Goodness-of-fit tests via minimum L_1-distance kernel density estimation

Ricardo Cao
Department of Mathematics, Universidade da Corua
Campus de Elvia, s/n
15071 A Corua, Spain
rcao@udc.es

Gabor Lugosi
Department of Economics, Universitat Pompeu Fabra
C/ Ramn Tras Fargas, 25-27
08005 Barcelona, Spain
lugosi@upf.es

1. Introduction

Given a class of densities $\mathcal{F}$ on $\mathbb{R}$ and a sample of n i.i.d. random variables $X_1, \ldots, X_n$ drawn from an unknown density f, the problem is to decide whether the null hypothesis $H_0 : f \in \mathcal{F}$ is true or not. To do this we use goodness-of-fit tests of level $\alpha \in (0,1)$.

The tests we propose are based on the kernel density estimate. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function with $\int K = 1$. For convenience we assume that K is nonnegative. For any smoothing factor $h > 0$, the kernel density estimate $f_{n,h}$ is defined as $f_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$, where $K_h(\cdot) = (1/h)K(\cdot/h)$ (see Akaike (1954), Parzen (1962) and Rosenblatt (1956)). The composite goodness-of-fit tests we investigate in this paper all have the form: accept H_0 if and only if $T_n \leq c_\alpha$, where c_α is an appropriate constant and the test statistics T_n has the form $T_n = \inf_{g\in\mathcal{F}} \int |f_{n,h} - g|$. Different versions of these tests differ in their different choices of the smoothing factor h and the constant c_α. Both data-dependent and data-independent choices of h can be considered.

2. Main results

First we investigate tests based on the test statistics $T_n = \inf_{h>0} \inf_{g\in\mathcal{F}} \int |f_{n,h} - g|$. Recall that this coresponds to the data-dependent smoothing factor is

$$H \stackrel{\text{def}}{=} \arg\min_{h>0} \inf_{g\in\mathcal{F}} \int |f_{n,h} - g|.$$

Then we may compute

$$c_\alpha = \inf\left\{ c : \sup_{f\in\mathcal{F}} \mathbb{P}_f\left[T_n > c\right] \leq \alpha \right\}.$$

This constant may be approximated by arbitrary precision using Monte-Carlo simulations. The definition of c_α immediately guarantees that if the data are indeed drawn from a density in $\mathcal{F}$, then the rejection probability is bounded by α, as desired. It remains to see how the test behaves when $f \notin \mathcal{F}$. We prove the following bound on the probability of acceptance:

Theorem 1 *Let* $b_n \stackrel{\text{def}}{=} \sup_{f\in\mathcal{F}} \inf_{h>0} \mathbb{E}_f \int |f_{n,h} - f|$. *Then for any density* f *and* $\delta > 0$, *if*

$$\mathbb{E}_f T_n > b_n + \delta + \sqrt{\frac{2}{n}\log\frac{1}{\alpha}},$$

then

$$\mathbb{P}_f\left[H_0 \text{ is accepted}\right] \leq e^{-n\delta^2/2} .$$

Using this result and the following theorem, the type II error probability can be controled.

Theorem 2 *Assume that* $f \notin \mathcal{F}$ *is such that* $H \to 0$ *and* $nH \to \infty$ *almost surely as* $n \to \infty$. *Then* $\liminf_{n\to\infty} \mathbb{E}_f T_n > 0$.

Different simulation studies were carried out to compare the test with the one proposed by Fan (1994), as well as with the Kolmogorov-Smirnov test, when testing for normality. For the null hypothesis scenario we used a standard normal distribution, while we considered six distributions of the lambda family as well as some normal mixture distributions for the alternative. The simulations results confirm that, in general, the minimum L_1-distance method performs better than Fan's test and also slightly better than Kolmogorov-Smirnov test.

REFERENCES

Akaike, H. (1954). An approximation to the density function. *Ann. Inst. Statist. Math.*, **6**, 127-132.

Fan, Y. (1994). Testing the goodness of fit of a parametric density function by kernel method. *Econometric Theory*, **10**, 316-356.

Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Statist.*, **33**, 1065-1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, **27** , 832-837.

RESUME

Soit donnée une échantillon aléatoire également distribue qui est obtenue à partir d'une densité f sur la droite réelle. On essaie de vérifier si f appartien a une classe donnée de densités. Nous cherchons des procédures test qui sont construits par minimisation de la distance L_1 entre l'estimateur du type noyau de la densité et l'ensemble des densités admisibles. Plusieurs études de simulation sont développés pour comparer le comportament de la méthode avec celui du test de Kolmogorov-Smirnov and quelques approximations L_2 de la densité dans le travail de Fan (1994).