Modified 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-BarcelonaTECH, Spain

Abstract

A modified version of the Kolmogorov-Smirnov (KS) test is presented as a tool to assess whether a specified, although arbitrary, probability model is unsuitable to describe the underlying distribution of a set of observations. The KS test computes distances between points of the sample cumulative distribution function and the hypothetical one as absolute differences between them, and then considering the supreme distance as test statistics. The modification here proposed consists of computing the mentioned distances as Aitchison distances of the probabilities as two part compositions.

In this contribution, we investigate by simulation the asymptotic distribution of the proposed test statistic, checking the appropriateness of the Gumbel distribution. The properties of the asymptotic distribution are studied for samples coming from generic distributions such as uniform, normal, lognormal, gamma, beta and exponential with different values of the parameters. A brief Monte Carlo investigation is made of the type I error and power of the test.

1 Introduction

The main purpose of this paper is to develop a goodness of fit test to assess the appropriateness of a certain theoretical distribution to the empirical one given a sample. We propose a modified version of the Kolmogorov-Smirnov test, which considers the largest absolute difference between two cumulative distribution functions (CDFs) as a dissimilarity. Section 2 presents the modified KS statistic which we propose in this paper. Section 3 deals with a Monte Carlo simulation study in order to investigate the asymptotic distribution of the proposed statistic and also to investigate the type I error and power of the test. Section 4 reports some comments on our proposal, which is just a first attempt to provide a log-ratio approach to a goodness of fit test, and suggests possible relationships between the sample size and the form of the test statistics.

2 The modified Kolmogorov-Smirnov statistic

Consider an independent sample, denoted $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$, coming from a continuous random variable X. Let the hypothetical CDF be $F(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ are the parameters of F. We formulate the hypothesis

$$H_0 : \; X \sim F(\cdot|\boldsymbol{\theta}) \,,$$

against the alternative that the random variable does not follow the claimed distribution.

H_0 can be tested using the well-known Kolmogorov-Smirnov (KS) statistic introduced by Kolmogorov (1933), which is a tool to assess whether a specified probability model is suitable to describe the underlying distribution of a set of observations. The expression of the KS statistic is

$$D_{KS} = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \,,$$

ISBN: 978-84-947240-0-8

where

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}\,,$$

is the empirical distribution function (EDF) of the sample and counts the proportion of the sample points less than or equal to x, and where $\mathbf{1}\{A\}$ is the indicator of event A. In the context of tests of fit the Kolmogorov-Smirnov statistic can be formulated as follows. Suppose that $F(x)$ is a continuous distribution, to be tested as the parent distribution of a given random sample $X_1, \ldots, X_n$. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics ($i = 1, \ldots, n$), and consider the largest difference at the points where the EDF is greater than $F(x)$ and the largest difference at the points where the EDF is smaller than $F(x)$ as

$$\begin{aligned} D_{KS}^{+} &= \max_{i=1,\ldots,n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\}, \\ D_{KS}^{-} &= \max_{i=1,\ldots,n} \left\{ F(X_{(i)}) - \frac{(i-1)}{n} \right\}, \end{aligned} \tag{1}$$

then, the Kolmogorov-Smirnov statistic is

$$D_{KS} = \max\left\{D_{KS}^{+}, D_{KS}^{-}\right\}. \tag{2}$$

The distribution of this statistic is known, even for finite samples (Birnbaum, 1952; Darling, 1957), and tables are available (Owen, 1962; D'Agostino and Stephens, 1986).

The modification of the KS test statistics in Equation (2) proposed here consists in replacing the absolute difference between the sample and the hypothetical CDF, which is a distance between real numbers, by a suitable difference for probabilities. Probabilities, like for instance i/n and $F(x_{(i)})$, can be considered as two part compositions, like for instance $(i/n, 1-i/n)$ and $(F(x_{(i)}), 1-F(x_{(i)}))$. In this case, a natural way of measuring the distance between probabilities is adopting the Aitchison distance (Aitchison, 1983; Aitchison et al., 2001). For 2-part compositions the Aitchison square distance between $\mathbf{p}_1 = (p_1, 1-p_1)$ and $\mathbf{p}_2 = (p_2, 1-p_2)$ is

$$\mathrm{d}_a^2(\mathbf{p}_1, \mathbf{p}_2) = \left(\frac{1}{\sqrt{2}} \ln \frac{p_1}{1-p_1} - \frac{1}{\sqrt{2}} \ln \frac{p_2}{1-p_2} \right)^2,$$

