Optoelectronics, Instrumentation and Data Processing
No. 2, pp. 3-12, 2002

Avtometriya

UDC 519.2

METHODS FOR DATA PROCESSING

APPLICATION OF THE NONPARAMETRIC GOODNESS-OF-FIT TESTS TO TESTING NONPARAMETRIC MODEL ADEQUACY*

B.Yu.Lemeshko, S.N.Postovalov, and A.V.Frantsuzov

Novosibirsk

The possibility of applying the nonparametric Kolmogorov and also Mises ω^2 and Ω^2 goodness-of-fit tests to testing adequacy of the nonparametric models of distribution laws to observed data is shown. It is demonstrated that when using the nonparametric estimators the distributions of statistics of goodness-of-fit tests are affected by some factors determining the composite tested hypothesis H_0, such as a random variable distribution law corresponding to H_0, a kind of the kernel function, a sample size, and a method of estimating the fuzziness parameters.

INTRODUCTION

Creation of a measurement system at different stages of development and investigation, certification of measurement devices, and metrological support of a technological process are necessarily associated with the need for determining the measurement accuracy, with statistical analysis of observation results, and construction of models of error distribution laws. Unfortunately, measurement errors are not always described by the normal distribution law. Moreover, it is not always possible to select from multiple frequently applied distribution laws a parametric model describing adequately the observed variable. In this case, one tries to use some nonparametric model law, a nonparametric estimator.

Construction of a probabilistic model for some object includes normally two stages. One chooses the type of model and estimates if necessary parameters of the model at the first stage, and tests adequacy of the model to the observed data at the second stage. In parametric statistics, two basic types of problems correspond to these stages, namely, estimation of parameters and statistical hypothesis testing.

In the simplest situation, when we deal with the observed random variable, in the parametric approach at the first stage we hypothesize the kind of model of the distribution law, and using samples extracted from the general set estimate the model parameters. At the second stage, adequacy of the model to observed data is tested by means of Pearson's χ^2, Kolmogorov's, Mises ω^2, etc. goodness-of-fit tests.

* This research was supported by the Russian Foundation for Basic Research, Project no. 00-01-00913.

© 2002 by Allerton Press, Inc.
Authorization to photocopy individual items for internal or personal use, or the internal or personal use of specific clients, is granted by Allerton Press, Inc. for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of $50.00 per copy is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923.

Naturally, the limited set of parametric models used most frequently in practice does not always allow adequate description of actually existing random variables. Recent decades are characterized by active development of nonparametric statistics and more frequent applications of the nonparametric methods. The nonparametric methods are sometimes opposed with parametric ones. The opposition is usually accompanied by not always reasonable criticism of the parametric approach. In the course of this criticism it is overlooked that application of the nonparametric estimators of the observed distribution laws has "bottlenecks" either. For example, there is a problem concerned with the best choice of a fuzziness parameter (parameters) of the employed kernel estimators of density function, and it becomes especially acute in the case of a limited domain of the observed random variable and finite sample sizes. In the latter case the kernel estimator of the distribution function differs often substantially from the empirical distribution function at the boundaries of domain (on "tails" of distribution). The most important point is that the problems of testing adequacy of the nonparametric models are still an open question.

To our mind, too negative treatment of the parametric methods and opposition to the nonparametric ones by some authors is not justified: one should apply methods leading to the best results in a concrete situation. The parametric and nonparametric methods are not mutually exclusive, and the boundary between them starts blurring as the mathematical tools develop. For example, as soon as we begin speaking about the optimal choice of the fuzziness parameter in a nonparametric model, the principal distinction between such a model and a parametric model disappears. On the other hand, the nonparametric models and methods have some concrete merits determining growing interest in them in various applications.

At the moment, statisticians focuse maximum effort at finding of more accurate nonparametric estimators and investigation of properties of the estimators. Researchers do not pay attention to the problem of adequacy of the obtained nonparametric model of the law. Test of adequacy of the model is the final stage of statistical analysis, it follows after construction of a parametric or nonparametric model of the observed law and substantiates applicability of the model in a concrete application. The absence of a tool for testing the nonparametric model adequacy hinders wide application of the methods of nonparametric statistics in practice.

