vsgoftest: An R Package for Goodness-of-Fit Testing Based on Kullback-Leibler Divergence

Justine Lequesne, Philippe Regnault

vsgoftest: An R Package for Goodness-of-Fit Testing Based on Kullback-Leibler Divergence

Justine Lequesne
Centre Henri Becquerel

Philippe Regnault
Université de Reims Champagne-Ardenne

Abstract

The R-package **vsgoftest** performs goodness-of-fit (GOF) tests, based on Shannon entropy and Kullback-Leibler divergence, developed by Vasicek (1976) and Song (2002), of various classical families of distributions. The theoretical framework of the so-called Vasicek-Song (VS) tests is summarized and followed by a detailed description of the different features of the package. The power and computational time performances of VS tests are studied through their comparison with other GOF tests. Application to real datasets illustrates the easy-to-use functionalities of the **vsgoftest** package.

Keywords: R statistical computing environment, goodness-of-fit tests, Shannon entropy, Kullback-Leibler divergence, sample spacing based estimation.

1. Introduction

Goodness-of-fit (GOF) tests constitute a classical tool in deciding of the compatibility of data with a theoretical (probability) distribution. The present work proposes a package for the R statistical computing environment R Core Team (2017) performing GOF tests based on Shannon entropy and Kullback-Leibler divergence, together with a methodological guide and applications.

Precisely, we consider fitting numeric (real valued) data either to a unique distribution, the so-called simple null hypothesis test

$$H_0 : P = P_0(\theta) \quad \text{against} \quad H_1 : P \neq P_0(\theta), \tag{1}$$

or to a parametric family, the so-called composite null hypothesis test

$$H_0 : P \in \mathcal{P}_0(\Theta) \quad \text{against} \quad H_1 : P \notin \mathcal{P}_0(\Theta). \tag{2}$$

The set $\mathcal{P}_0(\Theta) = \{P \in \mathcal{D} : P = P_0(\theta), \theta \in \Theta\}$, with $\Theta \subset \overline{\mathbb{R}}^d$, is a parametric subfamily of the set $\mathcal{D}$ of all probability distributions absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$, *i.e.*, probability distributions with a density function. Decision is to be taken from the observation $x_1^n = (x_1, \ldots, x_n)$ of a sample $X_1^n = (X_1, \ldots, X_n)$ of size n of independent and identically distributed random variables drawn from $P \in \mathcal{D}$.

Classically, a GOF test procedure is derived by computing some distance-like functional between the observations and the null distribution, or family of distributions, the null hypothesis being rejected when the distance is larger than a critical value. Kolmogorov-Smirnov, Cramér-von Mises and Anderson-Darling tests constitute some of the most commonly used GOF tests.

Their test statistics measure discrepancy between the empirical cumulative distribution function of the sample and the cumulative distribution function of the null distribution; these tests are refered to as EDF tests in the following; see Stephens (1974). The tests in this paper are based on the Kullback-Leibler (KL) divergence of the density of the sample with respect to the null density.

GOF tests based on KL divergence have been introduced by Vasicek (1976) for testing normality. Vasicek normality test relies on the maximum entropy property satisfied by the normal distribution: amoung all distributions with density and finite variance, Shannon entropy is maximized by the normal distribution. Vasicek test statistic is a monotone function of the entropy difference between the null normal distribution and the observed one. This has been subsequently extended to GOF tests of uncategorical data for numerous families of distributions satisfying a maximum entropy property – say maximum entropy (ME) distributions; see Section 2 for references. GOF tests based on entropy differences are known to have higher power than classical GOF tests in numerous cases; see Section 4.1. Song (2002) considers GOF tests based on KL divergence, for a large class of distributions including all classical distribution families. The test statistic is an estimate of the KL divergence between the sample and the null distributions. It is asymptotically normally distributed. When applied to ME distributions, it is equal to the difference between the entropies of the null distribution and the sample one, yielding the same decision rule as Vasicek test. This paper presents the implementation of Vasicek and Song tests (VS tests) for various families of distributions: uniform, normal, log-normal, exponential, gamma, Weibull, Pareto, Fisher, Laplace and Beta distributions. For further details on the theoretical aspects of VS tests, see Girardin and Lequesne (2017) in which a unifying framework for tests based on entropy difference and KL divergence is provided.

Numerous R packages perform GOF tests for various families of distributions. The functions `chisq.test`, `ks.test` and `shapiro.test` of the **stats** package perform respectively the chi-squared test of adequacy to a discrete distribution, the Kolmogorov-Smirnov GOF test for any theoretical continuous distribution and the Shapiro-Wilk normality test. The packages **goftest** developed in Faraway, Marsaglia, Marsaglia, and Baddeley (2015) and **goft** in Gonzalez-Estrada and Villasenor-Alva (2016) perform respectively Cramér-von Mises and Anderson-Darling GOF tests, and tests based on the ratios of variance and other moment estimators. **KScorrect** in Novack-Gottshall and Wang (2016) performs the Lilliefors-corrected Kolmogorov-Smirnov GOF test. Numerous GOF tests of the exponential or two-parameter Weibull distributions are available in **EWGoF** Krit (2015) while **nortest** and **normtests** are dedicated to testing normality. The **dbEmpLikeGOF** package developed in Miecznikowski, Vexler, and Shepherd (2013) proposes GOF normality and uniformity tests based on empirical likelihood ratio. These tests are closely related with VS tests; similarities and differences between them are highlighted in the following sections.

The test procedure implemented in **vsgoftest** uses either the asymptotic distribution of the test statistic or Monte-Carlo simulation, depending on sample size or user's choice. Optional arguments are included for handling particular situations such as samples with numerous ties. They also contribute to make the procedure flexible and fully parameterizable. Besides these practical aspects, the paper presents a comprehensive review of the literature dealing with power properties of VS tests. Monte Carlo simulations are conducted to illustrate their performance when applied to discriminate between close distributions.

The paper is organized as follows. The theoretical framework of GOF tests based on Shannon

entropy difference and KL divergence is briefly presented in Section 2. The functionalities of **vsgoftest** are presented in Section 3. The tests performed by **vsgoftest** are compared to other GOF tests in Section 4. More precisely, power comparisons to classical GOF tests are presented in Section 4.1; Section 4.2 focuses on the comparison of **vsgoftest** and **dbEmpLikeGOF** test procedures, which rely on very close theoretical frameworks but significantly differ in some of their features. Finally, applications to real data in Section 5 illustrate the usage of the proposed functionalities.

