda(L, F)}, \tag{4.6}$$

where $\lambda(L, F)$ is the largest eigenvalue of integral operator (2.5) with the kernel

$$H(x, y) = \mathbf{E}L(x, X) + \mathbf{E}L(X, y) - L(x, y) - \mathbf{E}L(X, X'),$$

where X, X' are independent random variables with cumulative distribution function $F(x)$. In other words $\lambda(L, F)$ is the largest coefficient of diagonalized quadratic form (2.11) of independent standard normal random variables (the distribution of this quadratic form coincides with the limit null distribution of T_n).

Let us in more detail consider the case of location alternative and compare proposed N-distance tests with classical criteria presented in Table 3 in (Nikitin, 1995) for two hypothesized distribution functions $G(x)$: normal (with density function $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp -x^2/2$) and logistic (with density function $g(x) = \exp x(1 + \exp x)^{-2}$). In this case H_1 is characterized by a shift parameter θ with $F(x) = G(x + \theta)$ and for $\theta = 0$ turns into H_0.

In case of N-distance statistics T_n functions $b(\theta)$ can be presented in the form

$$b(\theta) = \left[- \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x, y) \triangle g(x, \theta) \triangle g(y, \theta) \, dx \, dy \right]^{1/2}, \qquad (4.7)$$

where $\triangle g(x, \theta) = g(x + \theta) - g(x)$ and $g(x) = G^{'}(x)$.

In the most important case of close alternatives, approximate Bahadur slopes (4.5) can be replaced by local slopes, when $\theta \to 0$, and therefore exact ARE approximated by local ARE.

From (4.6) and (4.7) we have that the principal part of the local approximate Bahadur slopes, as $\theta \to 0$, of the sequence $\{[T_n]^{1/2}\}$ have the form

$$c^*_{T_n}(\theta) \sim - \frac{\theta^2}{\lambda(L, F)} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} L(x, y) g^{'}(x) g^{'}(y) \, dx \, dy.$$

All the classical tests considered in (Nikitin, 1995) also have principal parts of the form $const * \theta^2$, when $\theta \to 0$, thus for our study it is sufficient to compare only coefficients of θ^2, which are called the local indices. In the table below we present the local indices for classical and N-distance statistics. N-metric tests were considered with the following strongly negative definite kernels:

$$L_1(x, y) = |x - y|, \quad L_2(x, y) = |x - y|^{\frac{1}{2}}, \quad L_3(x, y) = |x - y|^{\frac{3}{2}},$$

$$L_4(x, y) = \frac{|x - y|}{1 + |x - y|}, \quad L_5(x, y) = 1 - \exp^{-(x-y)^2},$$

$$L_6(x, y) = |G(x) - G(y)|,$$

$$L_7(x, y) = G(x) \vee G(y), \quad L_8(x, y) = G(x \vee y),$$

$$L_9(x, y) = \log(1 + (x - y)^2), \quad L_{10}(x, y) = \frac{(x - y)^2}{1 + (x - y)^2}.$$

The local indices for N-distance test were calculated numerically, by comput-

Table 4.1. Bahadur Local Indices for the location alternatives

Statistics	Gaussian distribution	Logistic distribution
Likelihood ratio statistic	1.000	0.333
Kolmogorov-Smirnov D_n	0.640	0.250
Anderson-Darling A_n^2	0.960	0.333
Omega-square ω_n^1	0.955	0.333
Omega-square ω_n^2	0.906	0.329
Omega-square ω_n^3	0.870	0.320
Omega-square ω_n^4	0.560	0.300
Watson U_n^2	0.490	0.220
Khmaladze-Aki K_n	0.814	0.250
Khmaladze-Aki L_n^2	0.940	0.329
N-distance $T_n(L_1)$	0.949	0.324
N-distance $T_n(L_2)$	0.878	0.321
N-distance $T_n(L_3)$	0.991	0.327
N-distance $T_n(L_4)$	0.825	0.291
N-distance $T_n(L_5)$	0.758	0.230
N-distance $T_n(L_6)$	0.906	0.329
N-distance $T_n(L_7)$	0.906	0.329
N-distance $T_n(L_8)$	0.915	0.332
N-distance $T_n(L_9)$	0.918	0.324
N-distance $T_n(L_{10})$	0.782	0.267

ing the largest coefficients of limiting diagonalized quadratic forms from Theorem 9. The first row refers to the likelihood ratio statistics, which in Bahadur theory is asymptotically optimal and has the largest exact slope and local index. As it was mentioned in section 2.1.2, N-distance statistics with the kernels $L_{6,7}$ are very similar to classical Cramer-von Mises statistics ω_n^2, therefore is was quite natural to get equal local indices for all of them.

4.2. Empirical power comparison

Let us switch to a comparative Monte Carlo power study of goodness of fit (simple and composite hypothesis) and homogeneity tests in uni- and bivariate cases. N-distance tests with several strongly negative definite kernels are compared

with some classical criteria using a wide range of alternative hypotheses. Proposed alternatives gave us a variety of types of departure from null hypothesis and allowed to test the sensitivity of criteria to each of them. Results of simulations show that proposed tests are powerful competitors to existing classical ones, in the sense that they are consistent against all alternatives and have relatively good power against general alternatives compared with other tests. The possibility in the selection of the kernel for N-distance allows to create the test more sensitive to particular type of alternatives.

4.2.1. Simple hypothesis of goodness of fit

We start from the simple hypothesis of goodness of fit. For the comparative analysis we have chosen N-distance statistics based on strongly negative kernels represented in section 4.1 and three classical nonparametric criteria: Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM) and Anderson-Darling (AD).

Simulation design

In all the cases we investigate the behavior of above mentioned tests for sample sizes $n = 25, 50, 100, 200$ and significance level $\alpha = 0.05$. The first part of simulations (Tables 4.2–4.3) is devoted to univariate simple hypotheses of normality with $N(0, 1)$ as a hypothesized distribution. In the second part we consider hypothesis of exponentiality with $Exp(1)$ as null distribution.

The power of the tests was estimated from a simulation of 200 samples of alternative distributions: Logistic, Gamma, Lognormal, mixtures of Normal and Exponential distributions with different location and scale parameters.

Simulation results

Empirical results summarized in Tables 4.2–4.3 illustrate that none of the tests are universally superior. Against the traditional location alternative all N-distance tests have rather similar results in comparison with classical ones. $T_n(L_2, 4)$ tests, being less sensitive to the differences in the tails of distribution, showed really good results against the contamination of normal distribution $N(0, 1)$ with $N(0, 0.1)$. But in the similar case of exponential distribution their performance was not so powerful. $T_n(L_{1,2,3,9})$ tests were more sensitive against normal location/scale mixtures than Kolmogorov-Smirnov and Cramer-von Mises criteria, but less powerful in this comparison than Anderson-Darling test. On the other hand, $T_n(L_{2,3,4,9})$ were comparable to or better than Anderson-Darling statistics against the similar alternatives for hypothesis of exponentiality. Practically all proposed N-distance

tests were better than classical criteria against mixtures of null distributions with uniform distributions both with 0.1 and 0.2 mixing probabilities.

4.2.2. Composite hypothesis of goodness of fit

We continue with comparative Monte Carlo power study of parametric hypothesis of goodness of fit, where in particular consider normality and exponentiality tests. N-distance tests with several strongly negative definite kernels are compared with classical criteria: D'Agostino (A), Cramer-von Mises (CvM), Anderson-Darling (AD), Lilliefors (KS), Pearson (P), Shapiro-Wilk (SW), Shapiro-Francia (SF) in univariate case; and Mardia, Henze-Zirkler (HZ), Mahalanobis[1] in bivariate case.

Simulation design

In all the cases we investigate the behavior of above mentioned tests for sample sizes $n = 25, 50, 100, 200$ and significance level $\alpha = 0.05$. The first part of simulations (Tables 4.4–4.5) is devoted to univariate normality and exponentiality tests with hypothesized distributions $N(0, 1)$ and $Exp(1)$ correspondingly. In the second part of our study (Table 4.6) we consider bivariate normality test with normal distribution with zero mean vector and covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ as a null cumulative distribution function.

