GOODNESS OF FIT TESTS BASED ON KERNEL DENSITY ESTIMATORS

R. Rudzkis, A. Bakshaev
Vilnius University, Institute of Mathematics and Informatics
Akademijos 4, LT-08663 Vilnius, Lithuania
e-mail: rimantas.rudzkis@mii.vu.lt
e-mail: aleksej.bakshaev@gmail.com

Abstract

The paper is devoted to goodness of fit tests based on kernel estimators of probability density functions. In particular, univariate case is investigated. The test statistic is considered in the form of maximum of the normalized deviation of the estimate from its expected value. Produced comparative Monte Carlo power study show that the proposed test is a powerful competitor to the existing classical criteria testing goodness of fit against a specific type of alternative hypothesis. An analytical way for establishing the asymptotic distribution of the test statistic is proposed, using the theory of high excursions of Gaussian random processes and fields introduced by Rudzkis [17,18]. The extension of the proposed methods to the multivariate case are discussed.

1 Introduction.

The problem of testing goodness of fit is well known and has generated plenty of attention from researchers both in theoretical and applied statistical literature. In its typical purpose these tests are used to determine whether an underlying probability distribution differs from a hypothesized one. Both for one- and multidimensional cases a wide range of solutions of this problem have been provided. However, the choice of the most efficient test, among the available criteria, is regarded as one of the basic problems of statistics. 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, especially in multivariate case, sensitive to a particular type of hypotheses remains actual and in our days.

Classical approaches to solve the goodness of fit problem use the empirical process theory. Most of the popular tests such as the Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics are based on the empirical distribution function $F_n(x)$. In this paper, we consider another type of tests based on the kernel density estimator. The idea of using nonparametric kernel density estimators for goodness of fit tests goes back to Bickel and Rosenblatt [13,14]. Since that time a great number of publications has appeared, devote mostly to the L_p, $p = 1, 2$ distance between the density estimate $\widehat{f}(x) = \widehat{f}(x, X^n)$ of the underlying density $f_0(x)$ and its expected value under the null hypothesis. A review of the methods could be found in [1–7,10] and references therein. Thereby much less attention were devoted to a consideration of the deviations in the uniform metric as the loss function for $\widehat{f}(x)$, which is an object of this work.

2 Statement of the problem.

Let $X_1, ..., X_n$ be a sample of independent observations of a random variable X with an unknown probability density function $f(x)$, $x \in \mathbb{R}$. Using the given sample, it is required to test a simple hypothesis of goodness of fit

$$H_0 : f(x) = f_0(x)$$

against the complex alternative

$$H_1 : f(x) = (1 - \epsilon) f_0(x) + \epsilon g(x), \tag{1}$$

where $f_0(x)$ is a given probability density function, ϵ is small enough and $g(x)$ is an arbitrary distribution concentrated on a small interval, e.g. $\sigma_g^2 \ll \sigma_{f_0}^2$, where σ_f^2 is a variance of distribution f.
The choice of uniform metric for the loss function for $\widehat{f}(x)$ is justified by investigation of specific type of alternative hypothesis (1). Such alternatives are of a particular interest in some social and economic studies, e.g. determination of small high income clusters of people, in population income distribution. Meaningful applications could be also achieved in multivariate case, dealing with multimodal distributions for detection of tight clusters.
We consider a test based on the well-known Parzen-Rosenblatt kernel density estimator of f, defined for any $x \in \mathbb{R}$ by

$$\widehat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}), \tag{2}$$

where $K(\cdot)$ is a probability kernel and $h = h(n)$ is a bandwidth parameter.
The form of the alternative hypothesis motivates us to consider the test statistic in the form of maximum of the normalized deviation in the uniform metric of the estimate $\widehat{f}(x)$ from its expected value $\mathbb{E}_0 \widehat{f}(x)$