2. Entropy difference and KL divergence based GOF tests

The Shannon entropy of a distribution P with density function p on $\mathbb{R}$ has been defined in Shannon (1948) as

$$\mathbb{S}(P) := -\int_{\mathbb{R}} p(x)\log p(x)\mathrm{d}x. \tag{3}$$

Entropy measures the uncertainty or variability of a distribution. The maximum entropy principle under moment constraints, or ME method, favours distributions with highest entropy for their highest degree of uncertainty; see Shannon (1948) and Jaynes (1957). Among all distributions supported by a given finite length interval I in $\mathbb{R}$, entropy is maximum and equals $\log|I|$ for the uniform distribution, where $|I|$ denotes the length of I. Hence, the entropy difference $\log|I| - \mathbb{S}(P)$ can be thought as a distance-like measure between P and the uniform distribution.

Similarly, among all continuous distributions supported within $\mathbb{R}$ with mean μ and variance σ^2, Shannon entropy is maximum for the normal distribution $\mathcal{N}(\mu,\sigma^2)$ and equals

$$\mathbb{S}(\mathcal{N}(\mu,\sigma^2)) = \ln(\sigma\sqrt{2\pi e}). \tag{4}$$

The entropy difference $\mathbb{S}(\mathcal{N}(\mu,\sigma^2)) - \mathbb{S}(P)$ is nonnegative and thus defines a distance-like measure between any distribution with mean μ and variance σ^2 and the $\mathcal{N}(\mu,\sigma^2)$ distribution. Based on this property, Vasicek (1976) derives a normality test, with a test statistic expressed in terms of entropy differences, defined as follows

$$K_{mn} := \frac{n}{2mS}\left\{\prod_{i=1}^{n}(X_{(i+m)} - X_{(i-m)})\right\}^{1/n} = \sqrt{2\pi e}\exp\left(V_{mn} - \mathbb{S}\left(\mathcal{N}\left(\overline{X}, S^2\right)\right)\right),$$

where

$$\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

are the empirical estimators of respectively the mean and the variance of the sample $X_1^n := (X_1,\ldots,X_n)$, $X_{(1)} \leq \cdots \leq X_{(n)}$ denotes the order statistics associated to X_1^n and

$$V_{mn} := \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{n}{2m}\left[X_{(i+m)} - X_{(i-m)}\right]\right) \tag{5}$$

is the non-parametric Vasicek estimator of $\mathbb{S}(P)$ based on spacings, with $X_{(i)} = X_{(1)}$ if $i < 1$ and $X_{(i)} = X_{(n)}$ if $i > n$; the window size $m \in \mathbb{N}^*$ is smaller than $n/2$.

The test statistic has been extended to various families of ME distributions under moment constraints; see Dudewicz and Van Der Meulen (1981), Ebrahimi, Habibullah, and Soofi (1992), Choi and Kim (2006), Mergel (1999), Mudholkar and Tian (2002), among many others. A unifying framework for any exponential family of distributions is proposed in Girardin and Lequesne (2017), with asymptotic properties, consistency and application to biology; see also Lequesne (2013), Lequesne (2015b) and Lequesne (2015a) for power efficiencies, GOF tests of Pareto distributions and extension to generalized entropies.

The Kullback-Leibler (KL) divergence of a distribution P with respect to another one Q, is defined as

$$\mathbb{K}(P|Q) := \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x, \tag{6}$$

if P is absolutely continuous with respect to Q, with respective densities p and q, and as $+\infty$ if not; see Kullback and Leibler (1951). The KL divergence is linked to Shannon entropy through the relation

$$\mathbb{K}(P|Q) = -\mathbb{S}(P) - \int_{\mathbb{R}} p(x) \log q(x) \mathrm{d}x. \tag{7}$$

The KL divergence is not a mathematical distance because of lack of both symmetry and triangular inequality, but it satisfies $\mathbb{K}(P|Q) \geq 0$, with $\mathbb{K}(P|Q) = 0$ if and only if $P = Q$, and thus constitutes a natural measure of discrepancy for GOF tests. Song (2002) proposes GOF tests based on KL divergence for either simple (1) or composite (2) null hypothesis. Precisely, thanks to (7), the test statistic I_{mn} is the estimator of $\mathbb{K}(P|P_0(\theta))$, defined by

$$I_{mn} := -V_{mn} - \frac{1}{n} \sum_{i=1}^{n} \log p_0(X_i, \widehat{\theta}_n), \tag{8}$$

where V_{mn} is the Vasicek estimator (5) of $\mathbb{S}(P)$, and $\widehat{\theta}_n$ is either the maximum likelihood estimator (MLE) of θ satisfying

$$\frac{1}{n} \sum_{i=1}^{n} \log p_0(X_i, \widehat{\theta}_n) = \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_0(X_i, \theta),$$

or θ itself in case of a simple null hypothesis (1).

The KL divergence $\mathbb{K}(P|Q)$ for a maximum entropy distribution Q under moment constraints reduces to the entropy difference $\mathbb{S}(Q) - \mathbb{S}(P)$ for all P satisfying the same moment constraints; see Csiszár (1975). This Pythagorean equality allowed Girardin and Lequesne (2017) to establish that entropy difference GOF test for ME distributions coincide with Song test – we will refer to these tests as Vasicek-Song tests and keep on denoting them by VS tests. Especially, Vasicek and Song normality test statistics are linked through the equality

$$K_{mn} = \sqrt{2\pi e} \exp(-I_{mn}),$$

yielding identical decision rules.

Based on the asymptotic properties of V_{mn} proven by Dudewicz and Van Der Meulen (1981) for testing uniformity, Song (2002) establishes the asymptotic behavior of I_{mn}, independently of the null hypothesis: I_{mn} is consistent and asymptotically normally distributed provided the null distribution belongs to the class

$$\mathcal{F} = \left\{ P \in \mathcal{D} : \sup_{x:\, 0<F(x)<1} \frac{|p'(x)|}{p^2(x)} F(x)[1 - F(x)] < \gamma,\ \gamma > 0 \right\}, \tag{9}$$

where F is the cumulative distribution function of P and p its density with derivative p' (almost everywhere). The class $\mathcal{F}$ contains the most classical distributions such as uniform ($\gamma = 0$), normal, exponential and gamma ($\gamma = 1$), Fisher ($\gamma = (2 + \nu_2)/\nu_2$ where ν_2 is the second degree of freedom), Pareto ($\gamma = (\mu + 1)/\mu$, where μ is the shape parameter), etc. For $\mathcal{P}_0(\Theta) \subset \mathcal{F}$, if

$$m/\log n \xrightarrow[n\to\infty]{} 0 \quad \text{and} \quad m(\log n)^{2/3}/n^{1/3} \xrightarrow[n\to\infty]{} 0, \tag{10}$$

then

$$\sqrt{6mn}[I_{mn} - \log(2m) + \psi(2m)] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{11}$$

where $\psi(m)$ is the digamma function. The asymptotic bias $\log(2m) - \psi(2m)$ of I_{mn} is that of $-V_{mn}$. Song (2002) suggests a bias correction in the asymptotic distribution (11) for moderate sample sizes:

$$\sqrt{6mn}\,[I_{mn} - b_{mn}] \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \tag{12}$$

where

$$b_{mn} := \log(2m) - \log(n) - \psi(2m) + \psi(n+1) + \frac{2m}{n} R_{2m-1} - \frac{2}{n}\sum_{i=1}^{m} R_{i+m-2},$$

with $R_m := \sum_{j=1}^{m} 1/j$. From (12), an asymptotic p-value for the related VS test is given by

$$p = 1 - \Phi^{-1}\left(\sqrt{6mn}\,[I_{mn}(x_1^n) - b_{mn}]\right), \tag{13}$$

where $I_{mn}(x_1^n)$ denotes the value of the statistic I_{mn} for the observations $x_1^n = (x_1, \ldots, x_n)$, and Φ denotes the cumulative distribution function of the normal distribution. According to Song (2002), the asymptotic p-value (13) provides accurate results for sample sizes n larger than 80. For small sample sizes, Monte Carlo simulations should be preferred. A large number N of replications of X_1^n drawn from $P_0(\widehat{\theta}_n)$ (or $P_0(\theta)$ in case of simple null hypothesis) are generated. The test statistic I_{mn}^i is computed for each replication $i, 1 \le i \le N$. The p-value is then given by the empirical mean $(\sum_{i=1}^{N} \mathbb{1}_{\{I_{mn}^i > I_{mn}(x_1^n)\}})/N$.

For choosing m, Song (2002) proposes to minimize I_{mn} – that is to maximize V_{mn}, with respect to m, yielding the most conservative test. The KL divergence $\mathbb{K}(P|P_0(\theta))$ being nonnegative, values of m for which I_{mn} is negative are excluded, leading to choose m subject to the constraint

$$V_{mn} \le -\frac{1}{n}\sum_{i=1}^{n} \log p_0\left(.; \widehat{\theta}_n\right). \tag{14}$$

Finally, the window size proposed by Song (2002) – say the optimal window size, is

$$\widehat{m} := \min\left\{ m^* \in \operatorname*{argmax}_{m\in\mathbb{N}^*} \left\{ V_{mn} :\ V_{mn} \le -\frac{1}{n}\sum_{i=1}^{n} \log p_0\left(X_i, \widehat{\theta}_n\right) \right\} : 1 \le m^* < \lfloor n^{1/3-\delta} \rfloor \right\}, \tag{15}$$

for some $\delta < 1/3$ and the VS test statistic is then

$$I_{\widehat{m}n} = -V_{\widehat{m}n} - \frac{1}{n}\sum_{i=1}^{n} \log p_0\left(X_i; \widehat{\theta}_n\right). \tag{16}$$

The upper bound $n^{1/3-\delta}$ for the window size m is chosen so that conditions (10) are fulfilled and hence that asymptotic normality (11) holds. No optimal choice of δ exists; it depends on the family of distributions of the null hypothesis; see Section 3 for details.

The package **vsgoftest** presented below performs VS GOF tests for several parametric families of ME distributions: uniform, normal, log-normal, exponential, Pareto, Laplace, Weibull, Fisher, gamma and beta distributions. These families of distribution, all included in the class $\mathcal{F}$ given by (9), have been chosen so that the package covers a large variety of applications. Note that the package **dbEmpLikeGOF** performs uniformity and normality VS tests, with an alternative choice for the window size. Precisely, the test statistic is $nI_{mn} + 1/2$, and the window m is chosen, between 1 and $n^{1/2}$, minimizing nI_{mn}. The constraint (14) is not considered. The asymptotic distribution of nI_{mn} is not used, p-values being computed from a pre-calculated table for small sample sizes or via Monte-Carlo simulation; see Miecznikowski *et al.* (2013) and Vexler and Gurevich (2010). This alternative methodological approach leads to different decisions that may be less reliable, particularly when applied to heavy tailed samples. Other differences in the coding structure make **vsgoftest** faster than **dbEmpLikeGOF**, especially when Monte-Carlo simulation is performed. These points will be detailled in Section 4.2.

3. The package vsgoftest

The **vsgoftest** package provides functions for estimating Shannon entropy of absolutely continuous distributions and testing the goodness-of-fit of some theoretical family of distributions to a vector of real numbers. It also provides functions for computing the density, cumulative density and quantile functions of Pareto and Laplace distributions, as well as for generating samples from these distributions.

The **vsgoftest** package is available on CRAN mirrors and can be installed by executing the command

```
install.packages('vsgoftest')
```

Alternatively, the latest (under development) version of the **vsgoftest** package is also available and can be installed in R from the github repository of the project as follows:

```
#Package devtools must be installed
devtools::install_github(repo = 'pregnault/vsgoftest')
```

The package is structured around two functions, `entropy.estimate` and `vs.test`; the first one computes the spacing based estimator (5) from a numeric sample, the second one performs Vasicek-Song GOF test for usual parametric families of distributions based on the test statistic (16). A comprehensive presentation of their usage is proposed in Sections 3.1 and 3.2, with numerous examples. Section 3.3 provides further technical information about the structure of the package.

3.1. Function entropy.estimate for estimating Shannon entropy

The function `entropy.estimate` computes the spacing based estimate (5) of Shannon entropy (3) from a numeric sample. Two arguments have to be provided:

- `x`: the numeric sample;
- `window`: an integer between 1 and half of the sample size, specifying the window size of the spacing-based estimator (5).

It returns the estimate of Shannon entropy of the sample. Here is an example for a sample drawn from a normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$.

```
library('vsgoftest')

Loading required package:  fitdistrplus
Loading required package:  MASS
Loading required package:  survival

set.seed(2)      #set seed of PRNG
samp <- rnorm(n = 100, mean = 0, sd = 1) #sampling from normal distribution
entropy.estimate(x = samp, window = 8) #estimating entropy with window = 8

[1] 1.394728

log(2*pi*exp(1))/2 #the exact value of entropy

[1] 1.418939
```

The estimate returned by `entropy.estimate` obviously depends on the window selected by the user, as illustrated by the following chunck.

```
sapply(1:10, function(w) entropy.estimate(x = samp, window =w))

 [1] 1.205018 1.346352 1.378732 1.387337 1.391691 1.393512 1.394428
 [8] 1.394728 1.394486 1.392669
```

One may select the window size that maximizes the entropy estimate, as follows.

```
n <- 100 #sample size
V <- sapply(1:(n/2 - 1), function(w) entropy.estimate(x = samp, window =w))
which.max(V) #Choose window that maximizes entropy

[1] 8
```