The power of the tests was estimated from a simulation of 200 samples from alternative distributions: Logistic, Student, Gamma, Weibull, Lognormal and mixtures of Normal and Exponential with different location and scale parameters.

For comparative analysis we have chosen N-distance statistics based on six strongly negative kernels $L_{1,2,3,4,5,8}$ in univariate case (see section 4.1) and three kernels in bivariate case:

$$L_{11}(x, y) = \Phi(x_1 \vee y_1) + \Phi(x_2 \vee y_2) - \Phi(x_1 \vee y_1)\Phi(x_2 \vee y_2),$$

$$L_{12}(x, y) = \|x - y\|,$$

$$L_{13}(x, y) = 1 - exp(-\|x - y\|^2),$$

where $x, y \in \mathbb{R}^2$.

[1] After Mahalanibis transformation one-sample Kolmogorov-Smirnov test is applied

Table 4.2. Percentage of rejected simple hypothesis of normality

$Alternative$	n	$T_n(L_4)$	$T_n(L_2)$	$T_n(L_1)$	$T_n(L_3)$	$T_n(L_9)$	$T_n(L_8)$	$T_n(L_5)$	CvM	AD	KS
$Logistic(0,1)$	25	63	64	65	54	76	29	69	24	76	24
$Logistic(0,1)$	50	87	92	93	85	94	57	89	62	96	48
$Logistic(0,1)$	100	100	100	100	99	100	91	100	96	100	92
$Logistic(0,1)$	200	100	100	100	100	100	100	100	100	100	100
$N(0.3,1)$	25	22	23	24	32	36	34	31	30	27	25
$N(0.3,1)$	50	37	49	54	59	50	49	44	52	54	47
$N(0.3,1)$	100	77	82	89	90	81	79	71	86	90	83
$N(0.3,1)$	200	95	97	98	98	99	99	97	97	97	95
$0.9N(0,1)+0.1N(3,1)$	25	9	10	12	18	16	8	12	8	15	8
$0.9N(0,1)+0.1N(3,1)$	50	20	35	49	59	39	19	21	23	66	14
$0.9N(0,1)+0.1N(3,1)$	100	61	80	91	92	80	43	46	51	98	37
$0.9N(0,1)+0.1N(3,1)$	200	98	100	100	100	100	73	89	70	100	76
$0.8N(0,1)+0.2N(3,1)$	25	33	51	72	83	76	32	39	32	91	27
$0.8N(0,1)+0.2N(3,1)$	50	91	99	100	100	100	72	86	82	100	83
$0.8N(0,1)+0.2N(3,1)$	100	100	100	100	100	100	100	100	100	100	100
$0.8N(0,1)+0.2N(3,1)$	200	100	100	100	100	100	100	100	100	100	100
$0.9N(0,1)+0.1N(0,3)$	25	4	3	6	10	11	6	9	3	10	4
$0.9N(0,1)+0.1N(0,3)$	50	11	14	18	18	13	5	10	14	21	12
$0.9N(0,1)+0.1N(0,3)$	100	19	22	27	27	17	8	10	17	43	13
$0.9N(0,1)+0.1N(0,3)$	200	26	27	31	30	41	11	19	18	50	15
$0.8N(0,1)+0.2N(0,3)$	25	12	13	16	22	20	10	16	10	34	7
$0.8N(0,1)+0.2N(0,3)$	50	26	37	44	42	41	10	27	14	60	15
$0.8N(0,1)+0.2N(0,3)$	100	55	65	73	71	72	18	45	30	90	20
$0.8N(0,1)+0.2N(0,3)$	200	88	91	94	92	96	35	80	38	99	26
$0.9N(0,1)+0.1N(0,0.1)$	25	4	3	3	3	5	5	6	3	3	4
$0.9N(0,1)+0.1N(0,0.1)$	50	8	9	5	5	5	5	8	8	5	8
$0.9N(0,1)+0.1N(0,0.1)$	100	28	20	12	6	11	11	18	14	12	14
$0.9N(0,1)+0.1N(0,0.1)$	200	50	35	14	7	12	15	29	15	13	29
$0.8N(0,1)+0.2N(0,0.1)$	25	9	5	5	5	5	5	13	5	5	6
$0.8N(0,1)+0.2N(0,0.1)$	50	29	29	12	5	7	8	27	15	10	23
$0.8N(0,1)+0.2N(0,0.1)$	100	86	76	45	16	30	39	61	50	46	55
$0.8N(0,1)+0.2N(0,0.1)$	200	100	98	77	25	65	84	90	78	77	87
$0.9N(0,1)+0.1U_{[0,1]}$	25	6	6	7	8	5	8	9	8	7	8
$0.9N(0,1)+0.1U_{[0,1]}$	50	10	11	9	9	7	9	13	12	8	13
$0.9N(0,1)+0.1U_{[0,1]}$	100	19	16	16	12	12	14	17	19	15	17
$0.9N(0,1)+0.1U_{[0,1]}$	200	38	30	21	18	24	23	35	22	20	30
$0.8N(0,1)+0.2U_{[0,1]}$	25	9	10	16	9	7	7	7	10	7	9
$0.8N(0,1)+0.2U_{[0,1]}$	50	16	18	26	19	20	18	14	20	15	23
$0.8N(0,1)+0.2U_{[0,1]}$	100	32	33	52	56	47	40	27	47	39	45
$0.8N(0,1)+0.2U_{[0,1]}$	200	66	67	84	84	79	64	49	63	64	73

Table 4.3. Percentage of rejected simple hypothesis of exponentiality

$Alternative$	n	$T_n(L_4)$	$T_n(L_2)$	$T_n(L_1)$	$T_n(L_3)$	$T_n(L_9)$	C-v-M	A-D	K-S
$Exp(0.8)$	25	22	25	28	30	26	26	28	23
$Exp(0.8)$	50	32	34	36	39	35	32	35	30
$Exp(0.8)$	100	50	58	64	68	63	58	62	45
$Exp(0.8)$	200	82	84	88	90	88	80	86	75
$Lognormal(0,1)$	25	48	58	65	70	52	59	69	48
$Lognormal(0,1)$	50	75	82	84	88	78	84	92	74
$Lognormal(0,1)$	100	96	98	99	99	95	98	100	95
$Lognormal(0,1)$	200	100	100	100	100	100	100	100	100
$Gamma(0.8)$	25	23	22	21	18	16	35	38	29
$Gamma(0.8)$	50	32	29	26	23	25	42	48	35
$Gamma(0.8)$	100	66	67	68	67	62	76	78	65
$Gamma(0.8)$	200	95	92	90	89	90	96	97	94
$0.9Exp(1)+0.1Exp(0.2)$	25	6	7	17	26	10	8	9	9
$0.9Exp(1)+0.1Exp(0.2)$	50	12	19	32	51	24	13	26	12
$0.9Exp(1)+0.1Exp(0.2)$	100	22	40	72	84	41	26	57	16
$0.9Exp(1)+0.1Exp(0.2)$	200	45	72	93	97	83	35	81	33
$0.8Exp(1)+0.2Exp(0.2)$	25	28	47	70	81	46	27	57	21
$0.8Exp(1)+0.2Exp(0.2)$	50	43	70	86	91	73	36	78	27
$0.8Exp(1)+0.2Exp(0.2)$	100	76	97	100	100	99	71	99	62
$0.8Exp(1)+0.2Exp(0.2)$	200	99	100	100	100	100	99	100	98
$0.9Exp(1)+0.1Exp(5)$	25	8	6	6	6	6	11	8	11
$0.9Exp(1)+0.1Exp(5)$	50	13	11	6	6	9	19	18	14
$0.9Exp(1)+0.1Exp(5)$	100	16	15	14	14	13	27	27	16
$0.9Exp(1)+0.1Exp(5)$	200	32	28	27	22	32	36	38	34
$0.8Exp(1)+0.2Exp(5)$	25	19	19	16	12	14	28	28	25
$0.8Exp(1)+0.2Exp(5)$	50	31	26	21	18	21	41	39	35
$0.8Exp(1)+0.2Exp(5)$	100	53	51	51	43	45	69	70	54
$0.8Exp(1)+0.2Exp(5)$	200	92	90	85	79	83	95	96	95
$0.9Exp(1)+0.1U_{[0,3]}$	25	7	8	8	8	8	6	8	8
$0.9Exp(1)+0.1U_{[0,3]}$	50	8	8	8	8	8	9	9	8
$0.9Exp(1)+0.1U_{[0,3]}$	100	12	11	11	11	10	13	12	9
$0.9Exp(1)+0.1U_{[0,3]}$	200	22	19	19	16	23	18	18	16
$0.8Exp(1)+0.2U_{[0,3]}$	25	10	11	10	9	10	11	10	10
$0.8Exp(1)+0.2U_{[0,3]}$	50	12	13	10	10	10	12	11	11
$0.8Exp(1)+0.2U_{[0,3]}$	100	30	31	30	30	25	30	30	22
$0.8Exp(1)+0.2U_{[0,3]}$	200	58	60	55	52	45	55	52	47