$$\zeta_h = \max_{x \in I} |\xi_h(x)|, \tag{3}$$

where $\xi_h(x) = \frac{\widehat{f}_h(x) - \mathbb{E}_0 \widehat{f}_h(x)}{\sqrt{\mathbb{D}_0 \widehat{f}_h(x)}}$, $\mathbb{E}_0$ and $\mathbb{D}_0$ denote a mathematical expectation and variance defined in the case of null hypothesis and I is a fixed interval.
Efficient use of kernel estimators requires the choice of an appropriate kernel and a bandwidth parameter. It is well-known that selection of the smoothing parameter rather than the form of the kernel is critical, as under- or over-smoothing can substantially reduce precision. In this work, a certain method to avoid the problem of selection of a bandwidth parameter is proposed. It is suggested to examine the test statistic ζ_h with different choices of a smoothing parameter h and thereby make the decision of rejecting the null hypothesis, based on the maximum of ζ_h values with respect to h. This leads us to the following improved form of the test statistic

$$M = \max_{h \in J} \left[\frac{\max_{x \in I} |\xi_h(x)| - \mu(h)}{\gamma(h)} \right], \tag{4}$$

where

$$\mu(h) = \mathbb{E}_0 \max_{x\in I} |\xi_h(x)| \quad \gamma^2(h) = \mathbb{D}_0 \max_{x\in I} |\xi_h(x)| \tag{5}$$

and maximum with respect to h is calculated in a certain interval J defined by a researcher.

We should reject the null hypothesis in the case of large values of our test statistics, that is if $M > c_\alpha$, where c_α can be found from the equation

$$\mathbb{P}_0(M > c_\alpha) = \alpha, \tag{6}$$

where $\mathbb{P}_0$ is a probability distribution corresponding to the null hypothesis and α is a pre-specified size of the test.

In practice the critical region of the test could be established by means of Monte Carlo simulations. The problem of analytical approximation of the distribution of the test statistics under the null hypothesis is discussed in section 3, using the theory of high excursions of Gaussian (and, in some sense, close to Gaussian) random processes and fields developed by Rudzkis [17,18]. Besides some of the already mentioned references, the asymptotic distributions of deviations of kernel density estimators in uniform metric were also considered in [8,9,11,12].

3 Analytical approximation of the null distribution of the test statistic.

This section is devoted to the analytical approximation of the functions $\mu(\cdot)$ and $\gamma(\cdot)$ in (5) and the null distributions of statistics (3) and (4) to determine the critical region of the tests. First we will be concerned with the asymptotics of the probability

$$P_h(u) = \mathbb{P}_0 \left\{ \max_{x\in I} |\xi_h(x)| < u \right\} \tag{7}$$

as $n \to \infty$. Note that $\widehat{f}_h$ is a consistent estimator, its finite dimensional distributions are asymptotically normal, and

$$cor(\widehat{f}_h(x_1), \widehat{f}_h(x_2)) \to 0,$$

if $x_1 \neq x_2$ and $n \to \infty$. The fact that $\widehat{f}_h$ is close to the Gaussian random process in a certain sense suggests us to apply the results from the theory of high excursions of Gaussian processes introduced in [16] to approximate the probability $P_h(u)$. Let $\xi(x)$ be a differentiable in the mean square sense Gaussian random process with zero mean, unit variance and continuous trajectories, and $\mu_i(x), i = 1, 2$ are smooth enough functions. Rudzkis has shown that, under some smoothness and regularity conditions, the probability $\mathbb{P}\{-\mu_1(x) < \xi(x) < \mu_2(x), x \in [a, b]\}$, could be approximated by

$$\mathbb{P}\{-\mu_1(x) < \xi(x) < \mu_2(x), x \in [a, b]\} \cong G(\mu_1, \mu_2), \tag{8}$$

where

$$G(\mu_1, \mu_2) = [\Phi(\mu_1(a)) + \Phi(\mu_2(a)) - 1] \exp\left\{ -\sum_{i=1}^{2} \int_a^b q(\mu_i(t))dt \right\}, \tag{9}$$