Let us consider a sample drawn from a Pareto distribution with density

$$p(x; c, \mu) = \frac{\mu c^\mu}{x^{\mu+1}}, \quad x \geq c,$$

where $c > 0$ and $\mu > 0$, which can be obtained by making use of the function `rpareto` as illustrated below. Its Shannon entropy is

$$\mathbb{S}(p(.; c, \mu)) = -\ln \mu + \ln c + \frac{1}{\mu} + 1.$$

```
set.seed(5)
n <- 100 #Sample size
samp <- rpareto(n, c = 1, mu = 2) #sampling from Pareto distribution
entropy.estimate(x = samp, window = 3)

[1] 0.8480204

-log(2) + 3/2 #Exact value of entropy

[1] 0.8068528
```

3.2. Function vs.test for testing GOF to a specified model

The function `vs.test` performs the VS test, as described in Section 2; setting two non-optional arguments is required:

- `x`: the numeric sample;
- `densfun`: a character string specifying the theoretical family of distributions of the null hypothesis. Available families of distributions are: uniform, normal, log-normal, exponential, gamma, Weibull, Pareto, Fisher and Laplace distributions. They are referred to by the symbolic name in R of their density function. For example, set `densfun = 'dnorm'` to test GOF of the family of normal distributions; see Table 1 for details.

It returns an object of class `htest`, *i.e.*, a list whose main components are:

- `statistic`: the value of VS test statistic (16) for the sample, with optimal window size defined by (15);
- `parameter`: the optimal window size;
- `estimate`: the maximum likelihood estimate of the parameters of the null distribution (for the test (2) with composite null hypothesis);
- `p.value`: the p-value associated to the sample.

By default, `vs.test` performs the composite VS test of the family of distributions `densfun` for the sample `x`. The p-value is estimated by means of Monte-Carlo simulation if the sample size is smaller than 80, or through the asymptotic distribution (11) of the VS test statistic otherwise.

In the following example, a normally distributed sample is simulated. VS test rejects the null hypothesis that this sample is drawn from a Laplace distribution, but does not reject the normality hypothesis (for a significant level set to 0.05).

```
set.seed(5)
samp <- rnorm(50,2,3)
vs.test(x = samp, densfun = 'dlaplace')
```

Distribution	**Call** (`densfun`)	**Parameters**	**Density**	**Default** δ
Uniform	`"dunif"`	$a < b$	$\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	1/12
Normal	`"dnorm"`	$\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	1/12
Log-normal	`"dlnorm"`	$\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$	$\frac{1}{x\sqrt{2\pi}\sigma}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right)\mathbb{1}_{\mathbb{R}_+}(x)$	1/12
Exponential	`"dexp"`	$\lambda \in \mathbb{R}$	$\lambda\exp(-\lambda x)\mathbb{1}_{\mathbb{R}_+}(x)$	1/12
Pareto	`"dpareto"`	$c \in \mathbb{R}, \mu \in \mathbb{R}_+$	$\mu c^\mu \frac{1}{x^{\mu+1}}\mathbb{1}_{[c,\infty[}(x)$	1/12
Laplace	`"dlaplace"`	$\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+$	$\frac{1}{2\sigma}\exp\left(-\frac{\lvert x-\mu\rvert}{\sigma}\right)$	1/12
Weibull	`"dweibull"`	$a, b \in \mathbb{R}_+$	$\frac{a}{b^a}x^{a-1}\exp\left[-\left(\frac{x}{b}\right)^a\right]\mathbb{1}_{\mathbb{R}_+}(x)$	2/15
Fisher	`"df"`	$\nu_1, \nu_2 \in \mathbb{R}_+$	$\frac{\left(\frac{d_1 x}{d_1 x+d_2}\right)^{d_1/2}\left(1-\frac{d_1 x}{d_1 x+d_2}\right)^{d_2/2}}{xB(d_1/2, d_2/2)}\mathbb{1}_{\mathbb{R}_+}(x)$	2/15
Gamma	`"dgamma"`	$\alpha, \beta \in \mathbb{R}_+$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\exp(-\beta x)\mathbb{1}_{\mathbb{R}_+}(x)$	2/15
Beta	`"dbeta"`	$\alpha, \beta \in \mathbb{R}$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\mathbb{1}_{\mathbb{R}_+}(x)$	2/15

Table 1: Families of distributions supported as the null by the **vsgoftest** package. The column denoted "Default δ" corresponds to the default setting for the parameter δ; see (15). Note that $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$ is the Beta function.

```
Vasicek-Song GOF test for the Laplace distribution

data:  samp
Test statistic = 0.32437, Optimal window = 2, p-value = 0.0248
sample estimates:
   Shape    Scale
2.194803 2.687321
```

```
set.seed(4)
vs.test(x = samp, densfun = 'dnorm')


Vasicek-Song GOF test for the normal distribution

data:  samp
Test statistic = 0.21655, Optimal window = 2, p-value = 0.3704
sample estimates:
    Mean St. dev.
2.194803 3.173824
```

For performing a simple null hypothesis GOF test, the additional argument `param` has to be set to a numeric vector, consistent with the parameter requirements for the null distribution. In such case, the MLE of the parameter(s) of the null distribution has not to be computed and hence the component `estimate` in results is not available.

```
set.seed(26)
vs.test(x = samp, densfun = 'dnorm', param = c(2,3))


Vasicek-Song GOF test for the normal distribution with Mean=2,
St. dev.=3

data:  samp
Test statistic = 0.22196, Optimal window = 2, p-value = 0.331
```

If `param` is not consistent with the specified distribution – e.g., standard deviation for testing a normal distribution is missing or negative, the execution is stopped and an error message is returned.

```
set.seed(2)
samp <- rnorm(50, -2, 1)
vs.test(samp, densfun = 'dnorm', param = -2)

Error in vs.test(samp, densfun = "dnorm", param = -2):  "param":  invalid
parameter (not consistent with the specified distribution)
```

One can estimate the p-value of the sample by Monte-Carlo simulation, even when sample size is larger than 80, by setting the optional argument `simulate.p.value` to `TRUE` (`NULL` by default). The number of Monte-Carlo replicates can be fixed through the optional argument `B` (default is `B = 5000`).

```
set.seed(1)
samp <- rweibull(200, shape = 1.05, scale = 1)
set.seed(2)
vs.test(samp, densfun = 'dexp', simulate.p.value = TRUE, B = 10000)


Vasicek-Song GOF test for the exponential distribution

data:  samp
Test statistic = 0.10907, Optimal window = 3, p-value = 0.3504
sample estimates:
   Rate
1.15047
```