Simulation results

Empirical results summarized in Tables 4.4–4.6 illustrate that none of the tests are universally superior, but some general aspects of power performance are evident. Practically all proposed N-distance tests showed better results against equal mixtures of normal or exponential distributions with different location and scale parameters both in uni- and bivariate cases (see Fig. 4.1–4.2).

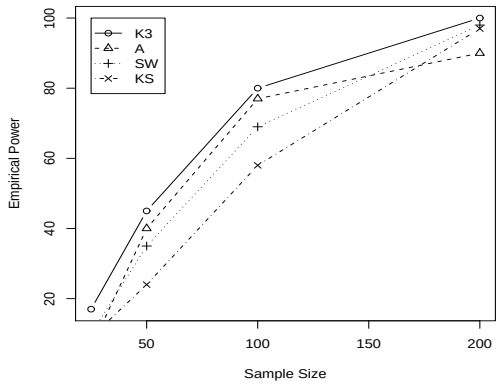

Figure 4.1. Empirical power of tests of univariate normality against location mixture $0.5N(0,1) + 0.5N(3,1)$

Similarly to simple goodness of fit test $T_n(L_4)$ test, being less sensitive to the differences in the tails of distribution, performs well against the contamination of hypothesized distribution $N(0,1)$ with $N(0,0.1)$. The same results against such alternative are shown by $T_n(L_2)$ test and are comparable only to Kolmogorov-Smirnov (KS) and Cramer-von Mises (CvM) tests among classical ones. In all the other cases, D'Agostino (A), Shapiro-Wilk (SW) and Shapiro-Francia (SF) tests were the most powerful with really impressive results against some non-normal alternatives. Their behavior was quite predictable, because mentioned tests are specified for testing normality only. However, in comparison with similar universal GoF tests like: Cramer-von Mises, Anderson-Darling and Kolmogorov-Smirnov (Lilliefors test) proposed N-distance criteria showed really competitive performance against all alternatives in case of normality test.

As for exponentiality criterion, N-distance tests were good against different mixtures of exponential distributions, but less sensitive to Lognormal, Weibull or

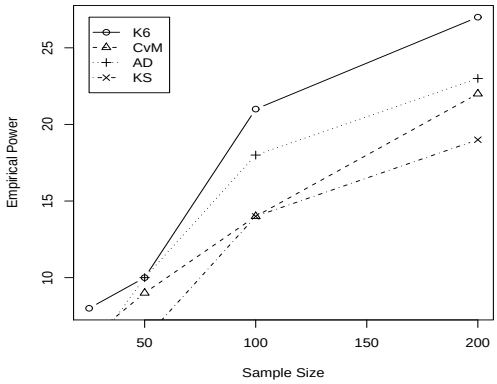

Figure 4.2. Empirical power of tests of exponentiality against location mixture $0.5Exp(1) + 0.5Exp(0.5)$

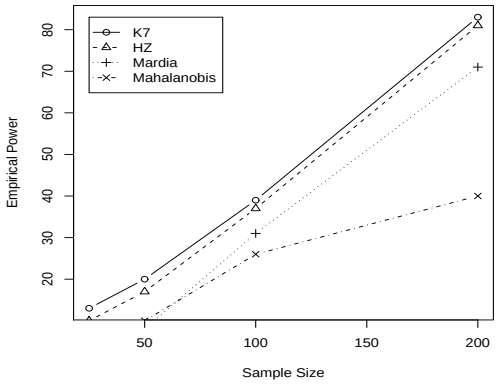

Figure 4.3. Empirical power of tests of bivariate normality against location mixture $0.5N(0, \Sigma) + 0.5N(2, I)$

Gamma alternatives than Anderson-Darling test.

In case of bivariate normality, one of the best results were shown by proposed $T_n(L_{11,12,13)}$ tests against mixtures of normal distributions with different location parameters (see Fig 4.3). $T_n(L_{12,13})$ tests were also good against equal mixtures of Normal distributions with different covariance matrixes.

All N-distance tests showed really impressive results in comparison with Mardia criterion against contamination of null distribution with $N(0, 0.1I))$. However, Mardia was the most powerful test against contamination of $N(0, \Sigma)$ with $N(0, 3I))$. Henze-Zirkler and $T_n(L_{13})$ tests, being very similar in their structure (Epps and Pulley, 1983; Henze and Zirkler, 1990; L.Baringhaus and H.Henze, 1998), predictably, showed comparable results against all considered alternatives.

4.2.3. Two-sample test

This section is devoted to a simulation power study of homogeneity tests in uni- and bivariate cases. N-distance tests with several strongly negative definite kernels are compared with four classical criteria: Kolmogorov-Smirnov (KS), Cramer-von Mises (CvM), Anderson-Darling (AD), Wilcoxon-Mann-Whitney (WMN).

Simulation design

In all the cases we investigate the behavior of above mentioned tests for sample sizes $n = 25, 50, 100, 200$ and significance level $\alpha = 0.05$.

We consider two-sample tests on the basis of standard normal distribution in univariate case (Table 4.7) and normal distribution with zero mean vector and covariance matrix $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ in bivariate case (Table 4.8).

The power of the tests was estimated from a simulation of 200 samples from alternative distributions: Logistic, Gaussian, a mixture of Normal distributions with different location and scale parameters.

For comparative analysis we have chosen N-distance statistics based on six strongly negative kernels $L_{1,2,3,4,8,9}$ in univariate case (see section 4.1) and four kernels in bivariate case:

$$L_{12}(x, y) = \|x - y\|, \quad L_{13}(x, y) = 1 - \exp(-\|x - y\|^2),$$

$$L_{14}(x, y) = \log(1 + \|x - y\|^2), \quad L_{15}(x, y) = \frac{\|x - y\|}{1 + \|x - y\|},$$

where $x, y \in \mathbb{R}^2$.