The goal of this paper is to investigate the possibility of testing adequacy of the nonparametric models with the use of nonparametric Kolmogorov's and Mises's ω^2 and Ω^2 goodness-of-fit tests.

Simple nonparametric density estimators. As nonparametric models, in this paper we consider nonparametric Rozenblat-Parzen density estimators [1] having the form

$$p_n(x)=\frac{1}{n\lambda_n}\sum_{i=1}^{n}\varphi\left(\frac{x-x_i}{\lambda_n}\right), \tag{1}$$

where x_i, $i=1,\ldots,n$, is the sample of observations of a one-dimensional continuous random variable, λ_n is the fuzziness parameter, $\varphi(u)$ is the bell-shaped (kernel) function satisfying the following regularity conditions:

$$\varphi(u)=\varphi(-u); \quad 0\le\varphi(u)\le\infty; \quad \int\varphi(u)du=1; \quad \int u^2\varphi(u)du=1;$$
$$\int u^m\varphi(u)du<\infty; \quad 0\le m<\infty. \tag{2}$$

Asymptotic properties of estimator (1), such as unbias, consistency, and convergence almost surely to the density $f(x)$, were analyzed in detail in [2–4]. It was shown, in particular, that the root-mean-square error of approximation of estimator (1), which is defined by the relationship

$$J=M\left\{\int[f(x)-p_n(x)]^2\,dx\right\}=M\left\{\|f(x)-p_n(x)\|^2\right\}, \tag{3}$$

depends substantially on the choice of the fuzziness parameter λ_n and is less dependent on the form of kernel function $\varphi(u)$.

In this investigation we use two kinds of kernel functions:

1) quadratic kernel function [4] having the best properties in minimizing the root-mean-square error of approximation (3) in the form

$$\varphi_1(u)=\begin{cases}\dfrac{3}{4\sqrt{5}}-\dfrac{3}{20\sqrt{5}}\,u^2, & \text{if } |u|\le\sqrt{5};\\ 0, & \text{if } |u|>\sqrt{5},\end{cases} \tag{4}$$

2) density function of the standard normal law

$$\varphi_2(u)=\frac{1}{\sqrt{2\pi}}\,e^{-u^2/2}. \tag{5}$$

The choice of the fuzziness parameter λ_n affects substantially the form of the nonparametric estimators of the density function. Also, the problem of finding the value of this parameter, which is optimal in a certain sense, exists everywhere. Proceeding from the condition of the minimum root-mean-square error of approximation (3), the optimal estimator of the fuzziness parameter takes on the form [4]:

$$\lambda^*=\left[\frac{\|\varphi\|^2}{n\|f''\|^2}\right]^{1/5}. \tag{6}$$

The drawback of estimator (6) is that for its determination we must know the density $f(x)$ of the true random variable distribution law that is, generally speaking, unknown.

For $n\to\infty$ expression (6) tends to $n^{-1/5}$. Hence, it is sometimes proposed to choose the fuzziness parameter equal to:

$$\tilde{\lambda}_n'=n^{-1/5}. \tag{7}$$

The fuzziness parameter (or parameters) can be chosen by various methods on a basis of different tests of optimality [5], e.g., using different measures of proximity of the empirical distribution function and its nonparametric estimator. In this case, however, the advantages of the nonparametric estimators over the parametric models are lost.

Investigation of behavior of distributions of statistics for the nonparametric goodness-of-fit tests in nonparametric estimation. Adequacy of the parametric model of the distribution law to the observed data is most frequently tested using Pearson's χ^2 goodness-of-fit tests or nonparametric Kolmogorov's and Mises ω^2 and Ω^2 tests.

Using the goodness-of-fit tests, one should distinguish simple and composite hypotheses. The tested simple hypothesis has the form H_0:$F(x) = F(x, \theta)$, where $F(x, \theta)$ is the probability distribution function to which the observed sample is tested for goodness of fit, and θ is the known value of the parameter (scalar or vector parameter). The composite tested hypothesis is written in the form H_0: $F(x)\in\{F(x,\theta),\ \theta\in\Theta\}$. We deal with the composite hypothesis in the case if it is tested by the same sample as that used to estimate the distribution law parameters.