Vasicek's estimates V_{mn} are computed for all m from 1 to $n^{1/3-\delta}$, where $\delta < 1/3$; the test statistic is $I_{\widehat{m}n}$ for $\widehat{m}$ the optimal window size, as defined in (15). The choice of δ depends on the family of distributions of the null hypothesis. Precisely, for Weibull, Pareto, Fisher, Laplace and Beta, δ is set by default to $2/15$, while for uniform, normal, log-normal, exponential and gamma, it is set to $1/12$. These default settings result from numerous experimentations. Still, the user can choose another value through the optional argument `delta`.

```
set.seed(63)
vs.test(samp, densfun = 'dexp', delta = 5/30)


Vasicek-Song GOF test for the exponential distribution

data:  samp
Test statistic = 0.16517, Optimal window = 2, p-value = 0.1538
sample estimates:
   Rate
1.15047
```

Note that upper-bounding the window size by $n^{1/3-\delta}$ is only required when the asymptotic normality of I_{mn} is used to compute asymptotic p-values from (11). When the p-value are computed by means of Monte-Carlo simulation, this upper-bound can be extended to $n/2$ by adding `extend = TRUE`, which may lead to a more reliable test, as illustrated below.

```
set.seed(8)
samp <- rexp(30, rate = 3)
vs.test(x = samp, densfun = "dlnorm")


Vasicek-Song GOF test for the log-normal distribution

data:  samp
Test statistic = 0.30717, Optimal window = 2, p-value = 0.1206
sample estimates:
 Location     Scale
-2.162290  1.683868
```

```
vs.test(x = samp, densfun = "dlnorm", extend = TRUE)


Vasicek-Song GOF test for the log-normal distribution

data:  samp
Test statistic = 0.3029, Optimal window = 3, p-value = 0.007
sample estimates:
 Location     Scale
-2.162290  1.683868
```

Enlarging the range of m is also pertinent if ties are present in the sample. Indeed, the presence of ties is particularly inappropriate for performing VS tests, because some spacings $X_{(i+m)} - X_{(i-m)}$ can be null. The window size m has thus to be greater than the maximal number of ties in the sample. Hence, if the upper-bound $n^{1/3-\delta}$ is less than the maximal

number of ties, the test statistic can not be computed. Setting `extend` to `TRUE` can avoid this behavior, as illustrated below.

```
samp <- c(samp, rep(4,3)) #add ties in the previous sample
vs.test(x = samp, densfun = "dexp")

Warning in vs.estimate(x, densfun, ESTIM, extend, delta, relax):  Ties should
not be present for Vasicek-Song test
Error in vs.estimate(x, densfun, ESTIM, extend, delta, relax):  Too many ties
to compute Vasicek estimate.
```

```
vs.test(x = samp, densfun = "dexp", extend = TRUE)

Warning in vs.estimate(x, densfun, ESTIM, extend, delta, relax):  Ties should
not be present for Vasicek-Song test


Vasicek-Song GOF test for the exponential distribution

data:  samp
Test statistic = 0.025702, Optimal window = 16, p-value = 0.9052
sample estimates:
    Rate
1.683785
```

Finally, Vasicek's estimate V_{mn} may exceed the parametric estimate of the entropy of the null distribution for all m between 1 and $n^{1/3-\delta}$. Then, no window size exists satisfying (15), as illustrated below.

```
set.seed(84)
ech <- rpareto(20, mu = 1/2, c = 1)
vs.test(x = ech, densfun = 'dpareto', param = c(1/2, 1))

Error in vs.estimate(x, densfun, ESTIM, extend, delta, relax):  The sample
entropy is greater than empirical maximal entropy for all possible window
sizes; the sample may be too small or is unlikely to be drawn from the null
distribution.
```

Enlarging the possible window sizes by setting `extend` to `TRUE` may enable Vasicek estimates to be smaller than empirical entropy.

Note that when computing the p-value by Monte-Carlo simulation, the constraint (14) may not be satisfied for some replicates, whatever be the window size. These replicates are then ignored and the p-value is computed from the remaining replicates. A warning message is added to the output, informing on the number of ignored replicates.

```
data(contaminants) #load data from package vsgoftest; see ?contaminants
set.seed(1)
vs.test(x = aluminium2, densfun = 'dpareto')

Warning in vs.test(x = aluminium2, densfun = "dpareto"):  For 176 simulations
(over 5000 ), entropy estimate is greater than empirical maximum entropy for
all window sizes.


Vasicek-Song GOF test for the Pareto distribution

data:  aluminium2
Test statistic = 1.3676, Optimal window = 2, p-value < 2.2e-16
sample estimates:
         mu           c
  0.3288148 360.0000000
```

A large proportion of such ignored replicates may indicate that the original sample is too small or the null distribution does not fit it.

The function `vs.test` also allows to avoid the constraint (14) when computing the optimal window size, by setting the optional argument `relax` to `TRUE`. This however should be used with special care, even when the p-value is computed by Monte-carlo simulation, because it may lead to spurious conclusions. Some examples will be discussed in Section 5. This option is to recover the non-parametric likelihood ratio GOF test developed by Vexler and Gurevich (2010) and performed by **dbEmpLikeGOF**; see Section 4.2.

3.3. Technical information on the internal structure of the vsgoftest package

While `entropy.estimate` is a stand-alone function – depending only on the **base** and **stats** packages, `vs.test` is supported by a set of internal functions – not available for users; the structure of the package and connections between functions are described in the organisational chart presented in Figure 1. Functions available for users are depicted by rectangles while internal functions are depicted by ellipses. An arrow connecting a function to another means that the first function (say master function) calls the second (slave) during execution. When such a call is optional (depending on arguments given in the master function), the arrow is dashed and annotated with the corresponding argument settings. The function `fitdist` depicted by a dashed rectangle is a function implemented in the **fitdistrplus** package Delignette-Muller and Dutang (2015). The double-lined ellipse depicts a C++ encoded function that has been integrated via the package **Rcpp** Eddelbuettel and Francois (2011).

The **vsgoftest** package is structured in such a way so as to:

- Allow easy access to the code source. Especially, the master function `vs.test` calls four slave functions corresponding to the following tasks (enumerated according to the organisational chart of Figure 1):

  1. computing the MLE of the parameter θ of the null distribution through the function