Table 4.4. Empirical power of tests of univariate normality

$Alternative$	n	$T_n(L_8)$	$T_n(L_5)$	$T_n(L_4)$	$T_n(L_2)$	$T_n(L_1)$	$T_n(L_3)$	A	CvM	AD	KS	SF	P	SW
$Logistic(0,1)$	25	9	9	9	9	10	11	15	11	13	8	19	9	15
$Logistic(0,1)$	50	12	15	14	15	15	16	23	15	17	10	27	8	21
$Logistic(0,1)$	100	19	26	23	25	25	28	36	24	28	20	40	9	32
$Logistic(0,1)$	200	35	30	27	31	35	37	57	30	35	23	54	15	45
$0.9N(0,1)+0.1N(3,1)$	25	19	20	20	20	20	22	37	22	23	17	31	10	26
$0.9N(0,1)+0.1N(3,1)$	50	47	53	42	42	53	56	52	50	57	34	60	17	59
$0.9N(0,1)+0.1N(3,1)$	100	80	84	78	77	82	84	86	81	82	68	86	33	86
$0.9N(0,1)+0.1N(3,1)$	200	97	98	97	97	99	100	99	97	100	93	100	72	100
$0.9N(0,1)+0.1N(0,3)$	25	21	23	21	22	24	27	44	22	26	22	40	18	33
$0.9N(0,1)+0.1N(0,3)$	50	31	49	42	42	49	52	61	47	52	35	65	29	61
$0.9N(0,1)+0.1N(0,3)$	100	65	67	60	61	66	71	87	58	70	50	85	29	81
$0.9N(0,1)+0.1N(0,3)$	200	86	90	89	90	93	94	97	92	94	82	100	53	98
$0.9N(0,1)+0.1N(0,0.1)$	25	6	4	5	5	4	5	5	6	5	6	11	9	9
$0.9N(0,1)+0.1N(0,0.1)$	50	10	12	12	12	11	11	7	12	13	15	11	14	12
$0.9N(0,1)+0.1N(0,0.1)$	100	22	14	24	24	20	17	9	22	20	21	11	21	9
$0.9N(0,1)+0.1N(0,0.1)$	200	45	25	48	48	39	32	13	46	38	48	21	44	20
$t(4)$	25	22	19	17	17	19	21	37	18	21	18	26	15	24
$t(4)$	50	35	43	38	39	43	46	50	41	46	34	60	20	51
$t(4)$	100	56	72	66	67	71	75	77	65	73	54	83	32	80
$t(4)$	200	83	85	83	83	86	88	93	87	88	75	93	48	92
$t(10)$	25	8	8	9	9	8	10	10	11	10	8	15	9	12
$t(10)$	50	10	10	10	10	12	12	17	13	13	12	20	6	14
$t(10)$	100	15	18	15	16	17	17	25	16	18	10	32	6	27
$t(10)$	200	23	20	18	18	20	26	40	19	22	13	41	6	34
$0.5N(0,1)+0.5N(3,1)$	25	14	10	17	16	13	10	6	13	12	9	5	11	10
$0.5N(0,1)+0.5N(3,1)$	50	37	40	45	41	36	30	40	39	39	24	18	22	35
$0.5N(0,1)+0.5N(3,1)$	100	81	79	80	80	77	72	77	75	76	58	53	47	69
$0.5N(0,1)+0.5N(3,1)$	200	100	100	100	99	99	99	99	100	99	97	97	85	98
$0.5N(0,1)+0.5N(2,1)$	25	4	4	7	8	6	4	3	6	7	6	2	5	4
$0.5N(0,1)+0.5N(2,1)$	50	5	6	9	8	6	5	4	8	7	7	4	8	7
$0.5N(0,1)+0.5N(2,1)$	100	14	13	14	13	12	11	15	12	12	8	4	8	10
$0.5N(0,1)+0.5N(2,1)$	200	25	29	26	25	29	27	28	26	27	19	14	12	22
$0.5N(0,1)+0.5N(0,2)$	25	13	12	14	16	12	13	17	15	13	13	19	10	15
$0.5N(0,1)+0.5N(0,2)$	50	15	16	18	19	18	18	21	20	19	16	24	12	21
$0.5N(0,1)+0.5N(0,2)$	100	26	30	29	28	31	33	38	29	32	21	43	15	35
$0.5N(0,1)+0.5N(0,2)$	200	53	55	50	50	57	58	55	52	57	36	69	18	61
$0.5N(0,1)+0.5N(0,3)$	25	30	30	30	32	31	31	25	31	31	27	36	17	30
$0.5N(0,1)+0.5N(0,3)$	50	46	52	51	51	51	52	46	53	55	40	62	31	49
$0.5N(0,1)+0.5N(0,3)$	100	88	89	88	88	89	88	70	89	89	74	90	48	85
$0.5N(0,1)+0.5N(0,3)$	200	100	100	100	100	100	100	93	100	100	100	99	85	99

Table 4.5. Empirical power of tests of univariate exponentiality

$Alternative$	n	$T_n(L_5)$	$T_n(L_4)$	$T_n(L_2)$	$T_n(L_1)$	$T_n(L_3)$	CvM	AD	KS
$0.9Exp(1)+0.1Exp(1/5)$	25	17	15	17	21	26	18	19	15
$0.9Exp(1)+0.1Exp(1/5)$	50	31	38	46	56	60	49	49	34
$0.9Exp(1)+0.1Exp(1/5)$	100	56	60	64	72	76	69	69	57
$0.9Exp(1)+0.1Exp(1/5)$	200	82	82	86	88	91	86	87	78
$0.8Exp(1)+0.2Exp(1/5)$	25	30	32	34	40	45	33	35	37
$0.8Exp(1)+0.2Exp(1/5)$	50	55	57	64	69	74	65	65	54
$0.8Exp(1)+0.2Exp(1/5)$	100	88	90	91	94	95	91	93	85
$0.8Exp(1)+0.2Exp(1/5)$	200	99	99	99	100	100	98	99	98
$0.9Exp(1)+0.1Exp(5)$	25	4	4	4	3	4	3	3	7
$0.9Exp(1)+0.1Exp(5)$	50	7	7	9	8	8	9	12	10
$0.9Exp(1)+0.1Exp(5)$	100	10	15	14	16	13	14	18	14
$0.9Exp(1)+0.1Exp(5)$	200	16	21	21	19	15	25	32	18
$0.8Exp(1)+0.2Exp(5)$	25	7	9	9	10	10	10	11	12
$0.8Exp(1)+0.2Exp(5)$	50	14	18	21	21	21	24	30	18
$0.8Exp(1)+0.2Exp(5)$	100	30	38	40	41	40	48	58	38
$0.8Exp(1)+0.2Exp(5)$	200	55	67	69	60	57	74	79	67
$0.5Exp(1)+0.5Exp(1/3)$	25	15	16	18	16	18	12	15	17
$0.5Exp(1)+0.5Exp(1/3)$	50	21	23	27	30	32	28	30	21
$0.5Exp(1)+0.5Exp(1/3)$	100	46	53	54	59	61	58	59	50
$0.5Exp(1)+0.5Exp(1/3)$	200	79	77	78	80	82	79	81	72
$0.5Exp(1)+0.5Exp(1/2)$	25	5	5	6	7	8	6	5	5
$0.5Exp(1)+0.5Exp(1/2)$	50	6	6	7	9	10	9	10	6
$0.5Exp(1)+0.5Exp(1/2)$	100	12	16	15	18	21	14	18	14
$0.5Exp(1)+0.5Exp(1/2)$	200	21	22	23	26	27	22	23	19
$Lognormal(0,1)$	25	12	14	14	17	21	13	11	15
$Lognormal(0,1)$	50	20 31	33	34	31	34	40	21	
$Lognormal(0,1)$	100	42	62	65	62	56	60	79	46
$Lognormal(0,1)$	200	71	90	91	81	75	87	98	81
$Weibull(1.2)$	25	11	13	14	12	11	14	11	13
$Weibull(1.2)$	50	14	20	24	23	22	26	27	18
$Weibull(1.2)$	100	34	41	40	46	42	47	53	37
$Weibull(1.2)$	200	77	78	80	78	79	86	88	73
$Gamma(0.8)$	25	6	5	5	7	8	6	10	8
$Gamma(0.8)$	50	13	19	24	24	24	26	38	19
$Gamma(0.8)$	100	21	26	27	28	28	30	44	29
$Gamma(0.8)$	200	38	46	48	42	43	55	70	51
$Gamma(1.4)$	25	11	12	13	11	9	13	10	19
$Gamma(1.4)$	50	23	29	31	30	29	33	37	29
$Gamma(1.4)$	100	40	48	49	49	49	57	64	47
$Gamma(1.4)$	200	71	80	81	75	73	88	93	80