As a distance between empirical and theoretical laws the Kolmogorov test uses the variable

$$D_n=\sup_{|x|<\infty}|F_n(x)-F(x,\theta)|,$$

where $F_n(x)$ is the empirical distribution function, and n is the sample size. Researchers use usually statistic of the form [6]:

$$S_K = \frac{6nD_n + 1}{6\sqrt{n}}, \tag{8}$$

where

$$D_n = \max(D_n^+, D_n^-); \quad D_n^+ = \max_{1 \le i \le n}\left\{\frac{i}{n} - F(x_i, \theta)\right\}; \quad D_n^- = \max_{1 \le i \le n}\left\{F(x_i, \theta) - \frac{i-1}{n}\right\};$$

n is the sample size; $x_1, x_2, \ldots, x_n$ are sample values in increasing order. The distribution of S_K in testing the simple hypothesis in the limit obeys the Kolmogorov law $K(S)$ [6].

In tests of the ω^2 type the distance between the hypothetical and true distributions is considered in the quadratic metric

$$\int_{-\infty}^{\infty} \{E[F_n(x)] - F(x)\}^2 \psi(F(x)) dF(x),$$

where $E[\cdot]$ is the mathematical expectation operator.

In choosing $\psi(t) \equiv 1$ in the Mises ω^2 tests, one uses statistic (Kramer-Mises-Smirnov statistic) of the form

$$S_\omega = n\omega_n^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left\{F(x_i, \theta) - \frac{2i-1}{2n}\right\}^2, \tag{9}$$

which obeys the distribution $a_1(S)$ [6] in testing a simple hypothesis.

In choosing $\psi(t) \equiv 1/t(1-t)$ in Mises's Ω^2 tests, the statistic (Anderson-Darling statistic) has the form

$$S_\Omega = n\Omega_n^2 = -n - 2\sum_{i=1}^{n}\left\{\frac{2i-1}{2n}\ln F(x_i, \theta) + \left(1 - \frac{2i-1}{2n}\right)\ln(1 - F(x_i, \theta))\right\}. \tag{10}$$

In the limit, this statistic obeys the distribution $a_2(S)$ [6].

In the case of simple hypotheses, the limiting statistic distributions of the nonparametric Kolmogorov, Mises ω^2 and Ω^2 tests do not depend on the kind of the distribution law observed and on its parameters. That is why these tests are called nonparametric tests and said to be "distribution-free" tests. However, in testing composite hypotheses the "distribution-free" property is lost [7]. It has been found that the conditional distribution laws of statistics $G(S \mid H_0)$ of nonparametric goodness-of-fit tests are affected by a number of factors determining "complexity" of the hypothesis: the type of the observed law $F(x, \theta)$ corresponding to the true hypothesis H_0; the type of parameter estimated and the number of parameters to be estimated; sometimes, a concrete value of the parameter (e.g., in the case of gamma distribution); the method used to estimate the parameters [8–11].

In this paper we present results on investigating the possibility of applying the nonparametric goodness-of-fit tests to test adequacy of the nonparametric models of distribution laws (nonparametric estimators) began in [12]. For this purpose, using the method of computer simulation of statistical regularities [8–11], we investigated: dependences of the distribution of statistics of the previously mentioned goodness-of-fit tests on the type of kernel functions; variations of the form of these distributions with growing sample size; dependences of statistic distribution on the law corresponding to the tested hypothesis H_0; influence of the fuzziness parameter on the distributions of estimation statistics.

In the nonparametric approach as well as in the parametric one we meet testing of simple and composite hypotheses. Let us assume that by some *earlier* observed sample $x_1, x_2, \ldots, x_n$ we obtained a nonparametric density estimator of the form (1). The values $\theta_1 = x_1$, $\theta_2 = x_2, \ldots, \theta_n = x_n$ may be interpreted as parameters of this model

$$p_n(x) = \frac{1}{n\lambda_n} \sum_{i=1}^{n} \varphi\left(\frac{x - \theta_i}{\lambda_n}\right). \tag{11}$$

If adequacy of the nonparametric estimator is tested by a *new* sample whose size is not necessarily the same, then it is obvious that we will deal with a simple tested hypothesis. This is a classical case in which the statistics of the Kolmogorov and Mises ω^2 and Ω^2 tests under validity of the tested hypothesis H_0 must obey the distributions $K(S)$, $a_1(S)$, and $a_2(S)$, respectively.