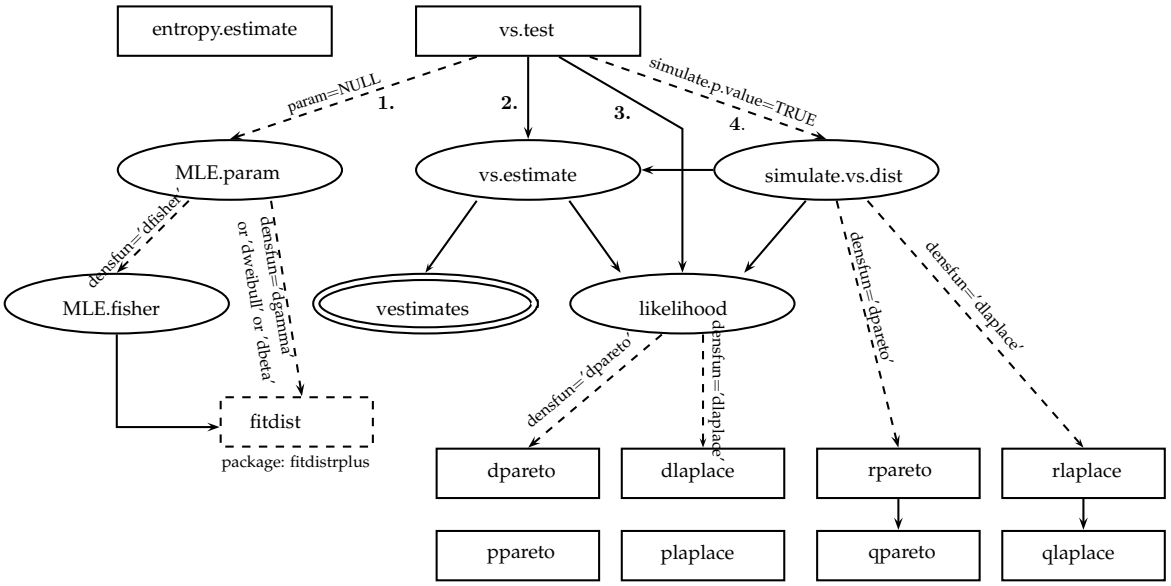

Figure 1: Organisational chart of the structure of the package and connections between functions available for users (depicted by rectangles) and internal functions (depicted by ellipses).

`MLE.param`.

2. Computing Vasicek estimate $V_{\widehat{m}n}$ of Shannon entropy for the sample with the optimal window $\widehat{m}$ given by (15).
3. Computing the VS test statistic $I_{\widehat{m}n}$.
4. Computing the p-value associated to the sample. If the sample size is either greater than 80 or the optional argument `simulate.p.value` is `TRUE`, then the p-value is estimated by means of Monte-Carlo simulation performed by the internal function `simulate.vs.dist`.

- Limit dependence to other packages. In this aim, density, cumulative density and quantile functions as well as random generators for Pareto and Laplace distributions have been encoded, even if they are available in other R packages such as **VGAM** in Yee (2010), **POT** in Ribatet and Dutang (2016) and **smoothmest** in Hennig (2012). The MLE $\widehat{\theta}_n$ of the parameter of the null distribution is computed thanks to the function `fitdist` of the **fitdistrplus** package only if no closed form expression is known for it, *i.e.*, for Gamma, Weibull, Beta and Fisher distributions. Otherwise, the closed form expression is used.

- Optimize time and resources, especially for Monte-Carlo simulation. To this end, the most time-consuming part of the procedure – namely, the computation of Vasicek estimate for all possible window sizes, has been converted to C++ and integrated to the package via **Rcpp**, in the internal function `vestimates`.

4. Performance of Vasicek-Song tests

First, a review of power studies of VS tests available in literature is presented in Section 4.1. Then, power comparisons of VS tests and classical GOF tests are proposed when applied to discriminate between close distributions, such as Pareto versus shifted log-normal and Exponential versus Weibull. Finally, the features of packages **vsgoftest** and **dbEmpLikeGOF** are compared in Section 4.2; the methodological differences are highlighted, the higher performance of **vsgoftest** both in terms of power and computational time is pointed out and illustrated.

4.1. Power computation

Comparisons of the power properties of VS tests are widely discussed in the literature. Various choices of null and alternative distribution families are considered. VS tests are shown to generally outperform classical GOF tests. A comprehensive list of these references is given in this section, with main conclusions summarized in Table 2. Especially, power properties of the VS test for normality have been discussed by Vasicek (1976), Arizono and Ohta (1989) and Gurevich and Davidson (2008) among many others. Compared with many tests, including Kolmogorov-Smirnov (KS), Cramér-von Mises (CvM), Anderson-Darling (AD) and Shapiro-Wilk (SW), the VS test exhibits higher power for most of alternative distributions. When the null distribution is an exponential distribution, the VS test is also shown in Ebrahimi *et al.* (1992) to be more powerful than the Van-Soest and Finkelstein and Schafer tests, which are modified versions of respectively CvM and KS tests, for various alternative distributions such as Weibull, gamma and log-normal. Choi and Kim (2006) for Laplace and Lequesne (2015b) for Pareto show that the VS test is more powerful than EDF tests, for various alternative distributions. The uniform VS GOF test is shown to outperform many other tests for alternative distributions having most of their mass near 0.5, but remains less powerful than CvM and Watson tests for other alternative distributions.

On the basis of power computation in the literature, we choose to compare the power of the VS test to the KS, CvM and AD tests, for close null and alternative distributions. In particular, difficulties in distinguishing a Pareto tail from that of a log-normal is an issue; see for example Malevergne, Pisarenko, and Sornette (2011). For illustration, we estimate through Monte-Carlo simulation the power of VS, KS, CvM and AD of Pareto distributions applied to samples drawn from a (shifted) log-normal distribution. We simulate 10000 replicates x_1^n of a random sample X_1^n drawn from a shifted log-normal distribution $\mathcal{LN}(0, \sigma)$ with support $[1, \infty[$ and $\sigma = 1, 1.25$, for $n \in \{20, 30, 50, 100\}$. Then, we apply the tests for the simple null hypothesis $H_0 : P = \mathcal{P}ar(1, \mu)$, for $\mu = 1$ when $\sigma = 1$ and $\mu = 0.8$ when $\sigma = 1.25$; the power is estimated by the proportion of rejections of the null hypothesis among the 10000 replicates. The following code chunk illustrates the procedure, for $\sigma = \mu = 1$ and $n = 20$, using the VS test. This procedure immediately adapts to other values of σ, μ and n and to other tests[1]. Results are presented in Table 3 (top).

```
N <- 10000
n <- 20
mu <- 1
set.seed(54)
```

[1]The seed of the pseudo-random number generator has been changed for each couple of μ and n; the whole procedure yielding Table 3 is available in the file *vsgoftest_performances.R*, in the directory *inst/doc* of the package source file.

