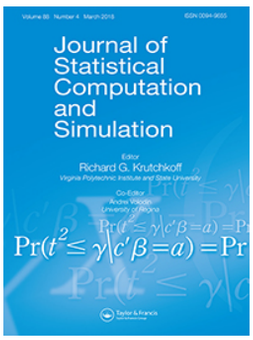

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

An R package for testing goodness of fit: goft

E. González-Estrada & J. A. Villaseñor

To cite this article: E. González-Estrada & J. A. Villaseñor (2018) An R package for testing goodness of fit: goft, Journal of Statistical Computation and Simulation, 88:4, 726-751, DOI: 10.1080/00949655.2017.1404604

To link to this article: https://doi.org/10.1080/00949655.2017.1404604

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION, 2018
VOL. 88, NO. 4, 726–751
https://doi.org/10.1080/00949655.2017.1404604

Taylor & Francis
Taylor & Francis Group

An R package for testing goodness of fit: goft

E. González-Estrada and J. A. Villaseñor

Colegio de Postgraduados, Texcoco, México

ABSTRACT

This R package implements three types of goodness-of-fit tests for some widely used probability distributions where there are unknown parameters, namely tests based on data transformations, on the ratio of two estimators of a dispersion parameter, and correlation tests. Most of the considered tests have been proved to be powerful against a wide range of alternatives and some new ones are proposed here. The package's functionality is illustrated with several examples by using some data sets from the areas of environmental studies, biology and finance, among others.

ARTICLE HISTORY
Received 9 July 2017
Accepted 9 November 2017

KEYWORDS
Tests of fit; hypothesis testing; Shapiro–Wilk test; Anderson–Darling test; correlation tests; data transformations; probability distributions

1. Introduction

Parametric statistical methods assume a specific model for the probability distribution of the observations. In several instances, statistical inferences will be valid only if the assumed distribution is a plausible model for the probability behaviour of the observations. Goodness-of-fit tests are the statistical methods used for testing distributional assumptions.

Several R [1] packages perform goodness-of-fit tests. `dbEmpLikeGOF` package [2] performs density-based empirical likelihood tests for normality and uniformity. `goftest` [3] implements Cramér–von Mises and Anderson–Darling tests of goodness of fit for continuous univariate distributions with known parameters. `fgof` [4] implements classical ECDF goodness-of-fit tests for one sample data with two bootstrap methods. `GofKernel` [5] provides tests based on a kernel smoothing of the data. `MVN` [6] and `energy` [7] perform multivariate normality tests. `nortest` [8] implements five omnibus tests for normality. `PoweR` [9] performs several tests for uniformity, normality and double exponentiality.

Here, we discuss package `goft` [10], which implements some recently published goodness-of-fit tests for probability distributions with unknown parameters that are frequently used in statistical modelling, such as the multivariate and univariate normal, generalized Pareto, extreme value (Gumbel, Fréchet and Weibull), gamma, inverse Gaussian, Laplace, Cauchy, lognormal and exponential distributions. Three classes of tests are implemented: (i) tests based on the ratio of two estimators of some dispersion parameter, (ii) tests based on data transformations, and (iii) correlation tests.

CONTACT E. González-Estrada egonzalez@colpos.mx Colegio de Postgraduados, Km. 36.5 Carr. México-Texcoco s/n,Montecillo, Texcoco 56230, México

In Section 2, notations are defined. In Sections 3–5, the tests implemented in package goft are described. The content of the package is presented in Section 6. In Section 7, several examples are provided by using data sets from the fields of environmental studies, finance, biology and some others. Finally, in Section 8 some conclusions are included.

2. Notation

Notations defined in Table 1 are frequently used along the whole document.

3. Tests based on the ratio of two dispersion estimators

Shapiro–Wilk test for univariate normality compares an estimator for the standard deviation using a linear combination of the order statistics to the usual sample standard deviation. Several simulation studies have shown that this is an omnibus test with good power properties against a wide range of alternatives [11].

In this section, a brief description of some tests that compare two estimators of some dispersion parameter is presented.

3.1. Shapiro–Wilk test for normality

Let $X_1, \ldots, X_n$ denote a random sample of size n. The test statistic of Shapiro and Wilk [12] is the ratio of two estimators of the population variance σ^2, namely the best estimator as a linear function of the order statistics $\tilde{\sigma}^2$ and S^2. This test rejects the normality hypothesis with unknown parameters at a significance level α if $W < k_\alpha$, where

$$W = \frac{\tilde{\sigma}^2}{S^2}. \tag{1}$$

Table 1. Notations.

$\sim$	'is distributed as'
$\approx$	'approximately equal to'
$\Re$	Set of real numbers
α	Significance level of a test, a number in the (0, 1) interval
$z_{1-\alpha}$	$100(1-\alpha)$% quantile of the standard normal distribution
$X_1, \ldots, X_n$	Random sample of size n
$\bar{X}$	Sample mean
S^2	Sample variance
$\Gamma(a)$	Mathematical gamma function evaluated at a
$\psi'(a)$	Trigamma function evaluated at a
$\lvert x \rvert$	Absolute value of x
cdf	Cumulative distribution function
pdf	Probability density function
ECDF	Empirical cumulative distribution function
MLE	Maximum likelihood estimator
iid	Independent and identically distributed
gPd	Generalized Pareto distribution
$\Phi(\cdot)$	Cumulative standard normal distribution
$N(\mu, \sigma^2)$	Normal distribution with parameters μ and σ^2
$U(a, b)$	Uniform distribution in the (a, b) interval
$Exp(b)$	Exponential distribution with scale parameter b

The critical value k_α can be obtained by using the following approximation provided by [13]:

$$k_\alpha \approx 1 - \exp\{\mu_n + \sigma_n z_{1-\alpha}\},$$

where $z_{1-\alpha}$ is the $100(1-\alpha)\%$ quantile of the standard normal distribution,

$$\mu_n = -1.5861 - 0.31082y - 0.083751y^2 + 0.0038915y^3 \tag{2}$$

and

$$\sigma_n = \exp(-0.4803 - 0.082676y + 0.0030302y^2), \tag{3}$$

where $y = \log n$ for $11 < n \leq 2000$.

This test is available in R through `shapiro.test` function of the `stats` package [1].

3.2. A generalization of Shapiro–Wilk test for multivariate normality

Let $\mathbf{N}^p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the multivariate normal distribution of dimension $p \geq 1$ with mean vector μ and covariance matrix Σ. If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are a random sample of dimension p, the sample mean and covariance matrix are defined as $\bar{\mathbf{X}} = (1/n)\sum_{j=1}^n \mathbf{X}_j$ and $\mathbf{S} = (1/n)\sum_{\mathbf{j=1}}^{\mathbf{n}} (\mathbf{X_j} - \bar{\mathbf{X}})(\mathbf{X_j} - \bar{\mathbf{X}})'$.

Let $\mathbf{S}^{-1/2}$ represent the symmetric positive-definite square root of $\mathbf{S}^{-1}$, the inverse matrix of $\mathbf{S}$. When $\mathbf{X}_1, \ldots, \mathbf{X}_n$ follow an $\mathbf{N}^p(\mu, \Sigma)$ distribution, the random vectors

$$\mathbf{Z}_j^* = \mathbf{S}^{-1/2}\left(\mathbf{X}_j - \bar{\mathbf{X}}\right) \tag{4}$$

have a distribution close to the p-dimensional standard normal distribution, for $j = 1, 2, \ldots, n$. This means that the coordinates of $\mathbf{Z}_j^*$ are approximately independent with univariate standard normal distribution. By using this property, Villaseñor and González-Estrada [14] proposed a test for multivariate normality with unknown parameters based on the following test statistic:

$$W^* = \frac{1}{p}\sum_{i=1}^{p} W_{Z_i}, \tag{5}$$

where W_{Z_i} is Shapiro–Wilk statistic defined in Equation (1), evaluated on the ith coordinate of the transformed observations $Z_{i1}, \ldots, Z_{in}$, $i = 1, \ldots, p$.

The p-value of the test is approximated by $\Phi((w_1^* - \mu_1)/\sigma_1)$, where $w_1^* = \log(1 - w^*)$, $\sigma_1^2 = \log((p - 1 + e^{\sigma_n^2})/p)$ and $\mu_1 = \mu_n + \sigma_n^2/2 - \sigma_1^2/2$, with μ_n and σ_n given in Equations (2) and (3), where w^* is the observed value of W^*.

As usual, the multivariate normality hypothesis is rejected at a significance level α if p-value $< \alpha$.

This test is implemented by `mvshapiro_test` function.

Remark 3.1: Notice that when $p = 1$, this test reduces to Shapiro–Wilk test for univariate normality.