In order to see that in the case of a simple hypothesis the distributions $K(S)$, $a_1(S)$, and $a_2(S)$ are really the limiting distributions of statistics $G(S \mid H_0)$, we simulated the random variable samples corresponding to hypothesis H_0 according to (11) by the method of inverse functions. Our simulation has verified that the obtained empirical distributions of statistics of the tests under investigation are in good agreement with the classical limiting laws $K(S)$, $a_1(S)$, and $a_2(S)$. For example, Fig. 1 represents results of simulating a distribution of the Kolmogorov statistic for the size of samples of the simulated random variables $n = 50$ and the number of such samples $N = 500$. The figure shows results on testing goodness-of-fit of the obtained empirical distribution to the distribution $K(S)$ by Pearson's χ^2, likelihood ratio, Kolmogorov's, and Mises ω^2 and Ω^2 tests. In what follows we present for each test the attained level of significance $P\{S > S^*\} = \int_{S^*}^{\infty} g(s \mid H_0)\,ds$, where S^* is the value of the test statistic calculated by the sample; $g(s \mid H_0)$ is the limiting distribution density of statistic of the respective test under validity of hypothesis H_0. ($P\{S > S^*\} = 0.37264$ for the likelihood ratio test; 0.35344 for the Pearson χ^2 test; 0.14299 for the Kolmogorov test; 0.45431 for the Mises ω^2 test; 0.57374 for the Mises Ω^2 test.)

In the parametric approach we meet a composite hypothesis if testing is preceded by estimation of the model parameters by the same sample. For a nonparametric model, such a situation in a habitual understanding

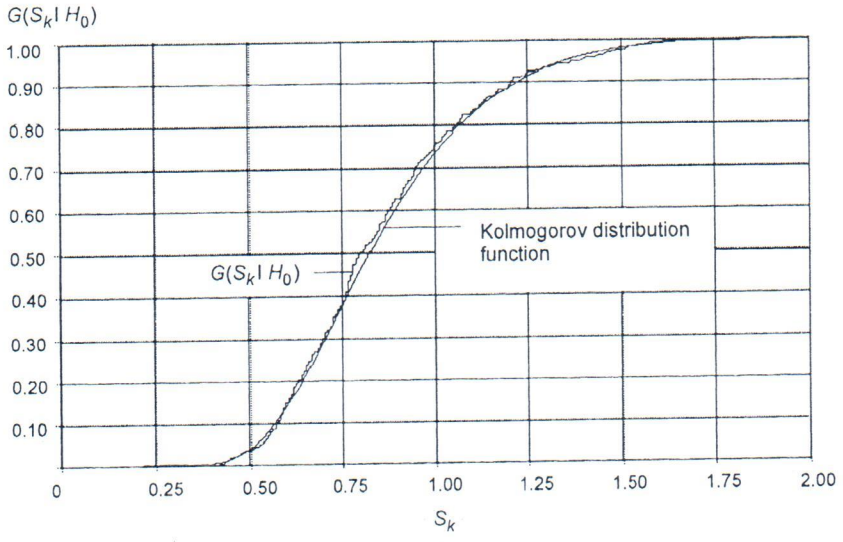

Fig. 1. The Kolmogorov statistic distribution for a simple hypothesis.

is consistent with the case of estimation by a concrete sample of the fuzziness parameter (or parameters). On the other hand, in constructing model (1) by the sample, its parameters are estimated implicitly $\theta_1 = x_1$, $\theta_2 = x_2$, ..., $\theta_n = x_n$ (see the model of the form (11)). Testing goodness-of-fit to this model by the same sample, even without estimating the fuzziness parameter, generally speaking, we are dealing with a composite hypothesis, too. In order to set apart this situation met most frequently in practice and underline that in this case no parameter estimation by any method occurs, we will call it a test of a "quasi-composite" hypothesis.