here

$$q(\mu_i(t)) = \phi(\mu_i(t)) \left[\beta(t)\phi\left(\frac{\mu_i'(t)}{\beta(t)}\right) - \mu_i'(t)\Phi\left(-\frac{\mu_i'(t)}{\beta(t)}\right)\right], \tag{10}$$

where $\Phi(\cdot)$ is a probability distribution function of the standard normal distribution, $\phi(x) = \Phi'(x)$ and $\beta^2(x) = \mathbb{D}\xi'(x)$.
Consider the empirical random process $\xi_h(x)$. Using approximation (8) we have

$$P_h(u) \cong [2\Phi(u) - 1] \exp\left\{-\exp(-u^2/2)/\pi \int_I \beta(z)dz\right\} =: \widehat{P_h}(u), \tag{11}$$

where

$$\beta^2(x) = \frac{\mathbb{D}_0\widehat{f_h'}(x)}{\sigma^2(x)} - \frac{[(\sigma^2(x))']^2}{4\sigma^4(x)}, \quad \sigma^2(x) = \mathbb{D}\widehat{f}_h(x). \tag{12}$$

For practical usage the exact expression for $\beta(x)$ could be found in [19].
From (11) it follows, that the functions $\mu(h)$ and $\gamma(h)$ could be approximated using the formulas

$$\mu(h) = \int u d\widehat{P_h}(u) \;\; \gamma^2(h) = \int u^2 d\widehat{P_h}(u) - \mu^2(h). \tag{13}$$

Produced simulation analysis of the accuracy of proposed approximations show that asymptotic distributions (11) provide a really good approximation to the null distribution of the statistic ζ_h even for small and moderate sample sizes. However practical experiments showed that the choice of the smoothing parameter could play a crucial role for the goodness of approximation especially for small sample sizes.
Coming back to the null distribution of statistic (4), for the approximations of the probability

$$P(u) = \mathbb{P}_0\left\{\max_{h\in J}\left[\frac{\max_{x\in I}|\xi_h(x)| - \mu(h)}{\gamma(h)}\right] < u\right\} \tag{14}$$

and determination of the critical region of test, the following obvious inequality could be used

$$P(u) \le \max_{h\in J} P_h(u). \tag{15}$$

Produced Monte Carlo simulations show that suggested estimate provide sufficiently good approximation for the null distribution of M statistic for small sizes of the test ($\alpha < 0.05$).

4 Multivariate case.

In this section a brief discussion about the extension of proposed testing procedures to multivariate case is presented. The precise investigations will be provided in our further research.
Since Pearson criteria, goodness of fit tests have been developed mostly for univariate distributions and, except for the case of multivariate normality, few references can be found in the literature about multivariate tests of fit. 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.

Therefore analytical establishment of the asymptotic distribution of test statistic may lead to substantial calculational difficulties. To extend the classical univariate goodness of fit tests, e.g. Kolmogorov-Smirnov, Cramer-von Mises, etc., to multivariate case usually initial sample is first transformed to p-dimensional cube. After that multivariate hypothesis of uniformity is verified. One of the most popular transformations is the Rosenblatt transformation introduced in [15]. However, this approach also have some disadvantages. One of them is lack of the uniqueness. The method is not invariant with respect to relabelling of the components of p-dimensional vector, which lead to a different Rosenblatt transformation and values of statistics. Another disadvantage is connected with the influence of any transformation to the power of the test, as transformation may significantly change the structure of alternative distribution. A generalization of the proposed approach for testing the null hypothesis against the alternative (1) to the multivariate case may help to avoid the mentioned problems.
Let now $x \in \mathbb{R}^p$ and I be a p-dimensional interval. It is worth noting, that after the replacement of univariate density estimate (2) with its multivariate analog

$$\widehat{f}_H(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K(H^{-1}(x - X_i)), \tag{16}$$