3.3. A test for the Gamma distribution

The gamma distribution is a versatile probabilistic model with applications in areas as reliability and survival analysis, hydrology, economics, metagenomics, protein expression, genetics, statistical analysis of cDNA microarray experiments, etc.

A random variable X has a gamma distribution with shape and scale parameters $a > 0$ and $b > 0$, denoted as $X \sim G(a, b)$, if its probability density function (pdf) is given by

$$f_X(x) = \frac{x^{a-1}\, \mathrm{e}^{-x/b}}{\Gamma(a) b^a}, \quad x > 0. \tag{6}$$

Villaseñor and González-Estrada [15] proved that if $X \sim G(a, b)$ then $\mathrm{cov}(X, \log X) = b$. Based on this property, they proposed the following estimators for a and b:

$$\check{b} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Z_i - \bar{Z}) \tag{7}$$

and

$$\check{a} = \bar{X}/\check{b}, \tag{8}$$

where $\bar{X}$ is the sample mean and $Z_i = \log X_i$. These estimators are implemented by function `gamma_fit`.

In the same paper, they introduced the following test for

$$H_0 : X_1, \ldots, X_n \sim G(a, b), \quad \text{where } a \text{ and } b \text{ are unknown.} \tag{9}$$

Reject H_0 in Equation (9) at an approximated significance level α if $|V^*| > \sqrt{2} z_{1-\alpha/2}$, where statistic V^* is defined as

$$V^* = \sqrt{n\check{a}} \left(S^2/\check{\sigma}^2 - 1 \right), \tag{10}$$

with $\check{\sigma}^2 = \bar{X}\check{b}$, an estimator of the population variance of the Gamma distribution based on estimators (7) and (8).

The p-value of the test is approximated by $2(1 - \Phi(|v^*|/\sqrt{2}))$, where v^* is the observed value of V^*.

This test is implemented by `gamma_test` function.

Remark 3.2: The exponential distribution is a gamma distribution with shape parameter equal to one. A well-known test for exponentiality is [16] test. The statistic of this test (CO) is a function of the ratio of two estimators for the scale parameter of the exponential distribution (b), namely $CO = n(1 - \check{b}/\bar{X})$, where $\check{b}$ is given in Equation (7). This test is implemented here by `exp_test` function with argument `method` = " ratio".

3.4. A test for the Inverse Gaussian distribution

The family of Inverse Gaussian distributions has applications in finance, physics, lifetime testing, etc. as a model for positive data sets with positive skewness. A random variable X

has an Inverse Gaussian distribution with parameters $\mu > 0$ and $\lambda > 0$, denoted by $X \sim \text{IG}(\mu, \lambda)$, if the cdf of X is given by

$$F(x; \mu, \lambda) = \Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} - 1\right)\right] + e^{2\lambda/\mu}\Phi\left[-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} + 1\right)\right], \quad x > 0. \tag{11}$$

Villaseñor and González-Estrada [17] proposed the following test for

$$H_0 : X_1, \ldots, X_n \sim \text{IG}(\mu, \lambda), \quad \text{where } \mu \text{ and } \lambda \text{ are unknown.} \tag{12}$$

Reject H_0 in Equation (12) at an approximated significance level α if $|T_1| > z_{1-\alpha/2}$, where statistic T_1 is defined as

$$T_1 = \sqrt{n\hat{\lambda}/6\bar{X}}(S^2/\hat{\sigma}^2 - 1), \tag{13}$$

$\hat{\sigma}^2 = \bar{X}^3/\hat{\lambda}$ is the MLE of the population variance of the IG distribution and $1/\hat{\lambda} = (1/n)\sum_{i=1}^{n}(1/X_i - 1/\bar{X})$.

The p-value of the test is approximated by $2(1 - \Phi(|t_1|))$, where t_1 is the observed value of statistic T_1.

This test is implemented in `ig_test` function with argument `method` = "ratio". Two additional tests for the IG distribution are included in Section 4.

3.5. A test for the Laplace or double-exponential distribution

The Laplace family of distributions is used in the areas of finance, economics, health sciences, hydrology, etc. for modelling symmetric observations. A random variable X has the Laplace distribution or double-exponential distribution with location and scale parameters $-\infty < \theta < \infty$ and $\beta > 0$, denoted by $X \sim L(\theta, \beta)$, if its cdf is given by

$$F_X(x) = \frac{1}{2}\exp\left\{-\frac{\theta - x}{\beta}\right\}, \quad x \le \theta, \tag{14}$$

$$= 1 - \frac{1}{2}\exp\left\{-\frac{x - \theta}{\beta}\right\}, \quad x \ge \theta. \tag{15}$$

An estimator for the scale parameter β is the sample mean average deviation (MAD) about the sample mean $\bar{X}$, defined as

$$\check{\beta} = \sum_{i=1}^{n} |X_i - \bar{X}|/n. \tag{16}$$

On the other hand, a moments estimator for β is $\tilde{\beta} = \sqrt{S^2/2}$.

González-Estrada and Villaseñor [18] proposed a test for

$$H_0 : X_1, \ldots, X_n \sim L(\theta, \beta), \quad \text{where } \theta \text{ and } \beta \text{ are unknown,} \tag{17}$$

based on the following ratio of β estimators:

$$T_2 = \tilde{\beta}/\check{\beta}. \tag{18}$$

This test rejects H_0 in (17) at a test size α if T_2 deviates away from 1. For small sample sizes, the critical values are computed by Monte Carlo simulation for every sample size n, since

T_2 is location-scale invariant. For large sample sizes, the critical values are obtained from the standard normal distribution, since $T_2^* = \sqrt{4n}(T_2 - 1) \sim N(0, 1)$ asymptotically. That is, if the sample size is large ($n \geq 500$), H_0 in Equation (17) is rejected at an approximated test size α if $|T_2^*| > z_{1-\alpha/2}$. In this case, the p-value of the test is approximated by $2(1 - \Phi(|t_2^*|))$, where t_2^* is the observed value of T_2^*.

This test is implemented by `laplace_test` function with argument `method= "ratio"`. A test for the Laplace distribution based on a data transformation to exponentiality is discussed in Section 4.

3.6. A new test for the Cauchy distribution

The Cauchy distribution has been used for modelling symmetric observations with heavy tails in the areas of finance and economics, among others. A random variable X follows the Cauchy distribution with location and scale parameters $-\infty < \theta < \infty$ and $\kappa > 0$, denoted by $X \sim C(\theta, \kappa)$, if its cdf is given by

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-\theta}{\kappa}\right), \quad -\infty < x < \infty. \tag{19}$$

The MLE $\hat{\theta}$ and $\hat{\kappa}$ are the solution to the following system of equations:

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{\kappa^2 + (x_i - \theta)^2} = 0$$

and

$$\frac{n}{\kappa} - \sum_{i=1}^{n} \frac{2\kappa}{\kappa^2 + (x_i - \theta)^2} = 0.$$

Another estimator for the scale parameter κ is the mean absolute deviation about $\hat{\theta}$:

$$\tilde{\kappa} = \frac{1}{n} \sum_{i=1}^{n} |X_i - \hat{\theta}|. \tag{20}$$

For testing the null hypothesis $H_0 : X_1, \ldots, X_n \sim$ Cauchy with unknown parameters, we propose a test based on the following ratio of two estimators for κ:

$$T_3 = \hat{\kappa}/\tilde{\kappa}. \tag{21}$$

This test rejects H_0 at a significance level α if T_3 is larger than a critical constant $c_{1-\alpha}$, which is computed by Monte Carlo simulation since T_3 is location-scale invariant. The p-value of the test is also approximated by Monte Carlo simulation.

This test is implemented by `cauchy_test` function. MLEs are computed using `mledist` function of `fitdistrplus` package [19].

3.7. A new test for extreme value distributions

There are three types of (Fisher-Tippett) extreme value distributions: Gumbel, Fréchet and Weibull. Under certain regularity conditions, these are the distributions of the maximum observation of a random sample.

X has the Gumbel distribution, also known as type I extreme value distribution, with location and scale parameters $-\infty < \mu < \infty$ and $\beta > 0$ if its cdf is given by

$$G_1(x) = \exp\left\{-\exp\left\{-\frac{x-\mu}{\beta}\right\}\right\}, \quad -\infty < x < \infty. \tag{22}$$

The Gumbel distribution has mean $\mu + \gamma_0\beta$ and variance $\sigma^2 = \pi\beta^2/6$, where $\gamma_0 \approx 0.5772157$ is Euler–Mascheroni Constant.

