Rethinking the Kolmogorov-Smirnov Test of Goodness of Fit in a Compositional Way

Il test di bontá di adattamento di Kolmogorov-Smirnov ripensato in chiave composizionale

G.S. Monti, G. Mateu-Figueras, M. I. Ortego, V. Pawlowsky-Glahn and J. J. Egozcue

Abstract The Kolmogorov Smirnov test (KS) is a well known test used to asses how a set of observations is significantly different from the probability model specified under the null hypothesis. The KS test statistic quantifies the distance between the empirical distribution function and the hypothetical one. The modification introduced in Monti et al. (2017) consists of computing the mentioned distances as Aitchison distances. In this contribution, we suggest a further modification of the latter test and investigate, by simulation, the asymptotic distribution of the proposed test statistic, checking the appropriateness of a Generalized Extreme Value (GEV) Distribution. The properties of the asymptotic distribution are studied via Monte Carlo simulations.

Abstract *Il test di Kolmogorov Smirnov (KS) é tra i piú noti test di bontá di adattamento di un modello ai dati. Il test KS é una funzione della distanza tra la distribuzione empirica dei dati e quella ipotizzata sotto l'ipotesi nulla. La modifica del test proposta in Monti et al. (2017) consiste nell'impiego della distanza di Aitchison come misura di tale scostamento. In questo contributo proponiamo una leggera modifica di quest'ultima statistica test, per la quale, attraverso simulazioni Monte Carlo, studieremo la distribuzione asintotica valutando l'accuratezza di una distribuzione generalizzata per valori estremi (GEV).*

Key words: Generalized Extreme Value Distribution, Aitchison distance, Monte Carlo Simulations

---

G.S. Monti
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy , e-mail: gianna.monti@unimib.it

G. Mateu-Figueras and V. Pawlowsky-Glahn
Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain

M. I. Ortego and J. J. Egozcue
Department of Civil and Environmental Engineering, Technical University of Catalonia-BarcelonaTECH, Spain

1 Modified Kolmogorov-Smirnov Test

Consider a random sample, denoted $\mathbf{x} = (x_1, \ldots, x_i, \ldots, x_n)$, coming from a continuous variable X. Let the hypothesized CDF be $F(x|\theta)$, where θ is the vector of parameters of F. We formulate the hypothesis $H_0: X \sim F(\cdot|\theta)$, against the alternative that the random variable does not follow the claimed distribution.

The Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933) consists of rejecting H_0 when the statistic

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

exceeds a critical value — which depends on the sample size n and on the significance level α — where, for all x, $F_n(x) = \frac{1}{n}\{\text{the number of } X_i\text{'s which are} \leq x\}$ is the empirical distribution function (EDF) of the sample. D_{KS} can be computed calculating first

$$D^+_{KS} = \max_{i=1,\ldots,n} \left\{ \frac{i}{n} - F(X_{(i)}) \right\} \quad \text{and} \quad D^-_{KS} = \max_{i=1,\ldots,n} \left\{ F(X_{(i)}) - \frac{(i-1)}{n} \right\}, \tag{1}$$

where $X_{(i)}$ is the ith order statistic; then the KS test statistic is $D_{KS} = \max\left\{D^+_{KS}, D^-_{KS}\right\}$. The distribution of this statistic is known, even for finite samples (Darling, 1957), and tables are available.

Here we consider a slight variation of the modified KS test statistic, denoted D_a, which has been defined and discussed previously (Monti et al., 2017). D_a consists in replacing the absolute difference between the sample and the hypothetical CDF, with the Aitchison distance (Aitchison, 1983) between two part compositions

$$\begin{aligned}
\mathbf{Z}_\ell(i) &= \left(\frac{i}{n+1}, 1 - \frac{i}{n+1} \right) = \left(\frac{i}{n+1}, \frac{n+1-i}{n+1} \right), \\
\mathbf{Z}_u(i) &= \left(\frac{i-1}{n+1}, 1 - \frac{i-1}{n+1} \right) = \left(\frac{i-1}{n+1}, \frac{n+2-i}{n+1} \right), \\
\mathbf{Z}_0(i) &= \left(F(x_{(i)}), 1 - F(x_{(i)}) \right),
\end{aligned}$$

that is $D_a = \max\left\{D^+_a, D^-_a\right\}$, where

$$D^+_a = \max_{i=1,\ldots,n} \left\{ \mathrm{d}_a\Big(\mathbf{Z}_\ell(i), \mathbf{Z}_0(i)\Big) \right\}, \quad D^-_a = \max_{i=1,\ldots,n} \left\{ \mathrm{d}_a\Big(\mathbf{Z}_0(i), \mathbf{Z}_u(i)\Big) \right\}. \tag{2}$$

Whereas in the previous version we considered the ratios $\frac{i}{n}$ in formula (2), in this version we adopt the median rank or the Weibull plotting position which are slightly more accurate than mean ranks.