Table 4.6. Empirical power of tests of bivariate normality

$Alternative$	n	$T_n(L_{11})$	$T_n(L_{12})$	$T_n(L_{13})$	HZ	Mardia	Mahalonobis
$0.9N(0,\Sigma)+0.1N(3,I)$	25	23	29	37	33	29	10
$0.9N(0,\Sigma)+0.1N(3,I)$	50	38	59	59	59	58	22
$0.9N(0,\Sigma)+0.1N(3,I)$	100	76	90	92	88	92	57
$0.9N(0,\Sigma)+0.1N(3,I)$	200	100	100	100	100	100	72
$0.8N(0,\Sigma)+0.2N(3,I)$	25	33	40	45	42	11	7
$0.8N(0,\Sigma)+0.2N(3,I)$	50	60	78	75	78	41	11
$0.8N(0,\Sigma)+0.2N(3,I)$	100	95	99	98	96	87	18
$0.8N(0,\Sigma)+0.2N(3,I)$	200	100	100	100	100	100	30
$0.9N(0,\Sigma)+0.1N(0,3I)$	25	42	44	55	47	63	35
$0.9N(0,\Sigma)+0.1N(0,3I)$	50	72	82	81	82	91	81
$0.9N(0,\Sigma)+0.1N(0,3I)$	100	92	96	97	95	100	98
$0.9N(0,\Sigma)+0.1N(0,3I)$	200	100	100	100	100	100	100
$0.8N(0,\Sigma)+0.2N(0,3I)$	25	66	78	79	81	78	69
$0.8N(0,\Sigma)+0.2N(0,3I)$	50	89	95	95	95	97	94
$0.8N(0,\Sigma)+0.2N(0,3I)$	100	99	100	100	100	100	100
$0.8N(0,\Sigma)+0.2N(0,3I)$	200	100	100	100	100	100	100
$0.5N(0,\Sigma)+0.5N(3,I)$	25	26	27	25	23	3	24
$0.5N(0,\Sigma)+0.5N(3,I)$	50	50	73	56	73	14	42
$0.5N(0,\Sigma)+0.5N(3,I)$	100	90	99	98	99	52	83
$0.5N(0,\Sigma)+0.5N(3,I)$	200	100	100	100	100	90	98
$0.5N(0,\Sigma)+0.5N(2,I)$	25	12	10	8	10	5	7
$0.5N(0,\Sigma)+0.5N(2,I)$	50	20	17	10	17	8	10
$0.5N(0,\Sigma)+0.5N(2,I)$	100	39	36	26	37	31	26
$0.5N(0,\Sigma)+0.5N(2,I)$	200	83	82	80	81	71	40
$0.5N(0,\Sigma)+0.5N(0,\Omega^a)$	25	10	10	10	9	10	9
$0.5N(0,\Sigma)+0.5N(0,\Omega)$	50	14	20	16	18	13	11
$0.5N(0,\Sigma)+0.5N(0,\Omega)$	100	17	26	24	23	21	20
$0.5N(0,\Sigma)+0.5N(0,\Omega)$	200	30	67	55	53	42	39
$0.9N(0,\Sigma)+0.1N(0,0.1I)$	25	11	9	8	9	5	6
$0.9N(0,\Sigma)+0.1N(0,0.1I)$	50	14	11	9	11	9	19
$0.9N(0,\Sigma)+0.1N(0,0.1I)$	100	31	20	24	25	14	57
$0.9N(0,\Sigma)+0.1N(0,0.1I)$	200	84	77	87	87	26	98
$0.8N(0,\Sigma)+0.2N(0,0.1I)$	25	38	31	39	29	12	32
$0.8N(0,\Sigma)+0.2N(0,0.1I)$	50	70	71	83	73	25	88
$0.8N(0,\Sigma)+0.2N(0,0.1I)$	100	100	98	100	99	44	100
$0.8N(0,\Sigma)+0.2N(0,0.1I)$	200	100	100	100	100	71	100

[a] Ω denotes a matrix with 1 on diagonal and 0.9 off diagonal.

Simulation results

The empirical results for homogeneity tests are summarized in Tables 4.7–4.8. In comparison with goodness of fit tests (simple hypothesis), where all the statistics showed more or less similar results, the performance of N-distance tests was impressive against normal location/scale mixtures. Especially it concerns the statistics, based on the kernels $L_{12,14,15}$ in bivariate case, which showed more than twice better results against such alternatives for all sample sizes. $T_n(L_4)$ test was also the most sensitive against the alternatives when the variance of contaminating distribution was smaller than the variance of the main distribution.

4.2.4. Test of uniformity on hypersphere S^{p-1}

In conclusion of our empirical power study we proposed a brief comparison of several criteria of uniformity on hypersphere. N-distance tests with strongly negative definite kernel $L(x, y) = \|x - y\|$ are compared with classical criteria: Rayleigh (R) (Figueiredo, 2007), Watson (W) (Watson, 1961; 1967), Gine (G) (Gine, 1975) and Ajne (A) (Ajne, 1968; Beran, 1968) for circular S^1 and spherical S^2 cases.

Simulation design

In all the cases we investigate the behavior of above mentioned tests for sample sizes $n = 50, 100$ and significance level $\alpha = 0.05$. The first part of simulations (Table 4.9) is devoted to the circular case. In the second part of our study (Table 4.10) we consider uniformity test on the sphere S^2.

The power of the tests was estimated from a simulation of 200 samples Z of alternative distributions on the circle and sphere, which were modeled using the formulas:

- Circular data

$$Z = (\cos 2\pi X,\ \sin 2\pi X),$$

  where X is a random variable with distributions from the first column of Table 4.9.

- Spherical data

$$Z = (\cos(2\pi X),\ \sin(2\pi X)(1 - 2Y),\ \sin(2\pi X)\sin(\arccos(1 - 2Y))),$$

Table 4.7. Percentage of rejected homogeneity hypothesis in univariate case

$Alternative$	n	$T_n(L_4)$	$T_n(L_2)$	$T_n(L_1)$	$T_n(L_3)$	$T_n(L_9)$	$T_n(L_8)$	KS	CvM	WMN	AD
$N(0.3,1)$	25	18	19	23	25	22	19	16	20	21	22
$N(0.3,1)$	50	20	24	32	31	26	24	12	25	29	31
$N(0.3,1)$	100	37	42	48	48	46	46	43	46	49	46
$N(0.3,1)$	200	73	79	85	85	82	84	81	83	84	85
$Logistic(0,1)$	25	31	36	36	28	45	14	8	11	-	21
$Logistic(0,1)$	50	60	65	69	44	73	16	9	15	-	47
$Logistic(0,1)$	100	96	96	96	84	96	51	44	50	-	83
$Logistic(0,1)$	200	100	100	100	100	100	96	91	96	-	100
$0.9N(0,1)+0.1N(3,1)$	25	4	5	6	9	6	5	4	5	5	11
$0.9N(0,1)+0.1N(3,1)$	50	8	10	20	24	15	10	3	9	10	14
$0.9N(0,1)+0.1N(3,1)$	100	23	32	49	53	39	20	21	19	22	31
$0.9N(0,1)+0.1N(3,1)$	200	43	73	92	92	82	43	39	39	39	74
$0.8N(0,1)+0.2N(3,1)$	25	20	31	48	55	39	20	11	20	22	31
$0.8N(0,1)+0.2N(3,1)$	50	50	67	87	85	79	38	18	37	39	63
$0.8N(0,1)+0.2N(3,1)$	100	96	99	100	100	100	75	81	77	67	97
$0.8N(0,1)+0.2N(3,1)$	200	100	100	100	100	100	99	100	100	96	100
$0.9N(0,1)+0.1N(0,5)$	25	9	6	10	10	7	6	7	6	-	6
$0.9N(0,1)+0.1N(0,5)$	50	9	6	12	14	8	6	7	6	-	7
$0.9N(0,1)+0.1N(0,5)$	100	12	12	19	18	14	7	7	8	-	10
$0.9N(0,1)+0.1N(0,5)$	200	16	27	52	45	34	13	13	12	-	20
$0.8N(0,1)+0.2N(0,5)$	25	8	13	23	28	17	7	7	7	-	10
$0.8N(0,1)+0.2N(0,5)$	50	15	25	49	48	36	7	7	7	-	12
$0.8N(0,1)+0.2N(0,5)$	100	38	64	84	79	79	11	12	9	-	21
$0.8N(0,1)+0.2N(0,5)$	200	76	99	100	100	100	26	23	20	-	71
$0.9N(0,1)+0.1N(0,0.1)$	25	6	6	6	5	6	6	9	5	-	5
$0.9N(0,1)+0.1N(0,0.1)$	50	12	9	6	6	6	6	12	6	-	8
$0.9N(0,1)+0.1N(0,0.1)$	100	12	11	9	7	6	6	13	6	-	10
$0.9N(0,1)+0.1N(0,0.1)$	200	19	15	12	8	7	9	17	12	-	14
$0.8N(0,1)+0.2N(0,0.1)$	25	6	6	6	5	6	6	10	7	-	7
$0.8N(0,1)+0.2N(0,0.1)$	50	17	10	6	6	6	6	14	7	-	7
$0.8N(0,1)+0.2N(0,0.1)$	100	39	28	12	9	11	12	21	14	-	13
$0.8N(0,1)+0.2N(0,0.1)$	200	77	72	37	11	21	38	59	35	-	41