As in the parametric case, in testing composite (quasi-composite) hypotheses the distributions of statistics of Kolmogorov's and Mises's ω^2 and Ω^2 tests depend on the true distribution law corresponding to hypothesis H_0 [7–10]. Figure 2 (*a*)–(*c*) that represents results of simulating the statistic distributions of Kolmogorov's and Mises's ω^2 and Ω^2 tests in testing "quasi-composite" hypotheses H_0 demonstrates dependence of the statistic distributions on the form of the true law corresponding to H_0. When simulating, we chose parameters $\theta_1 = x_1$, $\theta_2 = x_2$, ..., $\theta_n = x_n$ of model (11) corresponding to the true hypothesis H_0 such that the model is close to one of the following parametric models: the exponential law with the density

$$\exp(\mu,\sigma)=\frac{1}{\sigma}\exp\left\{-\frac{x-\mu}{\sigma}\right\},$$

the Cauchy distribution law with the density

$$\text{Cauchy}(\mu,\sigma)=\frac{\sigma}{\pi[\sigma^2+(x-\mu)^2]},$$

or the logistic distribution with the density

$$\log(\mu,\sigma)=\frac{\pi}{\sigma\sqrt{3}}\exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\}\Bigg/\left[1+\exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\}\right]^2.$$

Samples of random variables of the size $n = 50$ were simulated by (11). To form a sample of statistics we repeated our experiment $N = 500$. Obtained results evidently confirm that in testing the quasi-composite hypotheses the distributions of statistics depend substantially on the true law corresponding to H_0 and differ drastically from the limiting Kolmogorov distributions, $a_1(S)$ and $a_2(S)$, in the case of simple hypotheses.

In the parametric case, the distributions of statistics $G(S \mid H_0)$ of the nonparametric Kolmogorov and Mises ω^2 and Ω^2 goodness-of-fit tests depend on the sample size n, but with growing n they quickly converge to limiting ones for both the simple and composite hypotheses. Starting from the sample size $n \geq 15$–20 in testing simple hypotheses and $n \geq 20$–25 in testing composite hypotheses, one may use the limiting distributions of statistics [8].

In the case of the nonparametric models in testing composite (quasi-composite) hypotheses we see a stronger dependence on n and a slower convergence of statistic distributions to limiting ones. Figure 3 shows behavior of statistic distributions of the Mises ω^2 test under the "quasi-composite" hypothesis, depending on the size of random variable sample. In this case, when simulating we chose the parameters $\theta_1 = x_1$, $\theta_2 = x_2$, ..., $\theta_n = x_n$ of model (1) corresponding to the true hypothesis H_0 in such a manner that it is close to the logistic law. The statistic distributions of Kolmogorov and Mises Ω^2 tests change similarly with growing n.

The type of employed kernel functions also affects statistic distributions of the goodness-of-fit tests. For example, Fig. 4 illustrates dependence of the statistic distributions of the Kolmogorov test for the same "quasi-composite" hypothesis (close to the logistic law) and $n = 50$ on the type of employed kernel functions (4) and (5) that have a similar effect on behavior of the Mises ω^2 and Ω^2 statistic distributions. Investigation has shown that with growing sample size the difference in the statistic distributions of the goodness-of-fit tests under consideration for different kernel functions becomes more essential.

In testing composite hypotheses, when we estimate the fuzziness parameter by the observed sample, the statistic distributions of the goodness-of-fit tests under consideration depend on the same factors as in the case of