D_a is motivated by the fact that probabilities, like for instance $i/(n+1)$ and $F(x_{(i)})$ as well as $(i/(n+1), 1 - i/(n+1))$ and $(F(x_{(i)}), 1 - F(x_{(i)}))$, can be considered as two part compositions, and then the Aitchison distance (Aitchison, 1983; Aitchison et al., 2001) can be adopted as a natural similarity measure. We recall that for 2-part compositions, $\mathbf{p}_1 = (p_1, 1-p_1)$ and $\mathbf{p}_2 = (p_2, 1-p_2)$, the Aitchison

square distance between them is

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

It has been shown that D_a, as a test statistic, is invariant under a reversion of the orientation of the axis of the data (Monti et al., 2017).

Supported by a large number of Monte Carlo simulations, in Section 2 it will be shown that D_a follows reasonably well a Generalized Extreme Value Distribution (GEVD) for maxima and its location and scale parameters depend approximately on the sample size.

Recall that a random variable Z has a GEVD if its probability function can be written as

$$F_Z(z|\mu,\sigma,\xi)=\exp\left[-\left(1+\xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right]\ ,\ 1+\frac{\xi}{\sigma}(z-\mu)>0\ , \qquad (3)$$

where $\mu\in\mathbb{R}$ is a location parameter, $\sigma>0$ is a scale parameter, and $\xi\in\mathbb{R}$ is a shape parameter. The values of the shape parameter ξ define the three families of asymptotic distribution: type II for $\xi>0$, type III for $\xi<0$ and Gumbel in the limiting case $\xi=0$ in this parameterization (Fisher and Tippett, 1928; Embrechts et al., 1997).

2 Simulation results

In order to assess the accuracy of the GEV model to the D_a statistic defined in (2), we have conducted an intensive Monte Carlo (MC) simulation.

For each reference model – Normal, Uniform, Gamma, Beta, Exponential and lognormal with random parameters, i.e. we consider only the all-parameters-known case – and for each sample size – 1,000 different sample size values, ranging from 5 to 50,000 – we have simulated 1,000 random samples. For each simulated sample we have computed the D_a statistic in order to test the goodness of fit of the theoretical distribution. All the computations were carried out using the R statistical software program (R Core Team, 2017).

For each reference model and for each fixed sample size we have estimated the parameters of the Gumbel distribution, a subfamily of the GEV for $\xi=0$, and of the GEV model for the $1,000$ D_a values by maximum likelihood method.

Two linear regression models and three linear regression models of the 1,000 MC estimates of the Gumbel, μ (location) and σ (scale), and GEV parameters, μ (location), σ (scale) and ξ (shape), were estimated as a function of the log-size of the sample. The regression outputs are summarized in Table 1, which reports estimates, standards errors and p-values.

Table 1 Regression output for the different linear regression models.

Reference distribution: Normal			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1772 (0.0088; 0.0000)	0.7975 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7206 (0.0054; 0.0000)	-0.0009 (0.0006; 0.0911)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1817 (0.0092; 0.0000)	0.7973 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7226 (0.0057; 0.0000)	-0.0011 (0.0006; 0.0623)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0117 (0.0066; 0.076)	0.0008 (0.0007; 0.235)
Reference distribution: Uniform			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1782 (0.0095; 0.0000)	0.7974 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7184 (0.0054; 0.0000)	-0.0008 (0.0006; 0.165)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1851 (0.01; 0.0000)	0.7968 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7219 (0.006; 0.0000)	-0.0011 (0.0006; 0.058)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0174 (0.0066; 0.0089)	0.0015 (0.0007; 0.028)
Reference distribution: Gamma			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1798 (0.0091; 0.0000)	0.7971 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7205 (0.0053; 0.0000)	-0.001 (0.0005; 0.0656)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1858 (0.0094; 0.0000)	0.7966 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7233 (0.0055; 0.0000)	-0.0012 (0.0006; 0.0275)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0155 (0.0067; 0.0206)	0.0013 (0.0007; 0.0639)
Reference distribution: Beta			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1366 (0.0126; 0.0000)	0.8018 (0.0013; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.718 (0.0064; 0.0000)	-0.0007 (0.0007; 0.292)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1437 (0.0132; 0.0000)	0.8013 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7217 (0.0067; 0.0000)	-0.001 (0.0007; 0.148)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0183 (0.0082; 0.0262)	0.0014 (0.0009; 0.0982)
Reference distribution: Exponential			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1704 (0.0093; 0.0000)	0.7982 (0.0009; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7112 (0.0055; 0.0000)	-0.0001 (0.0006; 0.861)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1706 (0.0097; 0.0000)	0.7983 (0.001; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.711 (0.0058; 0.0000)	-0.0001 (0.0006; 0.919)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0003 (0.0071; 0.964)	-0.0003 (0.0007; 0.726)
Reference distribution: lognormal			
fitted distribution	linear model	Intercept (SE, pvalue)	Slope (SE, p-value)
Gumbel	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1459 (0.0134; 0.0000)	0.8007 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7106 (0.007; 0.0000)	-0.00002 (0.0007; 0.975)
GEV	$\mu = \beta_0 + \beta_1 \ln(n) + \varepsilon$	1.1507 (0.0141; 0.0000)	0.8003 (0.0014; 0.0000)
	$\sigma = \beta_0 + \beta_1 \ln(n) + \varepsilon$	0.7128 (0.0074; 0.0000)	-0.0002 (0.0008; 0.809)
	$\xi = \beta_0 + \beta_1 \ln(n) + \varepsilon$	-0.0121 (0.0088; 0.169)	0.0008 (0.0009; 0.368)

Likelihood ratio tests have been used to compare the two nested models for all simulation settings, and the proportions of simulated p-values less than 0.05 are reported in Table 2.