Table 4.8. Percentage of rejected hypothesis of homogeneity in bivariate case

$Alternative$	n	$T_n(L_{12})$	$T_n(L_{13})$	$T_n(L_{14})$	$T_n(L_{15})$	CvM	AD	KS
$N(0.3, \Sigma)$	25	29	12	26	17	23	25	15
$N(0.3, \Sigma)$	50	35	20	35	23	30	31	25
$N(0.3, \Sigma)$	100	63	42	61	53	60	61	40
$N(0.3, \Sigma)$	200	94	72	86	79	87	87	86
$N(0.5, \Sigma)$	25	54	20	55	39	50	50	36
$N(0.5, \Sigma)$	50	75	44	78	64	72	73	63
$N(0.5, \Sigma)$	100	97	85	97	94	95	95	88
$N(0.5, \Sigma)$	200	100	100	100	100	100	100	100
$0.9N(0, \Sigma) + 0.1N(3, I)$	25	25	7	11	5	7	7	8
$0.9N(0, \Sigma) + 0.1N(3, I)$	50	39	10	18	10	12	11	11
$0.9N(0, \Sigma) + 0.1N(3, I)$	100	55	19	44	30	17	21	17
$0.9N(0, \Sigma) + 0.1N(3, I)$	200	99	29	89	64	25	40	30
$0.8N(0, \Sigma) + 0.2N(3, I)$	25	67	13	52	25	18	22	15
$0.8N(0, \Sigma) + 0.2N(3, I)$	50	97	27	96	62	35	45	32
$0.8N(0, \Sigma) + 0.2N(3, I)$	100	100	69	100	99	68	90	59
$0.8N(0, \Sigma) + 0.2N(3, I)$	200	100	100	100	100	99	100	100
$0.9N(0, \Sigma) + 0.1N(0, 10I)$	25	26	7	9	6	6	6	7
$0.9N(0, \Sigma) + 0.1N(0, 10I)$	50	42	7	24	9	6	6	8
$0.9N(0, \Sigma) + 0.1N(0, 10I)$	100	86	11	45	22	9	11	10
$0.9N(0, \Sigma) + 0.1N(0, 10I)$	200	100	15	95	38	11	15	15
$0.8N(0, \Sigma) + 0.2N(0, 10I)$	25	96	9	63	20	7	8	8
$0.8N(0, \Sigma) + 0.2N(0, 10I)$	50	100	14	100	40	9	9	10
$0.8N(0, \Sigma) + 0.2N(0, 10I)$	100	100	36	100	94	20	35	27
$0.8N(0, \Sigma) + 0.2N(0, 10I)$	200	100	68	100	100	31	68	59
$0.9N(0, \Sigma) + 0.1N(0, 0.1I)$	25	5	5	5	5	5	5	5
$0.9N(0, \Sigma) + 0.1N(0, 0.1I)$	50	7	8	7	11	5	5	8
$0.9N(0, \Sigma) + 0.1N(0, 0.1I)$	100	8	19	9	17	6	7	9
$0.9N(0, \Sigma) + 0.1N(0, 0.1I)$	200	14	39	8	33	8	8	19
$0.8N(0, \Sigma) + 0.2N(0, 0.1I)$	25	8	13	7	13	8	5	11
$0.8N(0, \Sigma) + 0.2N(0, 0.1I)$	50	12	31	9	28	13	10	24
$0.8N(0, \Sigma) + 0.2N(0, 0.1I)$	100	15	66	17	80	17	15	31
$0.8N(0, \Sigma) + 0.2N(0, 0.1I)$	200	62	96	46	100	33	29	63

where X, Y are independent random variables with distributions from the first column of Table 4.10.

Simulation results

Empirical results summarized in Tables 4.9–4.10 illustrate that none of the tests are universally superior. In S^1 case proposed N-distance criteria, together with Watson test, showed one of the best results against all considered alternatives for moderate sample sizes.

The empirical results for spherical data are summarized in Table 4.10. In comparison with circular case, where all the criteria, except possibly Gine test, showed more or less similar results, the performance of N-distance test was really good for all sample sizes against truncated uniform and von Mises distributions. Gine test, which was not so powerful against considered alternatives in S^1 case, was really sensitive to contamination of hypothesized distribution with truncated uniform in case of spherical data.

Table 4.9. Empirical power of tests of uniformity on the circle

$Alternative$	n	W	A	R	G	T_n
$Unif[0, 0.9]$	50	13	13	13	12	13
$Unif[0, 0.9]$	100	30	30	28	23	30
$Unif[0, 0.8]$	50	74	60	57	45	70
$Unif[0, 0.8]$	100	99	93	91	72	98
$0.9Unif[0, 1] + 0.1Unif[0, 0.1]$	50	15	13	13	15	15
$0.9Unif[0, 1] + 0.1Unif[0, 0.1]$	100	30	28	23	27	29
$0.8Unif[0, 1] + 0.2Unif[0, 0.1]$	50	54	40	40	47	54
$0.8Unif[0, 1] + 0.2Unif[0, 0.1]$	100	94	82	74	85	92
$0.8Unif[0, 1] + 0.2Unif[0, 0.25]$	50	44	39	40	20	45
$0.8Unif[0, 1] + 0.2Unif[0, 0.25]$	100	73	67	66	32	71
$0.8Unif[0, 1] + 0.2Unif[0, 0.5]$	50	16	15	15	6	17
$0.8Unif[0, 1] + 0.2Unif[0, 0.5]$	100	41	42	42	9	41
$vonMises(0, 0.5)^a$	50	58	59	58	7	59
$vonMises(0, 0.5)$	100	88	88	88	10	88
$vonMises(0, 0.3)$	50	27	26	27	7	29
$vonMises(0, 0.3)$	100	50	51	50	10	51
$0.5Unif[0, 1] + 0.5vonMises(0, 0.5)$	50	19	19	19	9	21
$0.5Unif[0, 1] + 0.5vonMises(0, 0.5)$	100	31	34	33	10	32
$0.5Unif[0, 1] + 0.5vonMises(0, 0.8)$	50	40	40	40	6	42
$0.5Unif[0, 1] + 0.5vonMises(0, 0.8)$	100	65	67	67	9	65

[a] von Mises distribution (also known as the circular normal distribution) with location and concentration parameters