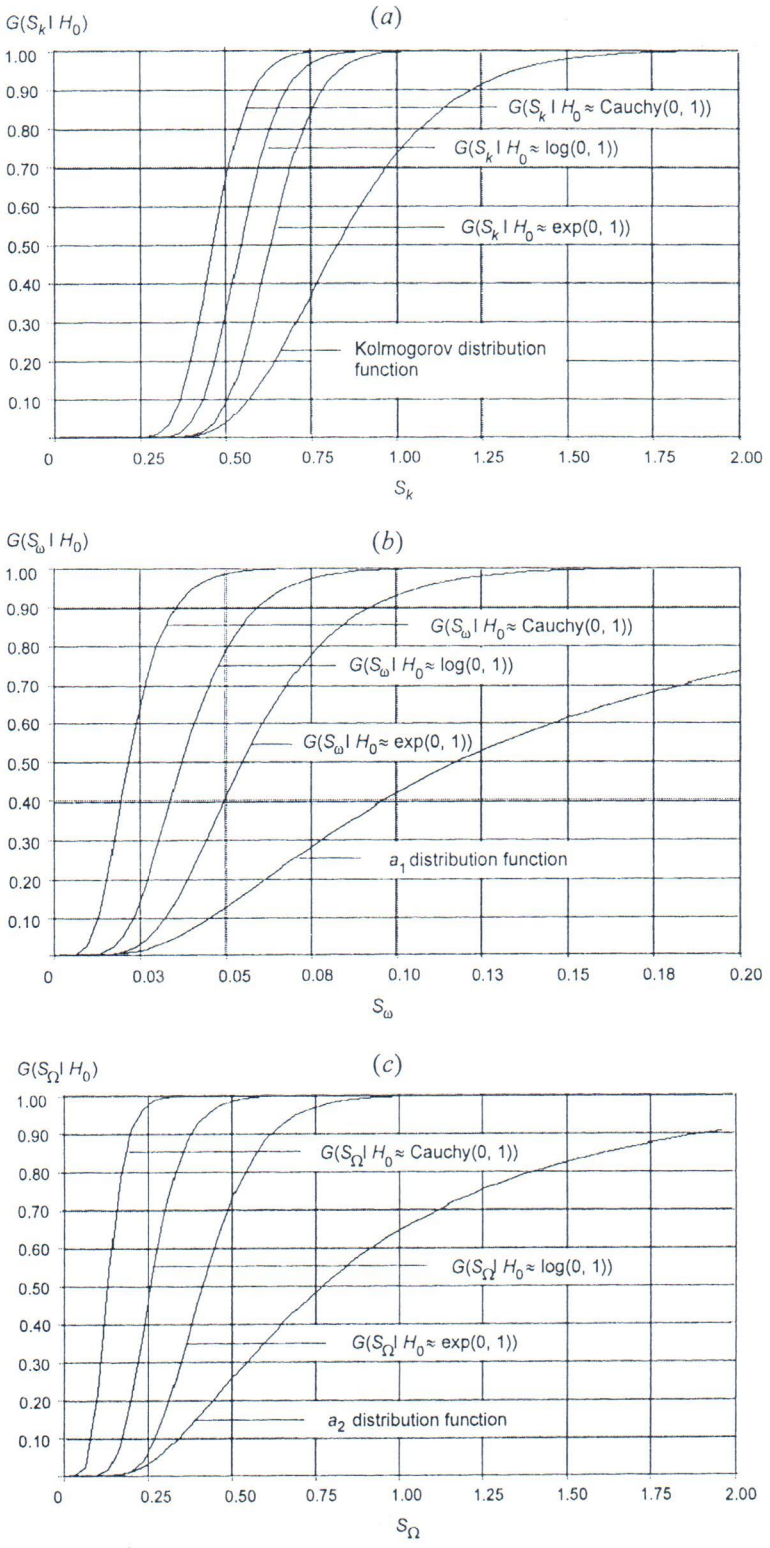

Fig. 2. Kolmogorov's (*a*), Mises's ω^2 (*b*) and Ω^2 (*c*) statistic distributions under the "quasi-composite" hypothesis, depending on the true law corresponding to hypothesis H_0.

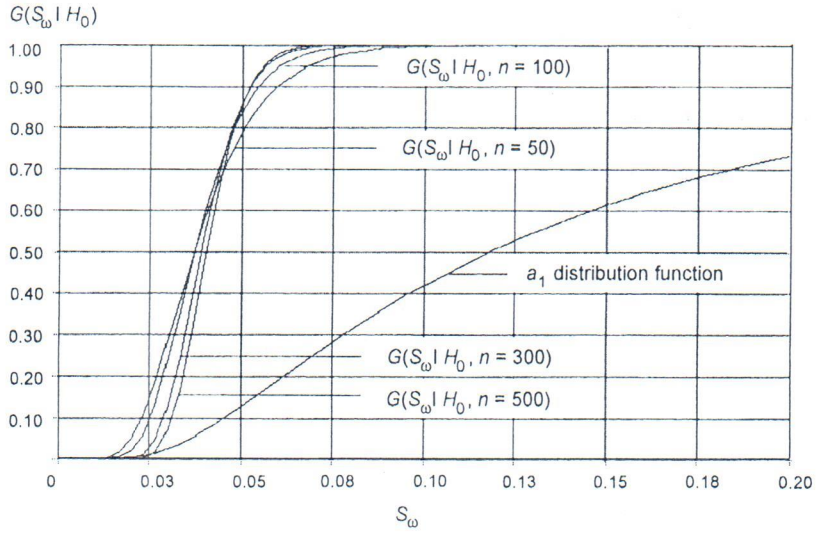

Fig. 3. The behavior of Mises ω^2 statistic distributions in testing the "quasi-composite" hypothesis, depending on the sample size n.