which is the square difference between the logit transforms of p_1 and p_2 up to the factors $1/\sqrt{2}$. Therefore, the Aitchison distance between two probabilities, $\mathrm{d}_a^2(p_1, p_2)$, can be identified with $\mathrm{d}_a^2(\mathbf{p}_1, \mathbf{p}_2)$. Under this perspective, we propose to consider

$$\begin{aligned} D_a^{+} &= \max_{i=1,\ldots,n-1} \left\{ \mathrm{d}_a\Big(\frac{i}{n}, F(X_{(i)})\Big) \right\}, \\ D_a^{-} &= \max_{i=2,\ldots,n} \left\{ \mathrm{d}_a\Big(F(X_{(i)}), \frac{(i-1)}{n}\Big) \right\}, \end{aligned} \tag{3}$$

and the modified KS statistic is

$$D_a = \max\left\{D_a^{+}, D_a^{-}\right\}. \tag{4}$$

Note that the ranges of the index i in Equations (3) have been modified with respect to Equation (1), thus excluding infinite distances. In fact, a probability equal to 0 or equal to 1 is always at an infinite distance of other probabilities considered as compositions.

An important property of D_a as a test statistics is that it is invariant under a reversion of the orientation of the axis of the data. This means that the CDFs $F(x|\theta)$, i/n, $(i-1)/n$ can be substituted by $1 - F(x|\theta)$, $1 - i/n$, $1 - (i-1)/n$ respectively in Equation (3) and the value of D_a does not change. This property is not fulfilled by the KS test statistics D_{KS}.

In order to complete a practical test, the distribution of the statistic in Equation (4) needs to be studied. However, the statistic (4) is the maximum of several distances. As a consequence,

ISBN: 978-84-947240-0-8

the asymptotic distribution of D_a is a generalized extreme value distribution (GEVD) (Embrechts et al., 1997). GEVD applies even in the case in which there is a weak dependence between the variables from which the maximum is computed. The appropriate type of GEVD is determined by the behaviour of the upper tail of the distances. In the present case, as the support of the distances is not bounded, the Weibull type of GEVD is excluded, and as the decay of the upper tails is exponential, the asymptotic distribution of the maximum is the Gumbel distribution (GEVD with $\xi = 0$) (Appendix A).

Supported by a large number of Monte Carlo simulations (not shown here), we have observed that the D_a statistic follows reasonably well a GEVD for maxima, that is, D_A asymptotically follows a Gumbel distribution (Gumbel, 1954) and its parameters approximately depend on the sample size, as shown in Section 3.

3 Simulation results

In order to investigate the parameters of the asymptotic distribution of the D_a statistic, we have conducted a Monte Carlo study. The Monte Carlo (MC) procedure is as follows. A case consists of a single maximum likelihood estimation of the parameters of the Gumbel distribution. This case is obtained from $m = 500$ simulated samples coming from a given distribution with fixed sample size and parameters, i.e. we consider only the all-parameters-known case. For each case, the distribution model and the sample size are randomly selected. This is repeated for 1,000 different cases, thus obtaining 1,000 estimations of the scale and shape parameters of the Gumbel distribution fitted to D_a. To obtain robust results, in these simulations a 5% trimmed D_a statistic was used.

The considered reference models were: normal and lognormal distribution with different mean and scale parameters, uniform distribution with several supports, exponential distribution with several rates, and gamma and beta distribution with great variety in the parameters. The sample sizes were randomly selected ranging from $n = 5$ up to $n = 15,000$.

Two regression models, linear and quadratic, of the 1,000 MC estimates of the Gumbel parameters, μ (location) and σ (scale), were estimated against the log-size of the sample. The results are displayed in Figure 1 and in Tables 1 and 2. Using the F-test to compare both models, the

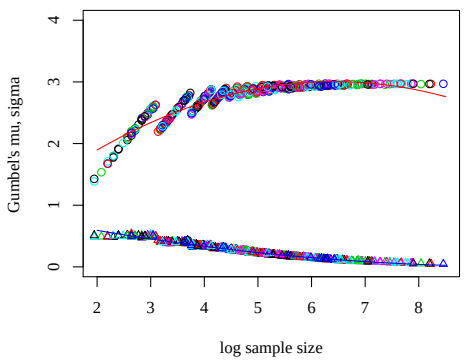

Figure 1: MC results for the Gumbel parameters. Horizontal alignment for scale parameter. Tilted alignment for location parameter. Colours indicate different distribution models. Red and blue lines represent the estimated quadratic models.