Table 4.10. Empirical power of tests of uniformity on the sphere

Alternative	n	A	R	G	T_n
$Unif[0, 0.9]$	50	23	20	18	24
$Unif[0, 0.9]$	100	53	50	43	58
$Unif[0, 0.8]$	50	86	85	57	93
$Unif[0, 0.8]$	100	99	99	91	100
$vonMises(0, 0.5)$	50	38	32	36	42
$vonMises(0, 0.5)$	100	83	83	73	90
$vonMises(0, 0.3)$	50	14	14	16	15
$vonMises(0, 0.3)$	100	39	38	28	44
$0.9Unif[0, 1] + 0.1Unif[0, 0.1]$	50	13	12	16	14
$0.9Unif[0, 1] + 0.1Unif[0, 0.1]$	100	35	30	41	36
$0.8Unif[0, 1] + 0.2Unif[0, 0.1]$	50	54	41	81	66
$0.8Unif[0, 1] + 0.2Unif[0, 0.1]$	100	96	92	99	99

4.3. Conclusions of Chapter 4

1. The results of the theoretical and empirical power comparison study show that N-metrics tests are powerful competitors to existing classical criteria, in the sense that they are consistent against all alternatives and have relatively good power against general alternatives compared with other tests.
2. The possibility in the selection of the strongly negative definite kernel for N-distance allows to create the test more sensitive to particular type of alternative hypothesis.

General conclusions

1. Based on N-distances, the construction of statistical tests of goodness of fit, homogeneity, symmetry and independence were proposed.
2. In the general case the limit null distribution of N-metrics statistics coincides with the distribution of infinite quadratic form of Gaussian random variables. Under the alternative hypothesis, proposed tests statistics are asymptotically normal.
3. The results of the theoretical and empirical power comparison study show that N-metrics tests are powerful competitors to existing classical criteria, in the sense that they are consistent against all alternatives and have relatively good power against general alternatives compared with other tests. The possibility in the selection of the strongly negative definite kernel for N-distance allows to create the test more sensitive to particular type of alternative hypothesis.
4. In the general case proposed N-metrics statistics are not distribution-free. In case of homogeneity hypothesis to avoid this problem bootstrap and permutation approaches are suggested to be used.
5. For normality and nonparametric hypotheses of goodness of fit in high dimensional cases, when it is difficult from computational point of view to

determine the limit null distribution of N-distance statistic analytically, the critical region of the test can be established by means of Monte Karlo simulations.

Even in such classical statistical problems like testing the homogeneity or goodness of fit a lot of questions are still open. Most of them are related to these tests in multi dimensional cases. For now it is not absolutely clear, how to establish the null limit distribution of the test statistic in case of composite hypotheses of goodness of fit, when initial parametric family is not Gaussian; is it possible to avoid the dependency of the distribution of the two-sample test statistic on the unknown distribution of the samples; is there a strict way for selection of the kernel of N-distance statistic to make the test the most powerful? These and many other empirical and theoretical questions are the subject for the further research in this field.

Bibliography

Ajne, B. 1968. A simple test for uniformity of a circular distribution, *Biometrika*, 55: 343–354.

Aki, S. 1986. Some test statistics based on the martingale term of the empirical distribution function, *Annals of the Institute of Statistical Mathematics*, 38(1): 1–21.

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

Anderson, T.; Darling, D. 1954. A test of goodness of fit, *Journal of The American Statistical Association*, 49(268): 765–769.

Bahadur, R. 1960. Stochastic comparison of tests, *The Annals of Statistics*, 31(2): 276–295.

Baringhaus, L.; Franz, C. 2004. On a new multivariate two-sample test, *Journal of Multivariate Analysis*, (88): 190–206.

Beran, R. 1968. Testing for uniformity on a compact homogeneous space, *Journal of Applied Probability*, 5: 177–195.

Best, D. J.; Rayner, J. C. W. 1985. Lancaster's test of normality, *Journal of Statis-*

tical Planning and Inference, 12(3): 395–400.

Bickel, P. 1969. A distribution free version of the smirnov two-sample test in the multivariate case, *Annals of Mathematical Statistics*, 40: 1–23.

Bickel, P.; Breiman, L. 1983. Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test, *Annals of Probability*, 11: 185–214.

Billingsley, P. 1968. *Convergence of probability measures*. Wiley, New York.

Bowman, K.; Shenton, L. R. 1975. Omnibus contours for departures from normality based on b1 and b2, *Biometrica*, 62: 243–250.

Bulinskii, A.; Shiryaev, A. 2005. *Theory of random processes*. Fizmatlit, Moscow.

Burke, M. 2000. Multivariate tests-of-fit and uniform confidence bands using a weighted bootstrap, *Statistics and probability letters*, 46: 13–20.

Csorgo, S. 1986. Testing for normality in arbitrary dimension, *Annals of Mathematical statistics*, 14(2): 708–723.

DAgostino, R. 1971. An omnibus test of normality for moderate and large size samples, *Biometrika*, 58: 341–348.

DAgostino, R.; Belanger, A.; Jr, R. D. 1990. An omnibus test of normality for moderate and large size samples, *The American Statistician*, 44: 316–322.

DAgostino, R. B.; Stephens, M. A. 1986. *Goodness-of-Fit Techniques*. Marcel Dekker, New York/ Basel.

Darling, D. 1957. The kolmogorov-smirnov, cramer-von mises test, *Annals of Mathematical Statistics*, 28(3): 823–838.

Darling, D. 1983a. On the asymptotic distribution of watson's statistics, *Annals of Statistics*, 11(6): 1263–1266.

Darling, D. 1983b. On the supremum of a certain gaussian process, *Annals of Statistics*, 11(4): 803–806.

Durbin, J. 1970. Asymptotic distirbutions of some statistics based on the bivariate sample distribution function, *Non-parametric techniques in the statistical inference* 435–449.

Durbin, J. 1973. Weak convergence of the sample distribution function when parameters are estimated, *Annals of Mathematical Statistics*, 1(2): 279–290.

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

Figueiredo, A. 2007. Comparison of tests of uniformity defined on the hypersphere, *Statistics and Probability Letters*, 77(3): 329–334.

Figueiredo, A.; Gomes, P. 2003. Power of tests of uniformity defined on the hypersphere, *Communication in Statistics: Simulation and Computation*, 32(1): 87–94.

Friedman, J.; Rafsky, L. 1979. Multivariate generalizations of the wolfowitz and smirnov two-sample tests, *Annals of Mathematical Statistics*, 7: 697–717.

Gine, E. 1975. Invariant tests for uniformity on compact riemannian manifolds based on sobolev norms, *Annals of Statistics*, 3: 1243–1266.

Hajek, J.; Sidak, Z. 1967. *Theory of rank tests*. Academic Press, New York.

Henze, N. 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences, *Annals of Mathematical Statistics*, 16(2): 772–783.

Henze, N. 1994. On mardias kurtosis test for multivariate normality, *Communications in Statistics - Theory and Methods*, 23: 1031–1045.

Henze, N.; Wagner, T. 1997. A new approach to the bhep tests for multivariate normality, *Journal of Multivariate Analysis*, 62: 1–23.

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

Hermans, M.; Rasson, J. P. 1985. A new sobolev test for uniformity on the circle, *Biometrika*, 72(3): 698–702.

Imhof, J. P. 1961. Computing the distribution of quadratic forms in normal variables, *Biometrika*, 48(3).

Jupp, P. 2005. Sobolev tests of goodness of fit of distributions on compact riemannian manifolds, *The Annals of Statistics*, 33(6): 2957–2966.

Justel, A.; Pena, D.; Zamar, R. 1997. A multivariate kolmogorov-smirnov test of goodness of fit, *Statistics and Probability Letters*, 35: 251–259.

Kac, M.; Kiefer, J.; ; Wolfowitz, J. 1955. On tests of normality and other tests of goodness of fit based on distance methods, *Annals of Mathematical statistics*, 26(2): 189–211.

Khmaladze, E. 1977. On omega-square tests for parametric hypotheses, *Theory Probab. Appl.*, 22(3): 627–629.