the "quasi-composite" hypothesis and also on the method of estimating the fuzziness parameter. For example, Fig. 5 shows statistic distributions of the Kolmogorov goodness-of-fit test under a composite hypothesis (close to the logistic law), $n = 50$, using different estimators of the fuzziness parameter, calculated according to (6) and (7). To compare, the same figure represents Kolmogorov statistic distributions in the case of testing the composite hypothesis on fitness to the logistic law (parametric model) with simultaneous calculation of maximum likelihood estimators (MLE) of two parameters of this law and with the use of MD (minimum distance) estimators ob-

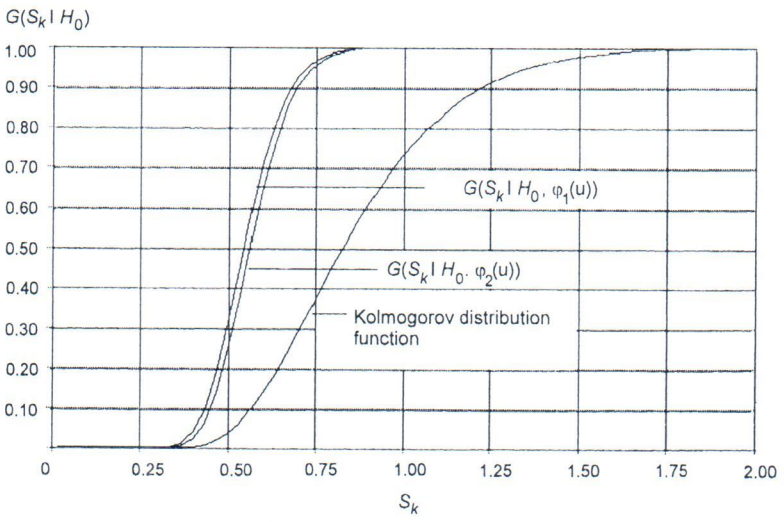

Fig. 4. Effect of the type of kernel function on the Kolmogorov statistic distributions in testing the "quasi-composite" hypothesis.

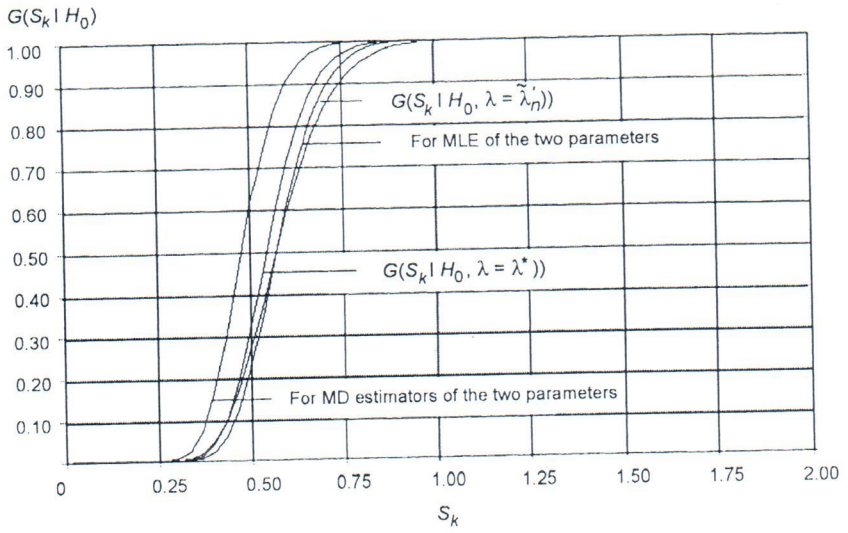

Fig. 5. Effect of the method of fuzziness parameter estimation on the Kolmogorov statistic distributions in testing a composite hypothesis.

tained by minimization of statistic of the corresponding goodness-of-fit test [9–11]. The choice of a fuzziness parameter estimator affects similarly the change in statistic distributions of the Mises ω^2 and Ω^2 tests.