Kimball [20] proposed the following estimator for β in terms of the order statistics:

$$\check{\beta} = \bar{X} - \frac{1}{n}\sum_{i=1}^{n} X_{(i)} \sum_{j=i}^{n} \frac{1}{j}. \tag{23}$$

Hence, an estimator for the population variance is $\check{\sigma}^2 = \pi^2\check{\beta}^2/6$. Another estimator for σ^2 is S^2.

Therefore, for testing the null hypothesis:

$$H_0 : X_1, \ldots, X_n \sim \textit{Gumbel} \text{ with unknown parameters.} \tag{24}$$

we propose the following test. Reject H_0 in Equation (24) if the following ratio deviates away from 1:

$$T_4 = \check{\sigma}^2/S^2. \tag{25}$$

Critical values and/or p-values needed to implement the test are obtained by Monte Carlo simulation since T_4 is a location-scale invariant statistic.

The Fréchet and Weibull extreme value distributions with shape and scale parameters $a > 0$ and $\beta > 0$ have the following cdf's:

$$G_2(x) = \exp\left\{-\left(\frac{x}{\beta}\right)^{-a}\right\}, \quad x > 0, \tag{26}$$

and

$$G_3(x) = \exp\left\{-\left(\frac{-x}{\beta}\right)^{-a}\right\}, \quad x \le 0. \tag{27}$$

If we wish to test goodness of fit for these distributions we propose to use statistic T_4 evaluated at $Y_i = \log X_i$ and $W_i = -\log(-X_i)$, since Y_i and W_i are Gumbel distributed whenever X_i is either Fréchet or Weibull distributed.

These tests are implemented by function `ev_test` with argument `method` = " ratio". In Section 5, a second test for extreme value distributions based on the sample correlation coefficient is described.

4. Tests based on data transformations

In this section we consider tests for the inverse Gaussian, lognormal, Laplace, Weibull and exponential distributions based on data transformations.

4.1. Tests for the Inverse Gaussian distribution

Two tests for the IG distribution based on data transformations to gamma and normal variables are described here.

(a) A test based on a transformation to Gamma variables

If $X \sim \text{IG}(\mu, \lambda)$ defined in Equation (11), then the random variable $\lambda(X-\mu)^2/\mu^2 X$ is distributed as χ^2_1, where χ^2_1 denotes the chi-square distribution with one degree of freedom [21]. Therefore, if $\beta = 2\mu^2/\lambda$, then the random variable

$$Z^{(\mu)} = (X-\mu)^2/X \sim G(1/2, \beta), \tag{28}$$

where $G(1/2, \beta)$ denotes the Gamma distribution with shape parameter 1/2 and scale parameter β.

Under H_0 in Equation (12), by property (28), the transformed observations:

$$Z_i = (X_i - \bar{X})^2/X_i, \quad i = 1, \ldots, n, \tag{29}$$

are asymptotically independent random variables with $G(1/2, \beta)$ distribution. Using this result, for testing the IG hypothesis as stated in Equation (12), Villaseñor and González-Estrada [17] proposed to test

$$H_0' : Z_1, \ldots, Z_n \sim G(1/2, \beta), \quad \text{with } \beta \text{ unknown}, \tag{30}$$

by using scale invariant tests like [22] test.

Notice that if H_0' in Equation (30) is rejected, then H_0 in (12) is rejected.

(b) A test based on a transformation to normality

When $X \sim \text{IG}(\mu, \lambda)$, the random variable $Y = |\sqrt{\lambda}(X-\mu)/\mu\sqrt{X}|$ follows the standard half-normal distribution, denoted by HN(0, 1), with pdf $f(y) = (2/\sqrt{2\pi})\, \mathrm{e}^{-y^2/2}$, $y > 0$. Therefore, $X' = |(X-\mu)/\sqrt{X}|$ is distributed as $\text{HN}(0, \lambda/\mu^2)$.

Then, if U has a Bernoulli distribution with success probability equal to 1/2, denoted as Bernoulli(1/2), the random variable

$$Z' = X'(1-U) + X'(-U) \sim N(0, \lambda/\mu^2), \tag{31}$$

where $N(0, \lambda/\mu^2)$ is the normal distribution with mean zero and variance λ/μ^2.

Notice that, by using Equation (31), a random variable with *IG* distribution is transformed to a random variable with normal distribution. Therefore, for testing H_0 in Equation (12), Ochoa [23] proposed the following test procedure.

(1) Simulate n observations, $U_1, \ldots, U_n$, from the Bernoulli(1/2) distribution.
(2) Compute $Z_1', \ldots, Z_n'$ replacing μ by $\bar{X} = \sum_{i=1}^n X_i/n$ in definition of X'.
(3) Test $H_0' : Z_1', \ldots, Z_n'$ are normally distributed, using Shapiro–Wilk test for normality, which is described in Section 3.
(4) Reject H_0 in Equation (12) at a significance level α if the normality hypothesis is rejected at the same significance level.

These tests are implemented by `ig_test` function with argument `method= "transf"`.

4.2. A test for the lognormal distribution

The lognormal distribution is used for modelling positive data sets with long right-hand tail. A random variable X has the lognormal distribution with parameters μ and σ if $Y = \log X$ follows a $N(\mu, \sigma^2)$ distribution. Therefore, the plausibility of the lognormal distribution hypothesis can be assessed using any test for univariate normality based on the transformed observations $Y_i = \log X_i$. Here, Shapiro–Wilk test is used for testing univariate normality on the Y_i's (see Section 3).

This test is implemented by `lnorm_test` function.

4.3. A new test for the Weibull distribution

The Weibull family of distributions is an alternative model for the gamma distribution. It is used for modelling positively skewed data sets in the areas of survival analysis, reliability, insurance, hydrology, etc. The random variable X has the Weibull distribution if its cdf is given by

$$F(x) = 1 - \mathrm{e}^{-(x/\lambda)^k}, \quad x \geq 0, \tag{32}$$

where $k > 0$ and $\lambda > 0$ are shape and scale parameters, which is denoted as $X \sim \text{Weibull}(k, \lambda)$.

For testing the Weibull distribution hypothesis with cdf given in Equation (32) when the parameters are unknown, we propose to use the ratio test for the Gumbel distribution described in Section 3.7, based on $-\log X_i$, since $-\log X$ is Gumbel distributed.

This test is implemented by `weibull_test` function.

4.4. A new test for the exponential distribution

The exponential distribution is an important model in reliability and survival analysis. The cdf of the exponential distribution with scale parameter $b > 0$ is $F(x) = 1 - \mathrm{e}^{-x/b}$, $x > 0$. If X is a random variable with cdf F then it is well known that $F(X)$ and $\bar{F}(X) = 1 - F(X)$ have a uniform distribution on the (0,1) interval, denoted as $U(0, 1)$. Since the exponential distribution is a particular case of the gamma distribution, then $\check{b}$ given in Equation (7) is an estimator for the scale parameter b. Therefore, if the random sample $X_1, \ldots, X_n$ comes from the exponential distribution, then the transformed observations $U_i = \mathrm{e}^{-X_i/\check{b}}$ are approximately iid random variables with $U(0, 1)$ distribution, $i = 1, \ldots, n$.

For testing exponentiality , Villaseñor and González-Estrada [24] proposed the following decision rule. Reject exponentiality if

$$U^* = \frac{\bar{U} - 1/2}{\sqrt{\tau^2/n}} \tag{33}$$

deviates away from 0, where $\bar{U} = \sum_{i=1}^{n} U_i/n$ and $\tau^2 = 1/12 + (\psi'(1) - 1)/16 + \log(2)$. Critical values and/or p-values needed to perform the test are obtained by Monte Carlo simulation since U^* is scale invariant. For large sample sizes ($n > 200$), these quantities are obtained from the standard normal distribution, which is the asymptotic null distribution of U^*. That is, the exponentiality hypothesis is rejected at a significance level α if

$|U^*| > z_{1-\alpha/2}$ or equivalently if p-value $< \alpha$, where p-value $= 2(1 - \Phi(|u^*|))$ and u^* is the observed value of U^*.

This test is implemented by `exp_test` function with argument `method = "transf"`.

4.5. Tests for the Laplace distribution

If $X \sim L(\theta, \beta)$ defined in Equation (15) then the random variable

$$Y^{(\theta)} = |X - \theta| \sim \text{Exp}(\beta). \tag{34}$$

Using this property, for testing H_0 in Equation (17), González-Estrada and Villaseñor [18] proposed to test $H_0' : Y_1, \ldots, Y_n \sim \text{Exp}(\beta)$, with β unknown, using Anderson-Darling test, where $Y_i = |X_i - \bar{X}|$. H_0 in Equation (17) is rejected at a significance level α if the exponentiality hypothesis is rejected at the same significance level using Anderson–Darling test.