Looking at the simulations results we can deduce that the D_a statistic follows a Gumbel distribution, whose location parameter μ is related to the logarithm of the

Reference Model	#p-values $< 0.05/1000$
Normal	0.046
Uniform	0.048
Gamma	0.049
Beta	0.052
Exp	0.067
lognormal	0.048

Table 2 Proportions of simulated p-values less than 0.05 for comparisons of Gumbel and GEV models via asymptotic likelihood ratio tests for each reference distribution.

sample size by a linear relationship. Furthermore, the estimated parameter values are stable with rather small variations among models.

To complete the work, a further Monte Carlo investigation was made on the size (type I error) and on the power of the test. 2,000 samples of fixed size $n = 10, 50, 100, 200, 500, 1000, 1500, 2000, 5000, 10000$, were drawn from each of several distributions. Figure 1 reports six different plots. In the first column the probability of rejecting the null hypothesis using the D_a statistic considering three underlying distributions are reported. The second column reports the probability of rejecting hypothesis $H_0 : X \sim N(1,4)$ against $H_1 : X \sim T(2)$ using the D_a statistic (case (a)); $H_0 : X \sim Ga(2,3)$ against $H_1 : X \sim Exp(2)$ (case (b)) and $H_0 : X \sim Unif(0,1)$ against $H_1 : X \sim Exp(2)$ (case (c)).

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.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises Tests. *The Annals of Mathematical Statistics 28*(4), 823–838.

Embrechts, P., T. Mikosch, and C. Klüppelberg (1997). *Modelling Extremal Events: For Insurance and Finance*. London, UK, UK: Springer-Verlag.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society 24*(2), 180190.

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

Monti, G. S., G. Mateu-Figueras, M. I. Ortego, V. Pawlowsky-Glahn, and J. J. Egozcue (2017). Modified Kolmogorov-Smirnov test of goodness of fit. In K. Hron and R. Tolosana-Delgado (Eds.), *Proceedings of CoDaWork 2017*, pp. 151–158. CoDA, http://www.coda-association.org/en/.

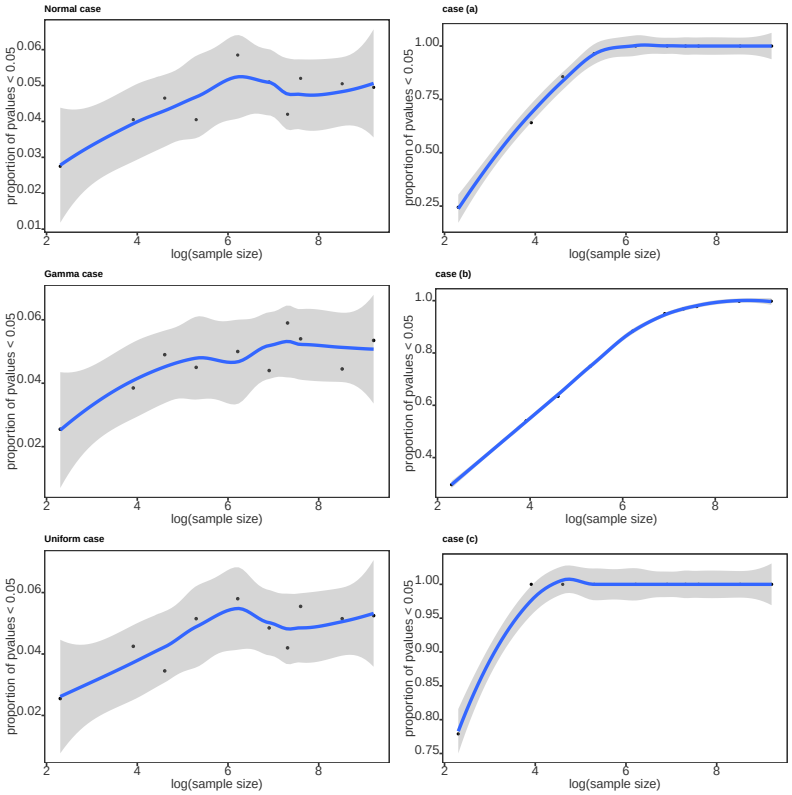

Fig. 1 MC results for probability of type I error (first column) and power of the test (second column). The blu lines represent a smoothing spline fitted to the data.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.