ON THE USAGE OF GOODNESS-OF-FIT CRITERIONS FOR TESTING ADEQUACY OF NONPARAMETRIC ESTIMATES OF THE DISTRIBUTION LAWS[1]

A.V. Frantsuzov, B.Yu. Lemeshko
Novosibirsk State Technical University
Novosibirsk, Russia
e-mail: `headrd@fpm.ami.nstu.ru`

Abstract

The distributions of Kolmogorov, ω^2 and Ω^2 Mises, χ^2 Pearson test statistic, using for testing adequacy of nonparametric models of distribution laws, have been investigated by methods of statistical simulation. It has been shown, that using of nonparametric estimates makes goodness-of-fit test statistic distributions be dependent on a number of factors. The statistic distributions of goodness-of-fit tests differs from the limiting laws taking place in a parametric case.

1 Introduction

Lately different nonparametric estimates are frequently used as distribution models of random variables under observation. It results from parametric model describing observed random variable in the best way is not always known. A lot of papers are devoted to nonparametric estimation, but the problem of obtained nonparametric model adequacy testing has not been considered enough yet. The goal of this paper is to investigate the possibility of testing adequacy of the nonparametric models with the use of different goodness-of-fit tests, such as Kolmogorov's, Mises's ω^2 and Ω^2 and Pearson χ^2 tests.

2 The nonparametric estimates of density

As nonparametric models, in this paper we consider nonparametric Rozenblat-Parzen density estimates [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 $x_i,\ \ i = 1, ..., n$ is the sample of observations of a one-dimensional continuous random value, λ_n is the fuzziness parameter, and $\varphi(u)$ is the bell-shaped (kernel) function satisfying the following regularity conditions:

$$\varphi(u) = \varphi(-u); \quad 0 \le \varphi(u) \le \infty; \quad \int \varphi(u)du = 1; \quad \int u^2\varphi(u)du = 1;$$

[1]The research is supported by the Ministry of Education of Russian Federation (project No. T02-3.3-3356)

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

Two kinds of kernel function have been used within the investigation:

1. quadratic kernel function [2], having the best properties in minimizing the root-mean-square error of approximation

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

1. density function of the standard normal law

$$\varphi_2(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{4}$$

3 Statistic distributions of nonparametric goodness-of-fit tests for nonparametric estimation

It doesn't matter, how considered law is specified in a case of simple hypothesis: as some parametric distribution $F(x,\theta)$ or as nonparametric estimator (1). In both cases the limiting distributions of considered statistics are the corresponding classical distributions. It should be pointed out that in adequacy testing of nonparametric model (1) we have a simple hypothesis under test if the model is obtained from some other sample.

In case of composite hypotheses considered test statistic distributions depend on the true distribution law corresponding to H_0 hypothesis as in parametric case. Figure 1 illustrates this dependence for the Kolmogorov statistic. Observation samples of size $n = 50$ have been simulated by following laws: 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 $Cauchy(\mu,\sigma) = \frac{\sigma}{\pi[\sigma^2+(x-\mu)^2]}$, and the logistic distribution with the density $Log(\mu,\sigma) = \frac{\pi}{\sigma\sqrt{3}}\exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\}\Big/\left[1+\exp\left\{-\frac{\pi(x-\mu)}{\sigma\sqrt{3}}\right\}\right]^2$. In this case, kernel functions (3) has been used. The obtained results confirm that in composite hypotheses testing the statistic distributions essentially depend on the true law corresponding to H_0 hypothesis under test. Similar regularities have been obtained for all test statistics.

In the parametric case, the distributions of statistics $G(S\,|H_0)$ 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. Our investigations have shown that in the case of the nonparametric models for testing composite hypotheses we have a stronger dependence on n and a slower convergence of statistic distributions to limiting ones.

Investigations have shown that the type of employed kernel functions also influence statistic distributions of the goodness-of-fit tests and with growing of the sample size the difference becomes more essential [3]. The distributions of statistics $G(S\,|H_0)$ also depend on the method of fuzziness parameter λ_n estimation.