Reference	Null distrib.	Alt. distrib.	VS compared with	Most powerful
Vasicek (1976)	Normal	exponential, gamma, uniform, beta, Cauchy	KS, CvM, Kuiper, Watson, AD, SW	AD (for Cauchy), VS (for others)
Arizono and Ohta (1989)		log-normal, uniform, χ^2, student	KS, CvM, χ^2	VS
Gurevich and Davidson (2008)		log-normal, χ^2, Student, uniform, exponential, gamma, beta, Cauchy	KS	KS (for Student and Cauchy), VS (for others)
Ebrahimi *et al.* (1992)	Exponential	Weibull, gamma, log-normal	Van-Soest, Finkelstein and Schafer	VS
Dudewicz and Van Der Meulen (1981)	Uniform	Distributions defined on $[0, 1]$	KS, CvM, Kuiper, Watson, AD, log-statistic, χ^2	VS (for alternative having most of its mass near 0.5), CvM or Watson (for others)
Choi and Kim (2006)	Laplace	Normal, Student, logistic, Cauchy, uniform, chi-squared, Weibull, log-normal, extreme value and inverse Gaussian	KS, CvM, AD, Kuiper and Watson	VS
Lequesne (2015b)	Pareto	Weibull, gamma, log-normal, two-parameter exponential	KS, AD	VS
Mudholkar and Tian (2002)	Inv. Gaussian	exponential, uniform, Weibull and log-normal	KS	VS (uniform and Weibull)
Alizadeh Noughabi, Alizadeh Noughabi, and Ebrahimi Moghaddam Behabadi (2014)	Rayleigh	Weibull, gamma, log-normal, half-normal, uniform, modified extreme value, linear increasing failure rate law, Dhillon's law and Chen's distribution	KS, CvM, AD, Kuiper and Watson	VS (for uniform) and AD (for other alternatives)
Perez-Rodriguez, Vaquera-Huerta, and Villaseñor-Alva (2009)	Gumbel	Weibull, log-normal, normal, logistic, Cauchy, Student, gamma and Fréchet	KS, CvM, AD, Kuiper and Kinnison	AD (for heavy tails), VS (for others)
Tsujitani, Ohta, and Kase (1980)	Extreme-Value	Normal and 3-parameter log-Weibull	KS, CvM, AD, Kuiper, Mann et modified Mann test	VS
Lund and Rao Jammalamadaka (2000)	von Mises	Mixtures of Von Mises distributions (bimodal, skewed, long-tailed and half), the cardoid and triangular distributions	Watson and integrated squared error test	Watson (for long-tailed), VS (for half) and fairly equal for other alternatives

Table 2: Power studies of VS tests performed in the literature.

```
res.pow <- replicate(n = N,
                     expr = vs.test(x = 1 + rlnorm(n,
                                                   meanlog = 0,
                                                   sdlog = 1),
                                    densfun = 'dpareto',
                                    param = c(1,1),
```

	VS	KS	AD	CvM	VS	KS	AD	CvM
	$H_0 : \mathcal{P}(1,1); H_1 : 1 + \mathcal{LN}(0,1)$				$H_0 : \mathcal{P}(1,0.8); H_1 : 1 + \mathcal{LN}(0,1.25)$			
n=20	59.79	8.93	6.72	7.02	40.62	16.36	13.44	16.21
n=30	77.66	15.79	22.83	16.61	55.50	27.05	26.54	26.98
n=50	94.02	37.39	68.02	46.86	76.83	50.70	58.87	52.36
n=100	99.99	85.90	99.83	96.36	98.42	89.22	97.58	92.07
	$H_0 : \mathcal{E}(1/2); H_1 : \mathcal{W}(1.2,2)$				$H_0 : \mathcal{E}(1/2); H_1 : \mathcal{W}(1.3,2)$			
n=20	9.97	5.06	3.65	4.63	14.67	5.27	3.28	4.41
n=30	12.05	6.10	4.29	5.19	19.93	7.40	5.65	6.45
n=50	13.47	7.37	6.50	6.80	25.86	11.28	11.30	10.53
n=100	25.23	11.35	14.04	11.42	67.14	21.91	34.67	24.60

Table 3: Power (expressed as percentage of true rejection) of VS, KS, CvM and AD tests for testing Pareto and exponential distributions against shifted log-normal (top) and Weibull distributions (bottom).

```
                                    simulate.p.value = TRUE,
                                    B = 1000)$p.value)
```

The power of VS, KS, CvM and AD tests is similarly computed for null exponential and alternative Weibull distributions. The Weibull distribution $\mathcal{W}(a,b)$ reduces to an exponential distribution $\mathcal{E}(1/a)$ when $b = 1$. The main aim is thus to determine which test better discriminates between these distributions when the shape parameter of the Weibull distribution is close to 1, precisely $b = 1.2$ and $b = 1.3$. Results are given in Table 3 (bottom), clearly showing that the VS test outperforms EDF tests.

Note that the above procedure for comparing the power of GOF tests adapts easily to other sets of null and alternative distributions.

4.2. vsgoftest versus dbEmpLikeGOF for testing uniformity and normality

The package **dbEmpLikeGOF** in Miecznikowski *et al.* (2013) performs uniformity and normality tests based on empirical likelihood ratios (ELR) – say ELR tests. These tests are strongly linked to VS tests. Precisely, for testing the normality of a sample $X_1, \dots X_n$, the ELR test statistic is $\log V_n$, where

$$V_n = \min_{1 \leq m < n^{1/2}} \left[(2\pi e \check{S}^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n \left[X_{(i+m)} - X_{(i-m)} \right]} \right] \quad \text{and} \quad \check{S}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \frac{1}{n} \sum_{j=1}^{n} X_j \right)^2 .$$

Mere algebra yields

$$\log V_n = n I_{\widetilde{m}n} + \frac{1}{2}, \tag{17}$$

with $\widetilde{m} \in \operatorname{argmax}_{1 \leq m < n^{1/2}} V_{mn}$. Hence, ELR and VS tests differ only in the window size choice: the upper bound is $n^{1/2}$ for the ELR test while it is (by default) $n^{1/4}$ for the VS test

	VS	ELR	VS	ELR
	$H_1 : \mathcal{L}(0,1)$		$H_1 : \mathcal{S}(4)$	
$n = 50$	17.8	16.0	16.1	14.6
$n = 200$	86.2	64.9	71.2	35.8

Table 4: Power comparisons between VS and ELR normality tests for samples drawn from a Laplace distribution (left) and a Student distribution (right). Power is estimated by means of Monte-Carlo simulation based on 1000 replicates of the samples.

and the constraint (14) is not taken into account by the EL test. Enlarging the upper bound from $n^{1/4}$ to $n^{1/2}$ may lead to a more powerful decision rule, as mentioned and illustrated in Section 3.2. Still, practically, the normality VS test tends to outperform the ELR test when applied to heavy tailed samples, as illustrated by Table 4[2]. Moreover, the upper bound $n^{1/4}$ legitimates the use of the asymptotic distribution of $I_{\widehat{m}n}$ in `vs.test`, which is not performed by `dbEmpLikeGOF`. As previously mentioned in Section 3.2, disabling the constraint (14) may lead to spurious conclusions.