where $K(\cdot)$ is the kernel function, H is a smoothing $p \times p$ symmetric and positive definite matrix and $|H|$ its determinant, all the formulas (3) - (5) after corresponding changes remain valid. After that, approximation of the null distribution of $\max_{x \in I} |\xi_H(x)|$ and functions $\mu(H)$, $\gamma(H)$ could be obtained by the direct application of the methods of high excursions of Gaussian fields presented in [18].
It has been shown that if a differentiable (in the mean square sense) Gaussian random field $\{\eta(t), t \in T\}$ with $\mathbb{E}\eta(t) \equiv 0, \mathbb{D}\eta(t) \equiv 1$ and continuous trajectories defined on the p-dimensional interval $T \subset \mathbb{R}^p$ satisfies certain smoothness and regularity conditions, then $\mathbb{P}\{-v_1(t) < \eta(t) < v_2(t),\ t \in T\} \cong e^{-Q}$, as $\forall t \in T\ v_{1,2}(t) > \chi,\ \chi \to \infty$, where $v_{1,2}(\cdot)$ are smooth enough functions and Q is a certain constructive functional depending on $v_{1,2}$, T and the matrix function $R(t) = cov(\eta'(t), \eta'(t))$. Stated result leads to the following approximation of probability $P_H(u)$

$$P_H(u) = \mathbb{P}_0 \left\{ \max_{x \in I} |\xi_H(x)| < u \right\} \cong e^{-2Q(u)} =: \widehat{P_H}(u), \tag{17}$$

where Q depends on u, I and the matrix function $R(x) = cov(\xi'_H(x), \xi'_H(x))$. The exact expression of functional Q in the general form could be found in [18]. In bivariate case Q has the form

$$Q(u) = \frac{1}{2\pi}(1 - \Phi(u) + u\phi(u)) \int_I det(R)^{1/2} dx_1 dx_2 +$$
$$+ \frac{\phi(u)}{2\sqrt{2\pi}} \left[\int_{I_1} (R_{1,1}^{1/2}(y, a_2) + R_{1,1}^{1/2}(x_1, b_2)) dx_1 + \int_{I_2} (R_{2,2}^{1/2}(a_1, x_2) + R_{2,2}^{1/2}(b_1, x_2)) dx_2 \right],$$

where $R = R(x_1, x_2) = cov(\xi'_H(x), \xi'_H(x))$ is the covariance matrix of the bivariate random field $\xi'_H(x)$ with elements $R_{i,j} = R_{i,j}(x_1, x_2)$, $i, j = 1, 2$ and $I = I_1 \times I_2 =$

$[a_1, b_1] \times [a_2, b_2]$. Here by $\phi(\cdot, \cdot|R)$ we denote the probability density functions of the bivariate normal distribution $N(0, R)$ with covariance function R.
Finally, the approximations for the functions $\mu(H)$ and $\gamma(H)$ could be derived from (13) using $\widehat{P_H}(u)$ instead of $\widehat{P_h}(u)$.
Considered method for approximation of the null distribution of test statistic provide a straightforward way for establishment of the critical region of the test without any initial transformations.

5 Simulation study.

Let us switch to a short description of the comparative Monte Carlo power study in detail presented in [19]. The analyzed test M is compared with the classical criteria: Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov, Shapiro-Wilk, D′Agostino (for the case of normality test) and Bickel-Rosenblatt using the stated type of alternative hypothesis. In the study we restrict to the usage of Epanechnikov kernel K in (2), as a kernel optimal in the minimum variance sense. The functions $\mu(\cdot)$ and $\gamma(\cdot)$ in (4) are calculated using the obtained approximations (13). The critical region of the test is established by means of Monte Carlo simulations. Considered several variants of the tightness of distribution cluster g and mixing probabilities ϵ in (1), give us a wide range of departures from the null hypothesis and allow us to test the sensitivity of criteria to each of them.
The results of simulations show that the proposed test is a powerful competitor to the existing classical ones. For small sample size ($n = 200$) the proposed test performance is very similar to the Bickel-Rosenblatt criterion being more powerful in comparison with all the other tests. In general, the Bickel-Rosenblatt criterion, as a test also based on the kernel density estimator, is considered to be the main competitor in the study. Detecting a small tight distribution cluster, using kernel estimators in uniform metrics, implies strong and expectable tendencies in increasing of the comparative power of the proposed test, while either sample size is growing and / or mixing distribution g becomes more concentrated for all mixing probabilities. As a result for large sample size ($n = 1000$) M test is the most powerful in our comparative analysis for all considered variants of alternative hypothesis.