ISBN: 978-84-947240-0-8

quadratic appears to be better than the linear one.

Table 1: Regression output for models $(1) : \mu = \beta_{01} + \beta_{11}\ln(n) + \varepsilon$ and $(2) : \mu = \beta_{01} + \beta_{11}\ln(n) + \beta_{21}(\ln(n))^2 + \varepsilon$.

	Dependent variable:	
	μ	
Coefficients	(1)	(2)
β_1	0.159^{***}	0.729^{***}
	(0.006)	(0.019)
β_2		-0.057^{***}
		(0.002)
β_0	1.989^{***}	0.664^{***}
	(0.027)	(0.047)
Observations	500	500
R^2	0.628	0.867
Adjusted R^2	0.627	0.866
Residual Std. Error	0.163 (df = 498)	0.098 (df = 497)
F Statistic	840.711^{***} (df = 1; 498)	$1{,}619.445^{***}$ (df = 2; 497)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 2: Regression output for models $(1) : \sigma = \beta_{02} + \beta_{12}\ln(n) + \varepsilon$ and $(2) : \sigma = \beta_{02} + \beta_{12}\ln(n) + \beta_{22}(\ln(n))^2 + \varepsilon$.

	Dependent variable:	
	σ	
Coefficients	(1)	(2)
β_1	-0.092^{***}	-0.185^{***}
	(0.001)	(0.004)
β_2		0.009^{***}
		(0.0004)
β_0	0.708^{***}	0.926^{***}
	(0.005)	(0.009)
Observations	500	500
R^2	0.947	0.977
Adjusted R^2	0.947	0.977
Residual Std. Error	0.029 (df = 498)	0.019 (df = 497)
F Statistic	$8{,}947.071^{***}$ (df = 1; 498)	$10{,}484.430^{***}$ (df = 2; 497)
Note:		*p<0.1; **p<0.05; ***p<0.01

A brief Monte Carlo investigation was made on the size (type I error) and on the power of the test. 5,000 samples of size $n = 30, 50, 100$ were drawn from each of several distributions. The probability of rejection using the modified Kolmogorov-Smirnov test (Tables 3 and 4) was determined. Results

ISBN: 978-84-947240-0-8

in Table 3 show a low conservative test in the sense that the actual significance level would be much greater than that given by the table, especially when the sample size is small.

Table 3: Probability of rejecting the null hypothesis using D_a (trim 5%) statistic with different sample sizes n. The numbers are the result of Monte Carlo simulations with 5,000 samples for each distribution.

Underlying distribution	Critical Level α	n=30	$n = 50$	$n = 100$
$N(0, 1)$	0.05	0.0268	0.0242	0.0310
$N(0, 1)$	0.1	0.0644	0.0664	0.0964
$Unif(1, 2)$	0.05	0.0244	0.0232	0.0362
$Unif(1, 2)$	0.1	0.0706	0.0704	0.1024
$Gamma(3, 5)$	0.05	0.0236	0.0276	0.0392
$Gamma(3, 5)$	0.1	0.0674	0.0730	0.0996
$Beta(2, 3)$	0.05	0.0258	0.0272	0.0382
$Beta(2, 3)$	0.1	0.0686	0.0732	0.0974

Table 4: Probability of rejecting hypothesis of Standard Normal distribution using D_a (trim 5%) statistic with different sample sizes n. The numbers are the result of Monte Carlo simulations with 5,000 samples for each distribution.

Underlying distribution	Critical Level α	n=30	$n = 50$	$n = 100$
$N(0, 4)$	0.05	0.9232	0.9876	1.0000
	0.1	0.9652	0.9962	1.0000
Student's t, 3 d.f.	0.05	0.4034	0.5360	0.8124
	0.1	0.5274	0.6628	0.8992
Exponential, rate=1	0.05	0.6122	0.8002	0.9792
	0.1	0.7276	0.8898	0.9922
Gamma, shape=rate=1	0.05	0.6160	0.8018	0.9788
	0.1	0.7398	0.8908	0.9936

4 Discussion

In this contribution we have proposed a modified version of the KS statistic. 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. However, our proposal is just a first tentative to provide a log-ratio approach to a goodness of fit test, and to suggest possible relationships between the sample size and the form of the test statistics. Further studies are required to arrive at any definitive conclusions.