Khmaladze, E. 1981. Martingale approach in the theory of goodness-of-fit tests, *Theory of Probability and its Applications*, 26(2): 246–265.

Klebanov, L. 2005. *N-distances and their applications*. Karolinum, Prague.

Kolmogorov, A. 1933. Sulla determinazione empririca di una legge di distribuzione, *Giorn. dell'Inst. Ital. degli Att.*, 4: 1–11.

Koroljuk, V.; Borovskich, Y. 1994. *Theory of U-statistics*. Kluwer Academic Publishers.

Koziol, J. 1986. Relative effciencies of goodness of fit procedures for assessing univariate normality, *Annals of the Institute of Statistical Mathematics*, 38: 121–132.

Koziol, J. A. 1983. On assessing multivariate normality, *J. Roy. Stat. Assoc.*, (3): 358–361.

Krivyakova, E.; Martynov, G.; Tyurin, Y. 1977. The distribution of the omega square statistics in the multivariate case, *Theory of Probability and its Applications*, 22(2): 415–420.

Kuiper, N. 1960. Tests concerning random points on the circle, *Proc. Kon. Ned. Akad. van Wet*, 63: 38–47.

L.Baringhaus, ; H.Henze, . 1992. Limit distributions for mardias measure of multivariate skewness, *Annals of Mathematical Statistics*, 20: 1889–1902.

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

Lee, A. 1990. *U-statistics: Theory and Practice*. Marcel Dekker, New York.

Lehmann, E. 1951. Consistency and unbiasedness of certain nonparametric tests, *Annals of Mathematical Statistics*, 22(2): 165–179.

Lilliefors, H. W. 1967. On the kolmogorov-smirnov test for normality with mean and variance unknown, *Journal of The American Statistical Association*, 62(318): 399–402.

Liu, R.; Yang, L. 2008. Kernel estimation of multivariate cumulative distribution function, *Journal of Nonparametric Statistics*, 0(0): 1–18.

Locke, C.; Spurrier, J. D. 1976. The use of u-statistics for testing normality against nonsymmetric alternatives, *Biometrika*, 63(1): 143–147.

Maag, U.; Stephens, M. 1968. The v(n,m) two-sample test, *Annals of Mathematical Statistics*, 39(3): 923–935.

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

Martynov, G. 1978. *Omega-Square criteria*. Nauka, Moscow.

Martynov, G. V. 1975. Computation of the distribution functions of quadratic forms in normal random variables, *Theory of Probability and its Applications*, 20(4): 797–809.

Nikitin, Y. 1995. *Asymptotic Efficiency of Nonparametric Tests*. Cambridge University Press, New York.

Park, S. 1999. A goodness-of fit test for normality based on the sample entropy of order statistics, *Statistics and Probability Letters*, 44: 359–363.

Pettitt, A. 1976. Two-sample anderson-darling rank statistics, *Biometrica*, 63(1): 161–168.

Pettitt, A. 1979. Two-sample cramer-von mises type rank statistics, *Journal of the Royal Statistical Society*, (1): 46–53.

Prescott, P. 1976. On test for normality based on sample entropy, *Journal of the Royal Statistical Society*, 38(3): 254–256.

Rosenblatt, M. 1952a. Limit theorems associated with variants of the von mises statistics, *Annals of Mathematical Statistics*, 23(4): 617–623.

Rosenblatt, M. 1952b. Remarks on a multivariate transformation, *Annals of Mathematical Statistics*, 23: 470–472.

Rothman, E. 1972. Tests for uniformity of a circular distribution, *Sankhya Ser. A*, 34: 23–32.

Shapiro, S.; Wilk, M. B. 1965. An analysis of variance test for normality, *Biometrika*, 52: 591–611.

Shapiro, S. S.; Francia, R. S. 1972. An approximate analysis of variance test for normality, *Journal of The American Statistical Association*, 67(337): 215–216.

Smirnov, N. 1939. On the estimation of the discrepancy between empirical curves

of the distribution for two independent samples, *Bull. Moscow Univ.*, 2: 3–6.

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

Spiegelhalter, O. J. 1977. A test for normality against symmetric alternatives, *Biometrika*, 64(2): 415–418.

Stute, W.; Gonzales-Manteiga, W.; Presedo-Quindimil, M. 1993. Bootstrap based goodness-of-fit-tests, *Metrika*, 40: 243–256.

Sukhatme, S. 1972. Fredholm determinant of a positive definite kernel of a special type and its application, *Annals of Mathematical Statistics*, 43(20): 1914–1926.

Szekely, G.; Rizzo, M. 2005. A new test for multivariate normality, *Journal of Multivariate Analysis*, 93: 58–80.

Szucs, G. 2008. Parametric bootstrap tests for continuous and discrete distributions, *Metrika*, 67: 63–81.

Towghi, N. 2002. Multidimensional extension of l.c. young's inequality, *Journal of Inequalities in Pure and Applied Mathematics*, 3(2). Prieiga per internetą: <http://jipam.vu.edu.au>.

Tyurin, Y. N. 1970. On testing parametric hypotheses by nonparametric methods, *Theory of Probability and its Applications*, 25(4): 745–749.

Tyurin, Y. N. 1984. On the limit distribution of kolmogorov-smirnov statistics for a composite hypothesis, *Izv. Akad. Nauk SSSR*, 48(6): 1314–1343.

Vaart, A. van der; Wellner, J. 1996. *Weak convergence and empirical processes*. Springer, New York.

Wald, A.; Wolfowitz, J. 1940. On the test whether two samples are from the same population, *Annals of Mathematical Statistics*, 11: 147–162.

Watson, G. 1961. Goodness-of-fit tests on the circle, *Biometrica*, 48: 109–114.

Watson, G. 1967. Another test for the uniformity of a circular distribution, *Biometrica*, 54: 675–677.

Watson, G. 1976. Optimal invariant tests for uniformity, *Studies in probability and statistics. Paper in honour of E.J.G. Pitman* 121–127.

Yamato, H. 1973. Uniform convergence of an estimator of a distribution function, *Bull. Math. Statist.*, (15): 69–78.

Zhang, P. 1999. Omnibus test of normality using the q statistic, *Journal of Applied Statistics*, 26: 519–528.

Zhu, L.-X.; Wong, H.; Fang, K.-T. 1995. A test for multivariate normality based on sample entropy and projection pursuit, *Journal of Statistical Planning and Inference*, 45: 373–385.

Zinger, A.; Klebanov, L.; Kakosyan, A. 1989. Characterization of distributions by mean values of statistics in connection with some probability metrics, *Stability Problems for Stochastic Models* 47–55.

List of Publications on the Topic of the Thesis

In the reviewed scientific periodical publications

Bakshaev, A. 2008. Nonparametric tests based on N-distances, *Lithuanian Mathematical Journal* 48(4): 368–379. ISSN 0363-1672 (ISI Master Journal List).

Bakshaev, A. 2009. Goodness of fit and homogeneity tests on the basis of N-distances, *Journal of Statistical Planning and Inference* 139(11): 3750–3758. ISSN 0378-3758 (ISI Master Journal List).

Bakshaev, A. 2010. N-distance tests for composite hypothesis of goodness of fit, *Lithuanian Mathematical Journal* 50(1): 14–34. ISSN 0363-1672 (ISI Master Journal List).

Bakshaev, A. 2010. N-distance tests for uniformity on the hypersphere, accepted for publication in *Nonlinear Analysis, Modelling and Control*. ISSN 1392-5113.

Aleksej BAKŠAJEV

STATISTICAL TESTS
BASED ON N-DISTANCES

Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Aleksej BAKŠAJEV

STATISTINIŲ HIPOTEZIŲ TIKRINIMAS,
NAUDOJANT N-METRIKAS

Daktaro disertacija
Fiziniai mokslai, matematika (01P)

2010 01 13. 12,0 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Baltijos kopija“,
Kareivių g. 13B, 09109 Vilnius, *http://www.kopija.lt*