This test is implemented by `laplace_test` function with argument `method = "transf"`.

5. Correlation tests

In this section we describe tests for the normal, extreme value and generalized Pareto distributions based on the sample correlation coefficient.

5.1. A test for normality based on the Lévy property

By the Lévy property that characterizes the normal distribution, if a random sample $X_1, X_2, \ldots, X_n$ of size n comes from a $N(\mu, \sigma^2)$ distribution, then the random variable

$$S_k = X_i + X_j \sim N(2\mu, 2\sigma^2), \tag{35}$$

for $i < j,\ j = 2, \ldots, n,\ k = 1, 2, \ldots, m = n(n-1)/2$. Hence, the cdf of S_k is $F_S(s) = \Phi((s - 2\mu)/\sqrt{2}\sigma)$. Therefore,

$$\Phi^{-1}\left(F_S\left(S\right)\right) = \frac{S - 2\mu}{\sqrt{2}\sigma}, \tag{36}$$

where Φ^{-1} denotes the inverse function of Φ.

From Equation (36), the random variable $Y^* \equiv \Phi^{-1}(F_S(S))$ is a linear function of the random variable S. Thus, the correlation coefficient $\rho(F)$ of Y^* and S must be equal to one.

An estimator of Y_k^* is $Y_k = \Phi^{-1}(F_{S,m}(S_k)), k = 1, 2, \ldots, m$, where $\tilde{F}_S(s) = F_{S,m}(s)$ is the ECDF of the S_k observations, which is given by $F_{S,m}(s) = (1/m) \sum_{k=1}^{m} \mathbf{1}(S_k \leq s)$, where $\mathbf{1}$ is the indicator function.

For testing univariate normality, Villaseñor and González-Estrada [25] proposed a test based on the sample correlation coefficient of Y_k and S_k, $k = 1, 2, \ldots, m$, which is given by

$$R_n = \sum_{k=1}^{m} \left(Y_k - \bar{Y}\right)\left(S_k - \bar{S}\right) \Big/ \sqrt{\sum_{k=1}^{m} \left(Y_k - \bar{Y}\right)^2 \sum_{k=1}^{m} \left(S_k - \bar{S}\right)^2},$$

where $\bar{Y}$ and $\bar{S}$ are the sample means of the $Y_k's$ and $S_k's$. Under univariate normality, the R_n statistic is expected to take on values close to one due to relation (36). Therefore, the normality hypothesis is rejected with a significance level α if $R_n < k_\alpha$, where k_α is approximated by

$$k_\alpha \approx 1 - \exp\{\mu_n + \sigma_n \Phi^{-1}(\alpha)\}, \quad 10 \leq n \leq 400, \tag{37}$$

where

$$\mu_n = -1.928 - 3.553 \log_{10}(n) + 0.7265 \log_{10}(n)^2 - 0.1243 \log_{10}(n)^3, \tag{38}$$

and

$$\sigma_n = 0.3994 + 0.6394 \log_{10}(n) - 0.2033 \log_{10}(n)^2, \quad \text{for } 10 \leq n \leq 30, \tag{39}$$

$$= 0.89994, \quad \text{for } 30 < n \leq 400. \tag{40}$$

The p-value of this test is approximated by $1 - \Phi((\log(1-r) - \mu_n)/\sigma_n)$, where r is the observed value of R_n.

This test is implemented by `normal_test` function.

5.2. A test for the extreme value distributions

Using the max-stability property of the Gumbel distribution, González-Estrada and Villaseñor [26] proposed a test for the null hypothesis given in (24) based on the sample correlation coefficient (R_n) of $Z_k = \max\{X_i, X_j\}$, $i < j$, $i,j = 1, 2, \ldots, n$, and $q(Z_k) = -\log(-\log(G_m(Z_k)))$, $k = 1, 2, \ldots, m = n(n-1)/2$, where $G_m(Z_k)$ is the ECDF of the $Z_k's$. H_0 is rejected for small values of R_n. If r is the observed value of R_n, the p-value is approximated by $1 - \Phi((\log(1-r) - \mu_n^*)/\sigma_n^*)$, where

$$\mu_n^* = -3.02921 - 0.03846n + 0.00023427n^2 - 0.000000471091n^3, \quad 20 \leq n \leq 250,$$

and

$$\sigma_n^* = 0.7588 - 0.01697n + 0.000399n^2 - 0.000003n^3, \quad \text{if } 20 \leq n \leq 60$$
$$\sigma_n^* = 0.53, \quad \text{if } 60 < n \leq 250.$$

By the distributional relationships among the extreme value distributions mentioned in Section 3.7, the correlation test evaluated at $Y_i = \log X_i$ and $W_i = -\log(-X_i)$ can be used for testing the Fréchet and Weibull extreme value distribution hypotheses with cdf's given in Equations (26) and (27).

These tests are implemented by function `ev_test` with argument `method` = " cor".

5.3. A test for the Generalized Pareto distribution

The generalized Pareto distribution (gPd) is a rich family of distributions with cdf given by

$$F(x; \sigma, \gamma) = 1 - \left(1 + \frac{\gamma}{\sigma} x\right)^{-1/\gamma}, \tag{41}$$

where $\sigma > 0$ and $\gamma \in \Re$ such that $x > 0$ for $\gamma \geq 0$ and $0 < x < -\sigma/\gamma$ for $\gamma < 0$.

This family of distributions contains heavy-tailed distributions, the exponential family of distributions as well as a subclass of Beta distributions and others with bounded support. When $\gamma = 0$, $F(x;\sigma,\gamma)$ reduces to the Exp(σ) distribution, and when $\gamma = -1$, $F(x;\sigma,\gamma) = x/\sigma$, which is the $U(0,\sigma)$ distribution. Due to its properties, the GPd has been used for modelling probabilities in different fields like for instance finance, environmental sciences and hydrology, among others [27].

Let $X_{(1)},\ldots,X_{(n)}$ represent the order statistics of $X_1,\ldots,X_n$. Let k be an integer between 1 and n, and let $W_j = \log X_{(j)}$. Villaseñor and González-Estrada [28] proposed two estimation methods for the parameters of the gPd: the asymptotic maximum likelihood estimation (AMLE) method and the combined method. The AMLE method is used to estimate the parameters of the subfamily of gPd with $\gamma \geq 0$. The estimators for γ and σ have the following expressions:

$$\hat{\gamma}_k = -\left(W_{n-k+1} - \frac{1}{k}\sum_{j=1}^{k} W_{n-j+1}\right) \tag{42}$$

and

$$\hat{\sigma}_k = \hat{\gamma}_k \exp\left\{W_{n-k+1} + \hat{\gamma}_k \log(k/n)\right\}. \tag{43}$$

These estimators are computed by `gp_fit` function with argument `method= "amle"`.

The combined method is used to estimate the parameters of the subfamily of gPd with $\gamma < 0$. The estimators have the following expressions:

$$\tilde{\gamma} = \frac{\bar{X}}{\bar{X} - X_{(n)}} \tag{44}$$

and

$$\tilde{\sigma} = -\tilde{\gamma} X_{(n)}, \tag{45}$$

where $X_{(n)}$ is the largest value of the data set. These estimators are computed by `gp_fit` function with argument `method= "combined"`.

If $X_1,\ldots,X_n$ is a random sample from a cdf F, for testing the null hypothesis:

$$H_0 : F \text{is a GP distribution}, \tag{46}$$

let us consider the following two subclasses of GP distributions:

$$A^+ = \left\{\text{all GP distributions with shape parameter } \gamma \geq 0\right\}$$

and

$$A^- = \left\{\text{all GP distributions with shape parameter } \gamma < 0\right\}.$$

Notice that the null hypothesis in Equation (46) is equivalent to $H_0 : F \in (A^+ \cup A^-)$.

In this setting, Villaseñor and González-Estrada [28] proposed an intersection-union test for H_0 by considering separately a test for

$$H_0^+ : F \in A^+ \tag{47}$$

and a test for

$$H_0^- : F \in A^-. \tag{48}$$

5.4. A test for gPd with positive shape parameter

Under H_0^+, there exists a linear relationship between the variables $Y = (\bar{F}(X;\sigma,\gamma))^{-\gamma}$ and X. The same kind of relationship exists between $Y^* = \log((\bar{F}(X;\sigma,\gamma))^{-\gamma} - 1)$ and $X^* = \log(X)$.