References

[1] Ahmad I.A., Cerrito P.B. (1993). Goodness of fit tests based on the L2-norm of multivariate probability density functions. *J. Nonparametr. Statist.*, Vol. **2**, pp. 169–181.

[2] Bowman A. W. (1992). Density based tests for goodness-of-fit normality. *J. Statist. Comput. Simul.*, Vol. **40**, pp. 1–13.

[3] Cao R., Lugosi G. (2005). Goodness of fit tests based on the kernel density estimator. *Board of the Foundation of the Scandinavian Journal of Statistics*, Vol. **32**, pp. 599–616.

[4] Fan Y. (1994). Testing the goodness of fit of a parametric density function by kernel method. *Econometric Theory*, Vol. **10**, pp. 316–356.

[5] Fan Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory*, Vol. **14**, pp. 604–621.

[6] Kim C., Hong C., Jeong M., Yang, M. (1997). Goodness-of-fit test for density estimation. *Communications in Statistics - Theory and Methods*, Vol. **26(11)**, pp. 2725–2741.

[7] Louani D. (2005). Uniform L1-Distance Large Deviations in Nonparametric Density estimation. *Sociedad de Estadistica e investigacion Operativa. Test*, Vol. **14(1)**, pp. 75–98.

[8] Muminov M. S. (2011). On limit distribution of maximal deviation of empirical distribution density and regression function I. *Theory of Probability and Its Applications*, Vol. **55(3)**, pp. 509–517.

[9] Muminov M. S. (2012). On limit distribution of maximal deviation of empirical distribution density and regression function II. *Theory of Probability and Its Applications*, Vol. **56(1)**, pp. 155–166.

[10] Nadaraya N., Babilua P., Sokhadze, G. (2009). On some goodness of fit tests based on kernel type Wolverton-Wagner estimates. *Bulletin of Georgian National Academy of Sciences*, Vol. **3(2)**, pp. 1–8.

[11] Konakov V.D., Piterbarg V. I. (1982). Rate of convergence of maximal deviation distributions for gaussian processes and empirical density functions I. *Theory of Probability and Its Applications*, Vol. **27(4)**, pp. 707–724.

[12] Konakov V.D., Piterbarg V. I. (1983). Rate of convergence of maximal deviation distributions for gaussian processes and empirical density functions II. *Theory of Probability and Its Applications*, Vol. **28(1)**, pp. 164–169.

[13] Bickel P.J., Rosenblatt M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics*, Vol. **1(6)**, pp. 1071–1095.

[14] Rosenblatt M. (1976). On the maximal deviation of k-dimensional density estimates. *The Annals of Probability*, Vol. **4(6)**, pp. 1009–1015.

[15] Rosenblatt M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, Vol. **23**, pp. 470–472.

[16] Rudzkis R. (1992). On the Distribution of Supremum-Type Functionals of Nonparametric Estimates of Probability and Spectral Densities. *Theory of Probability and Its Applications*, Vol. **37(2)**, pp. 236–249.

[17] Rudzkis R. (1992). Probabilities of large excursions of empirical processes and fields. *Soviet Math. Dokl.*, Vol. **45(1)**, pp. 226–228.

[18] Rudzkis R., Bakshaev A. (2012). Probabilities of high excursions of Gaussian fields. *Lithuanian Mathematical Journal*, Vol. **52(2)**, pp. 196–213.

[19] Rudzkis R., Bakshaev A. (2013). Goodness of fit tests based on kernel density estimators. *Informatica*, Vol. **24(1)**, pp. 1–14.