Acknowledgements

Research partially financially supported by the Italian Ministry of University and Research, FAR (Fondi di Ateneo per la Ricerca) 2015. The authors also gratefully acknowledge support by the Spanish Ministry of Education and Science under project 'CODA-RETOS' (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)) and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project 'COSDA' (Ref. 2014SGR551).

ISBN: 978-84-947240-0-8

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika 70*(1), 57–65.

Aitchison, J., C. Barceló-Vidal, A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Reply to letter to the editor by S. Rehder and U. Zier on Logratio analysis and compositional distance. *Mathematical Geology 33*(7), 849–860.

Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of kolmogorov's statistic for finite sample size. *Journal of the American Statistical Association 47*(259), 425–441.

Castillo, E. (1988). *Extreme Value Theory in Engineering*. Statistical Modeling and Decision Science. San Diego, Ca. (USA): Academic Press.

Castillo, E., A. Hadi, N. Balakrishnan, and J. Sarabia (2004). *Extreme value and related models with Applications in Engineering and Science*. London, GB: Wiley. 384 p.

D'Agostino, R. and M. Stephens (1986). *Goodness-of-fit Techniques*. Statistics, textbooks and monographs. New York (USA): Marcel Dekker, INC.

Darling, D. A. (1957). The kolmogorov-smirnov, cramer-von mises tests. *The Annals of Mathematical Statistics 28*(4), 823–838.

Embrechts, P., C. Kloppelberg, and T. Mikosch (1997). *Modelling extremal values*. Springer Verlag, Berlin.

Gumbel, E. (1954). *Statistical theory of extreme values and some practical applications*, Volume 33. U.S. Department of Commerce, National Bureau of Standards, Applied Mathematics Series. (1st ed.).

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari 4*, 83–91.

Kotz, S. and S. Nadarajah (2000). *Extreme value distributions. Theory and applications*. London, GB: Imperial College Press. 185 p.

Owen, D. (1962). *A Handbook of Statistical Tables*. Addison-Wesley, Reading, Mass.

A The Gumbel Distribution

The material in this appendix is well known and can be found in Embrechts et al. (1997) or in Castillo (1988), among others. The generalized extreme value distribution (GEVD) has the expression (Von Mises-Jenkinson formula; Embrechts et al. (1997); Castillo et al. (2004); Kotz and Nadarajah (2000))

$$F_Z(z|\mu,\sigma,\xi) = \exp\left[-\left(1+\xi\Big(\frac{z-\mu}{\sigma}\Big)\right)^{-1/\xi}\right] \ , \ 1+\frac{\xi}{\sigma}(z-\mu) > 0 \ , \tag{5}$$

where μ is a location parameter, σ is a scale parameter and ξ is a shape parameter. Parameters μ and ξ have support on the whole real line, and σ is positive. The values of the shape parameter ξ define the three families of asymptotic distribution: Weibull for $\xi < 0$, Fréchet for $\xi > 0$ and Gumbel in the limit case $\xi = 0$.

In particular, if $\xi = 0$ (Gumbel distribution), the expression (5) has the limit form

$$F_Z(z|\mu,\sigma,\xi=0) = \exp\left[-\exp\left(-\frac{z-\mu}{\sigma}\right)\right] \ , \ z \in \mathbb{R} \ . \tag{6}$$

ISBN: 978-84-947240-0-8

The corresponding probability density function is

$$f_Z(z|\mu,\sigma,\xi=0) = \frac{1}{\sigma}\exp\left(-\frac{z-\mu}{\sigma}\right)\exp\left(-\exp\left(-\frac{z-\mu}{\sigma}\right)\right) \ .$$

The mean and variance of the GEVD-Gumbel distribution are, respectively,

$$\mathrm{E}(Z) = \mu + \sigma^2\gamma \quad , \quad \mathrm{Var}(Z) = \frac{\pi^2}{6}\sigma^2 \, ,$$

where γ is the Euler-Mascheroni constant, $\gamma = -\int_0^\infty e^{-t}\ln t \, \mathrm{d}t$. The inverse CDF sampling technique could be used to generate a random sample from a Gumbel distribution: if $U \sim Unif(0,1)$ then $Y = F^{-1}(U) = -\ln(-\ln U)$ has the standard Gumbel distribution (with $\mu = 0$ and $\sigma = 1$). Given this result, the calculation of critical values of the test probability distribution is easy to compute.

ISBN: 978-84-947240-0-8