Let $Y_i = (\bar{F}_n(X_i))^{-\hat{\gamma}}$, $i = 1, 2, \ldots, n$, where F_n is the ECDF of $X_1, \ldots, X_n$ and $\hat{\gamma} = \hat{\gamma}_k$ is the estimator given in Equation (42). For $0 \leq \gamma < 0.5$, the second moment of X is finite; therefore, the sample correlation coefficient of X_i and Y_i, denoted by R_1, is an estimator of the linear correlation of Y and X when $0 \leq \hat{\gamma} < 0.5$, where

$$R_1 = \frac{\sum_{j=1}^{n} \left(X_j - \bar{X}\right)\left(Y_j - \bar{Y}\right)}{n\sqrt{S_X^2 S_Y^2}}$$

and $\bar{X}$, S_X^2 and $\bar{Y}$, S_Y^2 are the sample means and variances of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$.

Now, let $X_i^* = \log(X_i)$ and $Y_i^* = \log((\bar{F}_n(X_i))^{-\hat{\gamma}} - 1)$, $i = 1, 2, \ldots, n$. For $\gamma \geq 0.5$, the second moment of X^* is finite; therefore, the sample correlation coefficient of Y_i^* and X_i^*, denoted by R_2, is an estimator of the linear correlation of Y^* and X^*, for $\hat{\gamma} \geq 0.5$.

The test statistic for the null hypothesis H_0^+ is

$$R^+ = \begin{Bmatrix} R_1, & \text{if } 0 \leq \hat{\gamma} < 0.5, \\ R_2, & \text{if } \hat{\gamma} \geq 0.5. \end{Bmatrix}.$$

If H_0^+ is true, the value of R^+ is expected to be close to 1. Then H_0^+ is rejected if R^+ is smaller than a critical value.

The p-value of this test is approximated by parametric bootstrap.

5.5. A test for gPd with negative shape parameter

The sample correlation coefficient of X_i and $Z_i = (\bar{F}_n(X_i))^{-\tilde{\gamma}}$, denoted by R^-, is the statistic for testing H_0^-, where $\tilde{\gamma}$ is the estimator given in Equation (44).

If H_0^- is true, $|R^-|$ is expected to take on values close to 1. Therefore, H_0^- is rejected if $|R^-|$ is small.

The p-value of this test is also approximated by parametric bootstrap.

5.6. An Intersection-Union goodness-of-fit test for gPd

An intersection-union test for $H_0 : F$ *is a GP distribution* rejects H_0 whenever both hypotheses H_0^+ and H_0^- are rejected. In order to have a test of level α, H_0^+ and H_0^- are tested at a significance level equal to α.

This test is implemented by function `gp_test`.

6. The goft package

Package `goft` is available from the Comprehensive R Archive Network (CRAN) of R at https://CRAN.R-project.org/package = goft. Version 1.3.4 contains the functions listed in Table 2 for performing the goodness-of-fit tests described in Sections 3–5. All functions have the argument `x`, which is a numeric vector containing the observations for univariate

Table 2. Functions for testing goodness-of-fit.

`cauchy_test`	Test for Cauchy distribution
`ev_test`	Test for extreme value distributions
`exp_test`	Test for exponentiality
`gamma_test`	Test for gamma distribution
`gp_test`	Test for generalized Pareto distributions
`ig_test`	Test for inverse Gaussian distribution
`laplace_test`	Test for Laplace or double exponential distribution
`lnorm_test`	Test for Lognormal distribution
`mvshapiro_test`	Test for multivariate normality
`normal_test`	Test for univariate normality
`weibull_test`	Test for Weibull distribution

Table 3. Functions for parameter estimation.

`gamma_fit`	Parameter estimators for the gamma distribution given in Equations (8) and (7)
`gp_fit`	Parameter estimators for the generalized Pareto distribution given in Equations (44), (45), (42) and (43)
`ig_fit`	Maximum likelihood estimators for the inverse Gaussian distribution

distributions or a numeric matrix for function `mvshapiro_test`. If the null distribution is defined on the positive real numbers, `x` can only contain positive observations.

Functions containing more than one method for testing a distributional hypothesis (`ev_test`, `exp_test`, `ig_test` and `laplace_test`), also have the argument `method`, which specifies the particular goodness-of-fit test to be used. At least one of the following options are available: `ratio`, `transf` and `cor` for the ratio, transformation and correlation tests described in Sections 3–5.

Furthermore, `goft` package also contains the functions given in Table 3 for estimating the parameters of the gamma, inverse Gaussian and generalized Pareto distributions.

7. Examples

In this section, data sets coming from the areas of environmental analysis, finance and biology are used to illustrate `goft` package's functionality. Several of these data sets (`o3`, `o3max`, `strength`) are included as part of the package.

7.1. Modelling ozone data from Mexico City

Mexico City is one of the largest cities in the world with a population close to 20 million people. The combination of factors as overpopulation, industrial growth, proliferation of vehicles, geography (the city is situated in a valley surrounded by mountains, at 2240 meters above sea level), latitude (tropics), among others, produces high levels of air pollutants, which are a health hazard for the inhabitants.

7.1.1. Example 1

The `o3` data set contains the ozone concentrations in parts per million (ppm) registered in the Southwest of Mexico City from January 1rst, 2008 to April 30th, 2016 above 0.165 ppm. These ozone levels correspond to days when the air quality is pretty bad or extremely bad, according to the standards defined by the Mexican government.

Let X denote the ozone concentration and u denote a threshold. The difference $X-u$ is called 'excess'. Extreme value theory states that, given u, the generalized Pareto distribution defined in Equation (41) is the appropriate distribution for modelling the probability behaviour of the excesses if u is sufficiently large [29]. That is, $X-u$ given a large u follows a gPd.

Next, for testing that the excesses given $u=0.165$ ppm follow a gPd, the test described in Section 5 is used, which is implemented in `gp_test` function. The resulting p-value is high; therefore, the gPd is a plausible probability distribution for modelling probabilities of the excesses over the 0.165 ppm threshold, according to this test. The code is as follows.

```
> library(goft)
> data(o3)
> o3levels <- o3$ozone_level
> excess <- o3levels - 0.165
> results <- gp_test(excess, B = 9999); results

Bootstrap test of fit for the generalized Pareto
 distribution

data: excess p-value = 0.7351
```

In order to find out if the distribution of the excesses given $u=0.165$ is either a gPd with negative shape parameter (H_0^-) or a gPd with positive shape parameter (H_0^+), the p-values corresponding to H_0^- and H_0^+ are obtained as follows.

```
> results$pvalues

                                                          p.value      R
H_0^-: excess follows a gPd with NEGATIVE shape parameter  0.7351 0.9940
H_0^+: excess follows a gPd with POSITIVE shape parameter  0.2439 0.9597
```

Notice that the excesses are better modelled by a gPd with negative shape parameter since the p-value associated with H_0^- is larger than the p-value corresponding to H_0^+. Then, in order to fit a gPd with negative shape parameter, `gp_fit` function is used with argument `method = "combined"` as follows. This method computes the estimators given in Equations (44) and (45).

```
> fit <- gp_fit(excess, method = "combined"); fit

Parameter estimates
shape    -0.3145 scale    0.0183
```

Figure 1 depicts a plot of the ECDF of the excesses and the fitted gPd, which provides graphical evidence in favour of the plausibility of the gPd with negative shape parameter for modelling the excesses over the 0.165 ppm threshold.

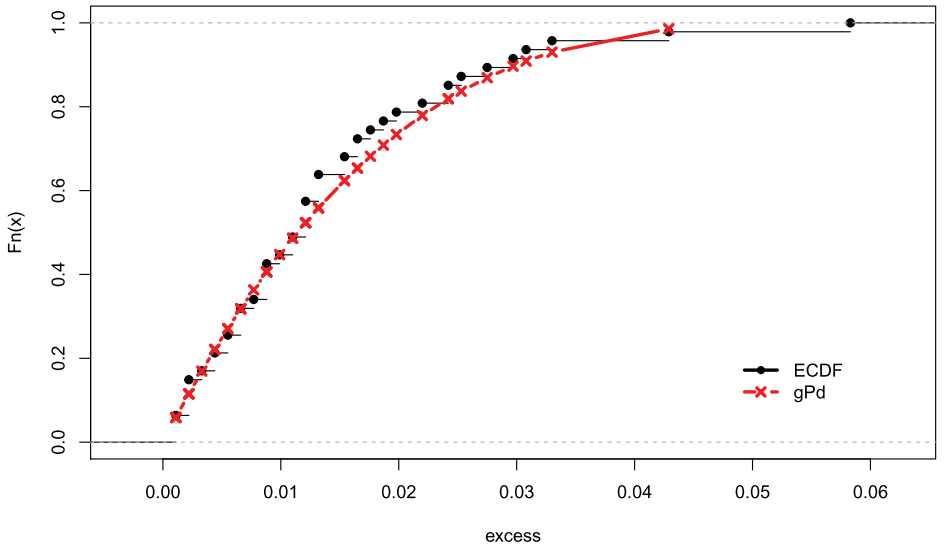

Figure 1. ECDF and fitted gPd plot for the excesses over the 0.165 ppm threshold.