CONCLUSION

Testing of adequacy of the nonparametric models used to describe measurement errors is, of course, necessary. It is obvious that this testing can be performed by using goodness-to-fit tests on the basis of the applied method.

In testing the adequacy of a nonparametric model we deal with a simple verified hypothesis only in the case if construction of the estimator and testing for fitness are performed by different samples or by different parts of a sample. Results of statistical simulation verified that in such situations the testing procedures must be based on classical results on the limiting statistic distributions of the Kolmogorov and Mises ω^2 and Ω^2 tests: the Kolmogorov distributions, $a_1(S)$ and $a_2(S)$, respectively.

Testing of composite hypotheses on goodness-of-fit by the Kolmogorov and Mises ω^2 and Ω^2 tests with the use of nonparametric models compared with application of parametric models is characterized by a great variety of factors determining "complexity" of the hypothesis. Statistic distributions of the tests under consideration are affected substantially by: the true distribution law of the observed random variable, which corresponds to the tested hypothesis H_0; the kind of the kernel function used; the sample size; the method of estimating (kind of estimator) the fuzziness parameter (or parameters). Compared with testing fitness with parametric models, we must underline specifically the greater dependence of the statistic distributions on the sample size. This is explained by the fact that each new sample element used in the nonparametric estimator is an additional "estimated" parameter of the model. It is exactly this fact that is the principal distinction of the problem of testing composite goodness-of-fit hypotheses with the use of the parametric models.

Our investigation has shown the possibility of using the nonparametric goodness-of-fit tests for testing adequacy of the nonparametric models of distribution laws with simple and composite hypotheses, and the possibility of constructing the models of statistic distributions of goodness-of-fit tests for different tested composite hypotheses.

Obviously, the variety of composite hypotheses is so great that it is impossible to construct beforehand the models of statistic distributions of the goodness-of-fit tests for each concrete type of composite hypothesis specified by the kind of nonparametric estimator. However, using the employed computer simulation method, for a concrete composite tested hypothesis (the law corresponding to H_0; a concrete type of nonparametric estimator; sample size; method for estimating the fuzziness parameter), one can always construct a model of statistic distribution of the goodness-of-fit test used and, therefore, ensure adequate testing of the nonparametric model.

REFERENCES

1. E.Parzen, *Ann. Math. Statist.*, vol. 33, p. 1065, 1962.
2. E.A.Nadaraya, *Soobshch. AN GSSR*, vol. 34, no. 2, p. 277, 1964.
3. E.A.Nadaraya, *Nonparametric Estimation of Probability Density and Regression Curve* (in Russian), Izd-vo TGU, Tbilisi, 1983.
4. V.A.Epanechnikov, *Teoriya Veroyatnostei i ee Primenenie*, vol. 14, no. 1, p. 156, 1969.
5. A.V.Lapko and S.V.Chentsov, *Nonparametric Data Processing Systems* (in Russian), Nauka, Moscow, 2000.
6. L.N.Bolshev and N.V.Smirnov, *Tables of Mathematical Statistics* (in Russian), Nauka, Moscow, 1983.
7. M.Kac, J.Kiefer, and J.Wolfowitz, *Ann. Math. Statist.*, vol. 26, p. 189, 1955.
8. B.Yu.Lemeshko and S.N.Postovalov, *Zavod. Lab.*, vol. 64, no. 3, p. 61, 1998.
9. B.Yu.Lemeshko and S.N.Postovalov, *Applied Statistics. Rules for Testing Fitness between Experimental and Theoretical Distributions: Methodical Advises*, part II, Izd-vo NGTU, Novosibirsk, 1999.
10. B.Yu.Lemeshko and S.N.Postovalov, *Optoelectr., Instrum. and Data Process.*, no. 2, p. 76, 2001.
11. B.Yu.Lemeshko and S.N.Postovalov, *Zavod. Lab. Diagnostika Materialov*, vol. 67, no. 7, p. 62, 2001.
12. B.Yu.Lemeshko, S.N.Postovalov, and A.V.Frantsuzov, *Proc. of Internat. Scient.-Techn. Conf. on Information Technologies and Problems of Telecommunications*, SibGUTI, Novosibirsk, 2001, p. 82.

The Novosibirsk State Technical University

E-mail: headrd@fpm.ami.nstu.ru

23 October 2001