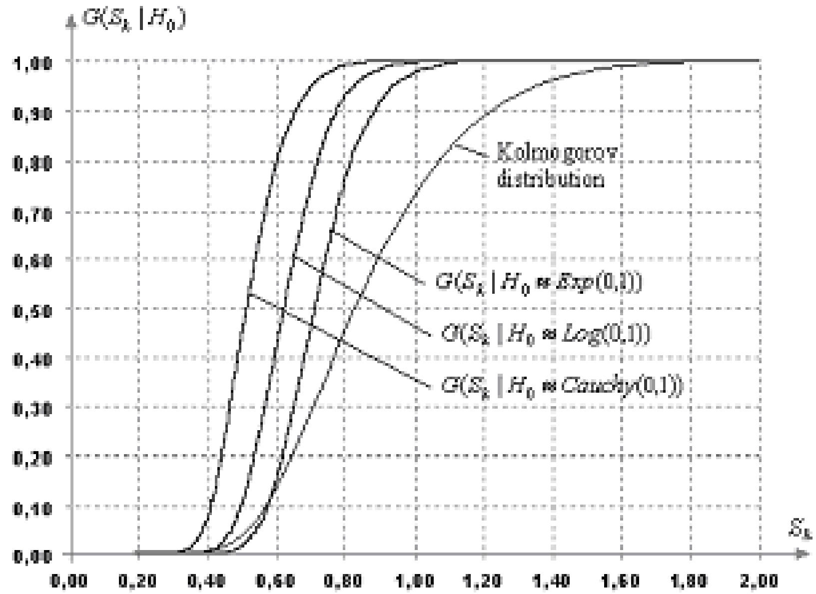

Figure 1: The Kolmogorov statistic distributions depending on the true law corresponding to composite hypothesis H_0

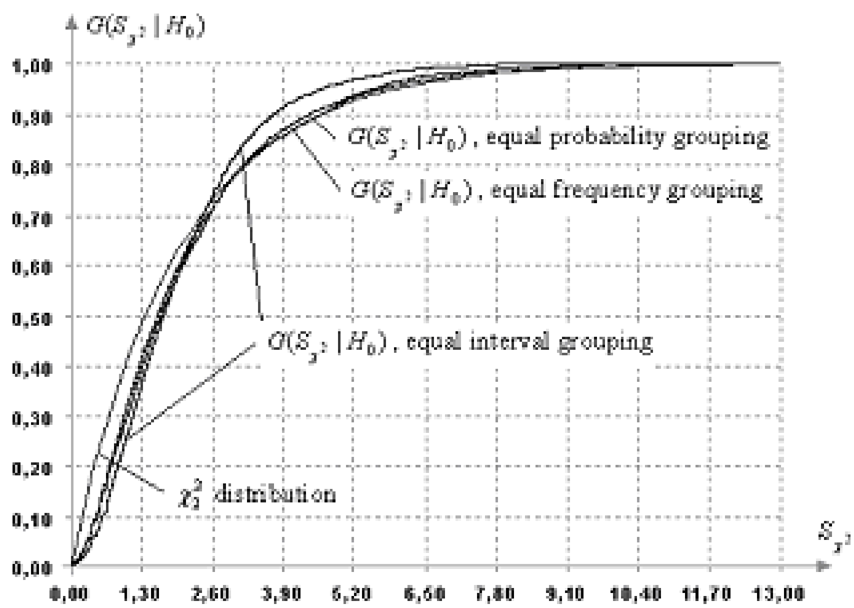

Figure 2: The dependence of χ^2 statistic distribution on the grouping type for composite hypothesis

4 The statistic distributions of χ^2 goodness-of-fit tests for nonparametric estimation

The statistic distributions of the χ^2 goodness-of-fit tests have been shown to be dependent on the same factors as the statistic distributions of nonparametric goodness-of-fit criteria for composite hypotheses. Moreover they depend on the grouping method. And obtained statistic distributions considerably differ from the limiting χ^2_{k-r-1}-distributions taking place in case of testing of goodness-of-fit to parametric models and parameter estimating by grouped data.

Figure 2 shows the influence of grouping kind on Pearsons χ^2 statistic distributions. In this case, samples of random variables of the size $n = 100$ have been simulated by normal law, 5 intervals of grouping and kernel functions (3) have been use.

5 Conclusions

In testing adequacy of nonparametric models we deal with a simple hypothesis only in the case when parameter estimating and goodness-of-fit testing are performed by different samples or by different parts of a sample. In such situations testing procedure must base on the classical limiting test statistic distributions and it doesn't depend on the observed law.

Testing of composite hypotheses 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 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 estimates) the fuzziness parameter (or parameters). Pearson χ^2 statistic distributions also depend on the grouping type with the number of intervals given. Statistic distributions poorly converge to the limiting ones, which depend on the true law.

Using computer simulation and analyzing the obtained empirical distributions, it is possible to construct models of limiting statistic distributions of goodness-of-fit tests for different tested composite hypotheses.

References

[1] Parzen E. (1962). On the estimation of probability density function and the mode. *Ann. Math. Stat.* Vol. **33**, pp. 1065-1076.

[2] Epanechnikov V.A. (1969). Nonparametric Estimation of Multi-dimensional Probability Density. *Teoriya Veroyatnostei i ee Primenenie*. Vol. **14**, N1, pp. 156-161.

[3] Lemeshko B.Yu., Postovalov S.N., Frantsuzov A.V. (2002). Application of the Nonparametric Goodness-of-fit Tests to Testing Nonparametric Model Adequacy. *Optoelectronics, Instrumentation and Data Processing*. N2, pp. 3-12 (in Russian).