7.1.2. Example 2

The `o3max` data set contains the maximum ozone concentrations per month (in ppm) registered in the Southwest of Mexico City during the months of March to May from 2008 to 2015. Fisher–Tippet theorem of extreme value theory states that under certain conditions the probability distribution of the maximum of a random sample is one of the so-called extreme value distributions: either Gumbel, Fréchet or Weibull. In order to find out if any of these distributions provides a reasonable fit to these data using the variance ratio test described in Section 3, `ev_test` function is used with argument `method="ratio"` as follows.

```
> data(o3max)
> o3_max <- o3max[,2] # maximum ozone concentrations
> ev_test(o3_max, dist = "gumbel", method = "ratio",
 N = 10000)

Variance ratio test of fit for the gumbel distribution

data: o3_max T = 1.0742, p-value = 0.1854

> ev_test(o3_max, dist = "frechet", method = "ratio",
 N = 10000)

Variance ratio test of fit for the frechet distribution

data: o3_max T = 1.1188, p-value = 0.0914

> ev_test(-o3_max, dist = "weibull", method = "ratio",
 N = 10000)
```

```
Variance ratio test of fit for the weibull distribution

data: -o3_max T = 1.0382, p-value = 0.3076
```

According to the results of this test, the extreme value Weibull distribution is a plausible model for the (negative) maximum ozone concentrations since the p-value corresponding to the Weibull hypothesis is the highest one. The same conclusion is reached when the correlation test described in Section 5 is used, as it is shown below.

```
> ev_test(o3_max, dist = "gumbel", method = "cor")

Correlation test of fit for the gumbel distribution

data: o3_max R = 0.93628, p-value = 0.02368

> ev_test(o3_max, dist = "frechet", method = "cor")

Correlation test of fit for the frechet distribution

data: o3_max R = 0.92955, p-value = 0.01504

> ev_test(-o3_max, dist = "weibull", method = "cor")

Correlation test of fit for the weibull distribution

data: -o3_max R = 0.97881, p-value = 0.5223
```

For fitting a Weibull distribution to the maximum ozone concentrations, `fitdist` function of `fitdistrplus` package (Delignette-Muller and Dutang [19]) is used . The fitted model has shape and scale parameters equal to 11.22 and 0.177.

```
> wfit <- fitdist(o3_max, "weibull")$estimate; wfit

shape  scale
11.2198575 0.1770641
```

Figure 2 shows a plot of the ECDF of the observations and the fitted Weibull distribution.

7.2. Modelling compressive strength of maize seeds

The inverse Gaussian (IG) distribution with cdf given in Equation (11) is frequently used in survival and reliability analysis for modelling positive data sets with positive asymmetry. The `strength` data set contains 90 measurements on the following two variables: compressive strength (in Newtons) and strain (in millimetres) of maize seeds with floury endosperm and 8% of moisture. For assessing the adequacy of the inverse Gaussian distribution for modelling the compressive strength, the tests based on data transformations described in Section 4 and the ratio test described in Section 3 are performed using

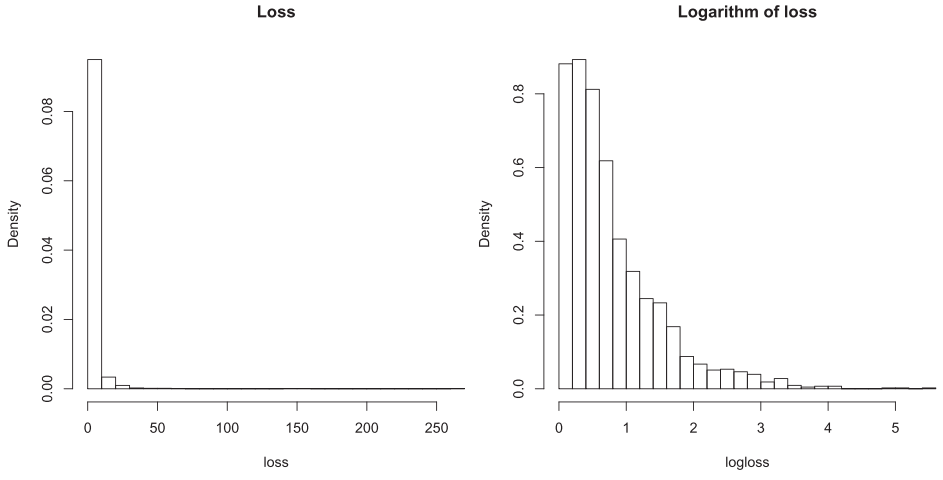

Figure 2. ECDF and fitted Weibull distribution for the monthly maximum ozone levels.

`ig_test` function. The code and results are shown below. Since the *p*-values of the three tests are high, the IG distribution is a plausible distribution for modelling the probability behaviour of maize seeds' compressive strength.

```
> data("strength")
> comp_strength <- strength$cstrength # compressive
  strength
> ig_test(comp_strength, method = "transf")

[[1]]

Test for Inverse Gaussian distributions using a
 transformation to normality

data: comp_strength W = 0.98737, p-value = 0.5403

[[2]]

Test for Inverse Gaussian distributions using a
 transformation to gamma variables

data: comp_strength AD = 0.40897, p-value = 0.691

> ig_test(comp_strength, method = "ratio")

Variance ratio test for Inverse Gaussian distributions

data: comp_strength T1 = 0.14779, p-value = 0.8825
```

For fitting an IG distribution by using the maximum likelihood estimation method, `ig_fit` function is used. The fitted model is an IG(229.26, 516.01) distribution. Figure 3 shows a plot of the ECDF of the observations and fitted IG distribution to the compressive strength variable.

```
> fit <- ig_fit(comp_strength); fit

Inverse Gaussian MLE
mu      229.2609 lambda    516.9174
```

The lognormal distribution is a well-known alternative model for the IG distribution. When applying the transformation test for the lognormal distribution hypothesis mentioned in Section 4 and implemented in `lnorm_test` function, it turns out that this distribution is also a plausible model for seeds' compressive strength. Figure 3 also shows the fitted lognormal distribution. Notice that both models provide rather similar fits on the whole range of the observations.

```
> lnorm_test(comp_strength)

Test for the lognormal distribution based on a
 transformation to normality

data: comp_strength p-value = 0.3734
```

Suppose we are interested in determining the maximum compression force that can be applied to the seeds such that at most 5% of them get broken. If X denote the compression

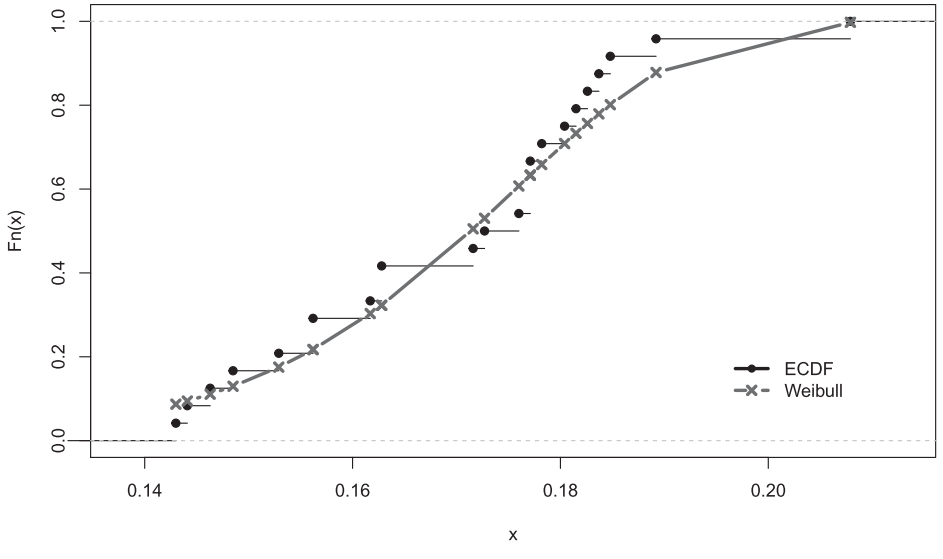

Figure 3. ECDF and fitted inverse Gaussian and lognormal distributions for the compressive strength variable.

strength, the quantity of interest is the 5% quantile of the probability distribution of X, denoted by $x_{0.05}$. An estimator for this quantity is provided by the 5% sample quantile computed from the ECDF as follows, which turns out to be $x_{0.05} = 70.93$ N.

```
> Fn <- ecdf(comp_strength)
> quantile(Fn, probs = .05)

5
70.93595
```

Notice that by using the IG(229.26, 516.01) distribution, $x_{0.05} = 70.98$ N, whereas using the fitted lognormal distribution LNorm(5.24, 0.614), $x_{0.05} = 68.92$ N.

```
> qlnorm(.05, mean(log(comp_strength)),
  sd(log(comp_strength)))

[1] 68.92352
```

Therefore, for having a percentage of damaged seeds less than 5%, maize seeds should not be subjected to a compression force larger than 71 N.

