Modified Multivariate Kolmogorov-Smirnov Test of Goodness of Fit

G.S. Monti[1] , G. Mateu-Figueras[2] , M. I. Ortego[3] , V. Pawlowsky-Glahn[2] , and J. J. Egozcue[3]

[1] Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy
gianna.monti@unimib.it

[2]Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain

[3]Department of Civil and Environmental Engineering, Technical University of Catalonia, Spain

Abstract

Contributions on a general multivariate goodness-of-fit test are scarce in the literature, although they are frequently used in applications. Particularly scarce are studies or generalizations of the Kolmogorov-Smirnov (KS) test. Here a modified version of the KS-test is presented as a tool to assess whether a specified, although arbitrary, multivariate probability model is unsuitable to describe the underlying distribution of a set of observations.

Consider an independent multivariate sample, denoted $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, coming from a k-variate continuous random vector $\mathbf{X}$. Let the hypothetical cumulative distribution function (cdf) be $F(\mathbf{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of F. We formulate the hypothesis $H_0 : \mathbf{X} \sim F(\cdot|\boldsymbol{\theta})$, against the alternative that the random variable does not follow the claimed distribution.

Let be $F^*(\mathbf{x}_i) = (1/n)\#^*[x_{i1}, x_{i2}, \ldots, x_{ik}]$ the empirical distribution (*superior value*), where $\#^*[\cdot]$ is the number of elements of the sample which coordinates are less than or equal to the corresponding coordinates of $\mathbf{x}_i$, including those of the sample $\mathbf{x}_i$. Define also the empirical distribution (*inferior value*) $F_*(\mathbf{x}_i) = (1/n)\#_*[x_{i1}, x_{i2}, \ldots, x_{ik}]$ where $\#_*[\cdot]$ is the number of elements of the sample which coordinates are less than the corresponding coordinates of $\mathbf{x}_i$. The KS statistic consists in looking for the maximum Euclidean distance between the empirical distributions values $F^*(\mathbf{x}_i)$, $F_*(\mathbf{x}_i)$ and the theoretical $F(\mathbf{x}_i|\boldsymbol{\theta})$. As any probability p can be considered in a 2-part simplex, we propose to compute that distance between probabilities as an Aitchison distance.

Define $\mathrm{ilr}(p) = \log(p/(1-p))/\sqrt{2}$ (i.e. the logit transform up to division by $\sqrt{2}$). Then, the Aitchison distances for which we look for the maximum are $|\mathrm{ilr}(F^*(\mathbf{x}_i)) - \mathrm{ilr}(F(\mathbf{x}_i|\boldsymbol{\theta}))|$ and $|\mathrm{ilr}(F_*(\mathbf{x}_i)) - \mathrm{ilr}(F(\mathbf{x}_i|\boldsymbol{\theta}))|$. The ilr-difference is evaluated at the sampling points where the empirical cdf is not null. We define our modified version of the KS statistic, D_a, as the maximum value of these Aitchison distances. In order to reduce the influence of the tails, we suggest for example to trim the sample, or the use of weighting techniques.

In this contribution we investigate by simulation the asymptotic distribution of D_a, checking the appropriateness of the Gumbel distribution, in order to carry out the test. The properties of the asymptotic distribution will be studied with special attention to invariance under affine transformations of the distribution and sample.

Although the test can be very useful in univariate statistics, the use in bivariate situations may be important, particularly to test goodness of fit for copulas.