7.3. Modelling Danish reinsurance claims

The `danishuni` data set provided in `fitdistrplus` package [19] comprises 2167 fire losses ≥ 1 million Danish Krone from 1980 to 1990, which were collected at Copenhagen Reinsurance and were adjusted to reflect 1985 values.

7.3.1. Example 1

Figure 4 (left) provides a histogram of variable `loss`, which indicates that the fire losses have a long-tailed distribution. Delignette-Muller and Dutang [19] fitted lognormal and type II Pareto distributions to the this variable and, although they show that none of these models provides a good fit to the observations, by using graphical tools they argue that the lognormal distribution provides a better fit to the right-hand tail of the empirical cumulative distribution function.

Here, the logarithms of the losses (`logloss`) are considered. A histogram of these data is provided in Figure 4 (right). Since the `logloss` data have positive skewness and the right-hand tail decays quickly to zero, the gamma and Weibull distributions with pdf and cdf given in Equations (6) and (32) might be plausible models for this variable.

```
> library(fitdistrplus)
> data("danishuni")
> loss <- danishuni$Loss   # losses
> logloss <- log(loss)   # logarithm of losses

> split.screen(c(1,2))
> screen(1)
```

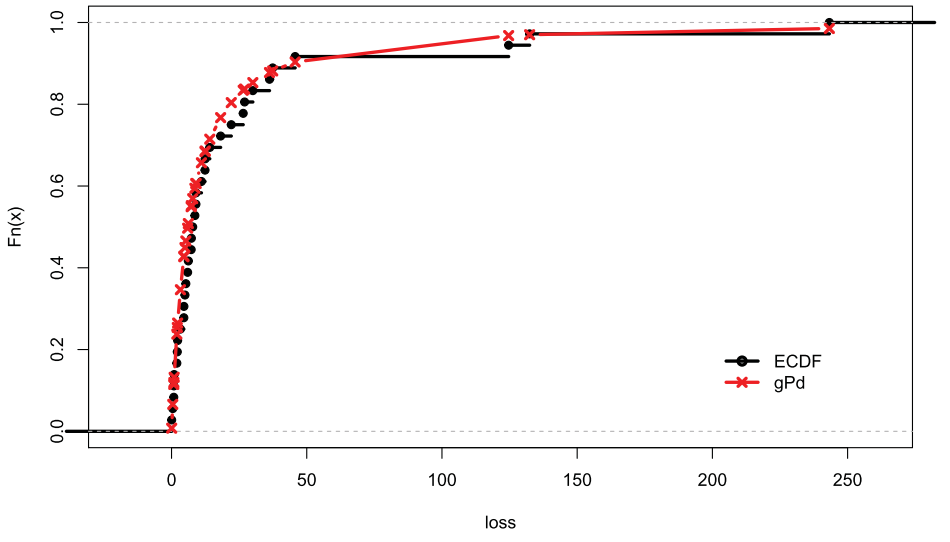

Figure 4. Histograms of loss (left) and logloss (right) variables.

```
> hist(loss, probability = TRUE, nclass = 22, main = "Loss",
   xlab = "loss")
> screen(2)
> hist(logloss, probability = TRUE, nclass = 22, main =
  "Logarithm of loss", +  xlab = "logloss")
> close.screen(all.screens = TRUE)
```

The adequacy of the gamma distribution for modelling the `logloss` variable is assessed using the ratio test described in Section 3 after removing the 11 observations that are equal to zero. This is done with the instructions below. Notice that this test produces a high p-value, supporting the gamma distribution hypothesis.

```
> logloss <- sort(logloss[logloss > 0]) # only positive
 observations are kept
> gamma_test(logloss)     # testing the gamma distribution
  hypothesis

Test of fit for the Gamma distribution

data: logloss V = -0.32803, p-value = 0.8166
```

For testing the Weibull distribution hypothesis, the transformation test mentioned in Section 4 is used. This distribution is also a plausible model for the `logloss` data; however, since the p-value is lower than that of the gamma test, the gamma distribution should be preferable.

```
> weibull_test(logloss, N = 1000) # testing the Weibull
  distribution hypothesis
```

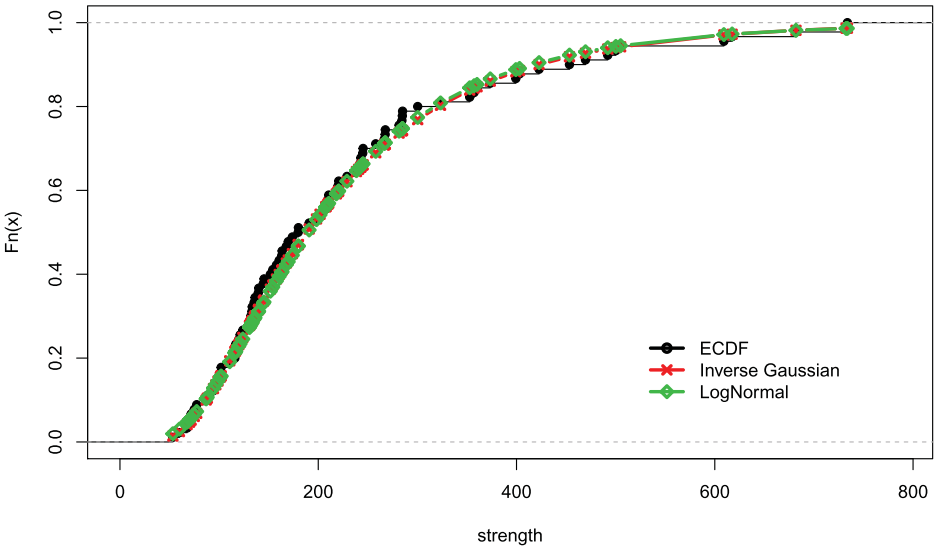

Figure 5. ECDF and fitted gamma distribution to the logarithms of the losses.

```
Test for the Weibull distribution

data: logloss p-value = 0.706
```

For fitting a gamma distribution to the `logloss` data using the parameter estimators given in Equations (8) and (7), `gamma_fit` function is used as shown below. The fitted model has shape and scale parameters equal to 1.2114 and 0.6529.

```
> gam.fit <- gamma_fit(logloss); gam.fit # fitting the
  gamma distribution

Parameter estimates
shape    1.2114035 scale    0.6529328
```

The plot of the ECDF of the observations and fitted gamma distribution shown in Figure 5 also provides evidence in favour of the plausibility of the gamma distribution for modelling `logloss` variable.

Let X be the loss and let $Y = \log X$. Suppose we are interested in estimating the probability of a loss larger than 20 million, which is denoted as $P(X > 20)$. Notice that $P(X > 20) = P(Y > \log 20)$. Thus, the last probability is estimated using the fitted gamma distribution as follows.

```
> pgamma(log(20), shape = gam.fit[1], scale = gam.fit[2],
  lower.tail = FALSE)

[1] 0.01595341
```

A second estimation of $P(X > 20)$ is provided by the ECDF of the losses, which is obtained as follows.

```
> Fn <- ecdf(danishuni$Loss) # ECDF of the losses
> 1 - Fn(20)       # P(X > 20) using the ECDF

[1] 0.01661283
```

These two estimates of $P(X > 20)$ are close to each other; however, the estimate obtained by using the lognormal fit (suggested by [19]) based on MLE is quite smaller. In fact,

```
> plnorm(20, mean(logloss), sd(logloss), lower.tail = FALSE)

[1] 0.001042484
```

On the other hand, notice that the gamma distribution provides good approximations even for very large values of X, as illustrates the following estimation of $P(X > 150)$.

```
> 1 - Fn(150)       # P(X > 150) using the ECDF

[1] 0.0009229349

> pgamma(log(150), shape = gam.fit[1], scale = gam.fit[2],
  lower.tail = FALSE)

[1] 0.0008009203

> plnorm(150, mean(logloss), sd(logloss), lower.tail =
  FALSE)

[1] 1.923298e-09
```

7.3.2. Example 2

Consider the `loss` variable of the `danishuni` data set. An alternative approach for modelling the probabilities of extremal events is by means of considering the excesses over a given high threshold. By using graphical tools, McNeil [30] argues that the excesses over the 20 million threshold can be modelled by a generalized Pareto distribution. In order to formally assess this assumption, `gp_test` function is used as indicated below. The approximated p-value of the gPd test is around 0.34, which supports the plausibility of the gPd hypothesis. Furthermore, the results indicate that the excesses follow a gPd with positive shape parameter since the p-value of the test for H_0^+ is high. By using `gp_fit` function with argument `method = "amle"`, the fitted distribution to the excesses turns out to be a GP(0.8209,6.4528).

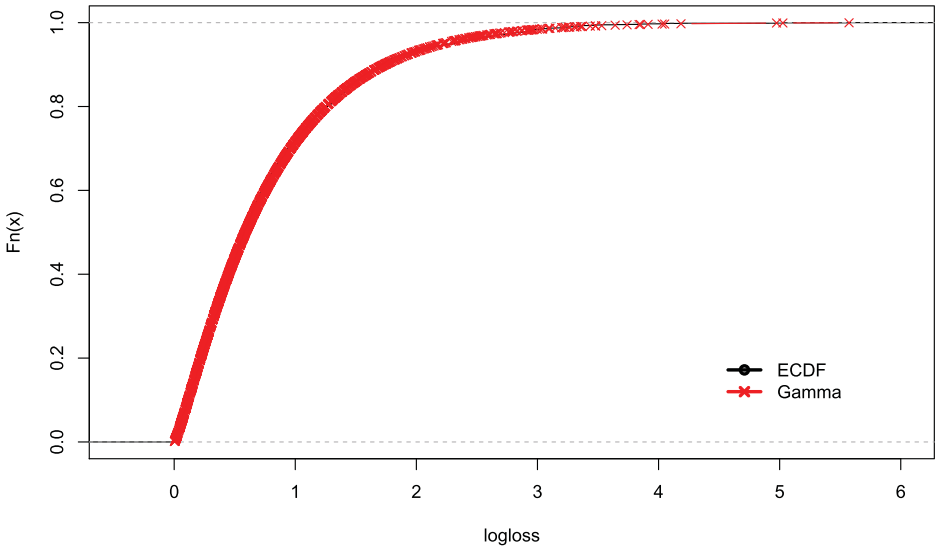

Figure 6. ECDF and fitted generalized Pareto distribution to the excesses over threshold $u = 20$ million.

```
> loss <- danishuni$Loss
> excess <- loss[loss > 20] -20
> results <- gp_test(excess); results

    Bootstrap test of fit for the generalized Pareto distribution

data:  excess
p-value = 0.3433

> results$pvalues

                                                                p.value      R
H_0^-: excess follows a gPd with NEGATIVE shape parameter  0.0000 0.8752
H_0^+: excess follows a gPd with POSITIVE shape parameter  0.3433 0.9800

> gpfit <- gp_fit(excess, "amle"); gpfit

      Parameter estimates
shape              0.8209
scale              6.4528
```

Figure 6 supports the conclusion about the plausibility of the gPd with positive shape parameter for modelling the conditional distribution of the excesses over the $u = 20$ million threshold.

Estimates of the probability of having an excess larger than 50 million given that $u = 20$ million are obtained as follows.

```
> Fn_excess <- ecdf(excess)    # ECDF of the excesses
> 1 - Fn_excess(50)      # P(excess > 50) using the ECDF
```

```
[1] 0.08333333

> 1 - pgp(50, gpfit[1], gpfit[2]) # P(excess > 50) using
 the fitted gPd

[1] 0.08788872

> 1 - pgp(50, .684, 9.63)  # P(excess > 50) using McNeil's
  fitted gPd

[1] 0.1090936
```

8. Conclusions

Package `goft` provides `R` functions for validating distributional assumptions on data sets to which in turn parametric fitting techniques might be applied, like those found in packages `fitdistrplus` [19] and `distrMod` [31]. Some additional extensions of the `goft` package are in progress of implementation such as tests for other important distributions, which are useful for modelling asymmetrical data, as well as other types of tests. Some new parameter estimation methods are also being considered.

Acknowledgments

The authors are grateful to an anonymous reviewer for her/his constructive comments on the original version of the manuscript. They are also grateful to Arturo Mancera-Rico and Ernesto J. Dorantes-Coronado for kindly allowing us to include the strength and goats data sets in this package.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

E. González-Estrada http://orcid.org/0000-0002-3086-0605
J. A. Villaseñor http://orcid.org/0000-0002-2091-6038

References

[1] R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2017.
[2] Miecznikowski JC, Vexler A, Shepherd L. dbEmpLikeGOF: an R package for nonparametric likelihood ratio tests for goodness-of-fit and two-sample comparisons based on sample entropy. J Stat Softw. 2013;54(3):1–19.
[3] Faraway J, Marsaglia G, Marsaglia J, et al. goftest: Classical goodness-of-fit tests for univariate distributions. R package version 1.1-1; 2017.
[4] Kojadinovic I, Yan J. fgof: Fast Goodness-of-fit Test. R package version 0.2-1; 2012.
[5] Pavia JM. Testing goodness-of-fit with the kernel density estimator: GoFKernel. J Stat Softw Code Snippets. 2015;66(1):1–27.
[6] Korkmaz S, Goksuluk D, Zararsiz G. Mvn: multivariate normality tests. R J. 2014;6(2): 151–162.

[7] Rizzo ML, Szekely GJ. energy: E-statistics (energy statistics). R package version 1.7-0; 2016.
[8] Gross J, Ligges U. nortest: Tests for normality. R package version 1.0-4; 2015.
[9] Lafaye de Micheaux P, Tran VA, PoweR: a reproducible research tool to ease monte carlo power simulation studies for goodness-of-fit tests in R. J Stat Softw. 2016;69(3):1–42.
[10] González-Estrada E, Villaseñor JA. goft: Tests of fit for some probability distributions. R package version 1.3.4; 2017.
[11] Thode H. Testing for normality. New York: Marcel Dekker; 2002.
[12] Shapiro SS, Wilk MB. An analysis of variance test for normality: complete samples. Biometrika. 1965;52(3):591–611.
[13] Royston P. Approximating the shapiro–wilk w test for non-normality. Stat Comput. 1992;2:117–119.
[14] Villaseñor JA, González-Estrada E. A generalization of shapiro-wilk's test for multivariate normality. Commun Statist Theory Methods. 2009b;38(11):1870–1883.
[15] Villaseñor JA, González-Estrada E. A variance ratio test of fit for gamma distributions. Statist Probab Lett. 2015c;96(11):281–286.
[16] Cox D, Oakes D. Analysis of survival data. Boca Raton: Chapman and Hall/CRC; 1984.
[17] Villaseñor JA, González-Estrada E. Tests of fit for inverse gaussian distributions. Statist Probab Lett. 2015b;105:189–194.
[18] González-Estrada E, Villaseñor JA. A ratio goodness-of-fit test for the laplace distribution. Statist Probab Lett. 2016;119:30–35.
[19] Delignette-Muller ML, Dutang C. fitdistrplus: An R package for fitting distributions. J Stat Softw. 2015;64(4):1–34.
[20] Kimball BF. The bias in certain estimates of the parameters of the extreme-value distribution. Ann Math Statist. 1956;27(3):758–767.
[21] Johnson N, Kotz S, Balakrishnan N. Continuous univariate distributions, Vol. 1. New York: Wiley; 1994.
[22] Anderson TW, Darling DA. Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes. Ann Math Statist. 1952;23:193–212.
[23] Ochoa A. Tests for the inverse gaussian distribution [Master's thesis]. Mexico: Colegio de Postgraduados; 2015.
[24] Villaseñor JA, González-Estrada E. A transformation test for exponentiality. Technical report, Colegio de Postgraduados; 2016.
[25] Villaseñor JA, González-Estrada E. A correlation test for normality based on the lévy characterization. Commun Stat Simul Comput. 2015a;44:1225–1238.
[26] González-Estrada E, Villaseñor JA. A goodness-of-fit test for location-scale max-stable distributions. Commun Stat Simul Comput. 2010;39:557–562.
[27] Reiss R, Thomas M. Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields. 2nd ed. Basel: Birkhäuser; 2001.
[28] Villaseñor JA, González-Estrada E. A bootstrap goodness of fit test for the generalized Pareto distribution. Comput Stat Data Anal. 2009a;53:3835–3841.
[29] Pickands J. Statistical inference using extreme order statistics. Ann Stat. 1975;3:119–131.
[30] McNeil AJ. Estimating the tails of loss severity distributions using extreme value theory. ASTIN Bull. 1997;27(1):117–137.
[31] Kohl M, Ruckdeschel P. R package distrMod: S4 classes and methods for probability models. J Stat Softw. 2010;35(10):1–27.