The ELR test can be performed using `vs.test`, by suitably setting its arguments, as illustrated by the following code chunk [3].

```
set.seed(1)
samp <- rnorm(50)
res.vs <- vs.test(x = samp, densfun = 'dnorm', delta = -1/6, relax = TRUE)
res.vs$statistic*50 +1/2

Test statistic
      7.970748

library(dbEmpLikeGOF)
res.el <- dbEmpLikeGOF(x = samp, testcall = 'normal', vrb = FALSE)
res.el$teststat

[1] 7.975815
```

Additionnaly, from a computational view point, some significant differences exist between the functions `dbEmpLikeGOF` and `vs.test`. Precisely, the p-value returned by `dbEmpLikeGOF` is computed by default by linear interpolation from a table of pre-computed p-values for various sample sizes (from $n = 10$ to 10000) and test statistic values; the p-value can be approximated by Monte-Carlo simulation by setting `pvl.Table = FALSE`. By default, `vs.test` computes the p-value by means of Monte-Carlo simulation or uses the asymptotic distribution (12),

[2]The comparison procedure is available in the file *vsgoftest_performances.R*, in the directory *inst/doc* of the package source file.

[3]Some slight difference remains between the two computed values, due to numerical inaccuracy in computation procedures: the estimated entropy of the null distribution is computed from the closed form expression (4) in `dbEmpLikeGOF` while it is computed as the empirical mean of the log-likelihood of the sample in `vs.test`.

depending on the sample size. In both cases, `vs.test` is approximately five times faster than `dbEmpLikeGOF`, as illustrated by Figure 2[4].

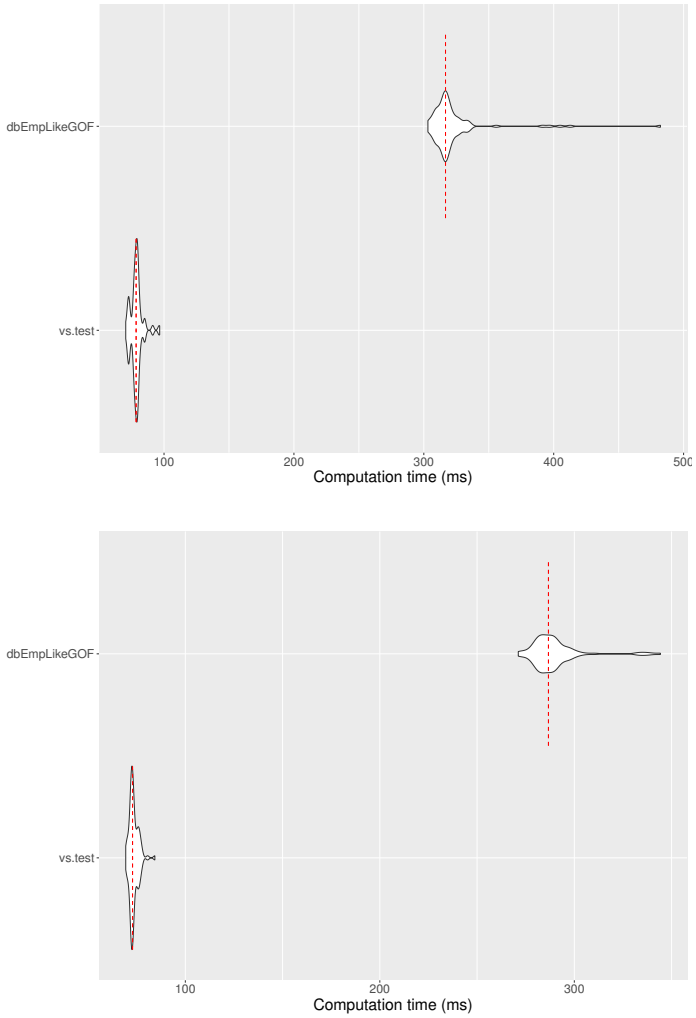

Figure 2: Distributions (violin plots) of computation time of 100 iterations of the functions dbEmpLikeGOF (from package **dbEmpLikeGOF**) and vs.test for testing normality of a sample of size 50 drawn from the standard normal distribution (top) and for testing uniformity of a sample of size 50 drawn from the uniform distribution on $[0, 1]$ (bottom). The p-values are estimated by Monte-Carlo simulation, based on 1000 replicates. The red dashed lines are the median computation times.

5. Application to real data

The **vs.test** package contains environmental data originating from a guidance report edited

[4]Simulations have been performed on a Dell Lattitude E5580 laptop, equipped with an Intel® Core™ i7-7600U CPU at 2.80GHz x 4, with 16GB RAM. R code for generating these figures is available in the file *vsgoftest_performances.R*, in the directory *doc* of the package source file.

by the Technology Support Center of the United States Environmental Protection Agency; see Singh, Singh, and Engelhardt (1997). According to Singh *et al.* (1997), environmental scientists take remediation decisions at suspected sites based on organic and inorganic contaminant concentration measurements. These decisions usually derive from the computation of confidence upper bounds for contaminant concentrations. Testing the goodness-of-fit of specified models hence appears of prior interest. Singh *et al.* (1997) also points out that contaminant concentration data from sites often appear to follow a skewed probability distribution, making the log-normal family a frequently-used model. The authors illustrate their purpose by applying Shapiro-Wilk test to the log-transformed of the samples `aluminium1`, `manganese`, `aluminium2` and `toluene` (stored in the present package)[5]; see the empirical skewness computed in the following chunk.

```
data(contaminants) #Load environmental data from package
#Package DescTools required for this chunk
unlist(lapply(X = list(aluminium1, manganese, aluminium2, toluene),
       FUN = DescTools::Skew))

[1] 2.323343 1.698686 1.996607 3.961129
```

The following code chunks intend to illustrate the use and behavior of the function `vs.test` for these environmental data. The significant level is fixed to 0.1 as in Singh *et al.* (1997). Note that warning messages notifying that there are ties in the samples have been dropped out from outputs.

```
set.seed(1)
vs.test(x = aluminium1, densfun = 'dlnorm')


Vasicek-Song GOF test for the log-normal distribution

data:  aluminium1
Test statistic = 0.31232, Optimal window = 2, p-value = 0.3372
sample estimates:
Location    Scale
6.225681 1.609719
```

The log-normal hypothesis is not rejected for `aluminium1`. Similar results are obtained for `manganese`. Log-normality is rejected for `aluminium2`.

```
set.seed(1)
vs.test(x = aluminium2, densfun = 'dlnorm')


Vasicek-Song GOF test for the log-normal distribution
```

[5] A succinct description of these data is available by executing the following R command: `?contaminants`

```
data:  aluminium2
Test statistic = 0.48369, Optimal window = 2, p-value = 0.0256
sample estimates:
 Location     Scale
8.9273293 0.8264409
```

Due to numerous ties in `toluene`, `vs.test` can not compute Vasicek entropy estimate unless `extend` is set to `TRUE`. Still, `vs.test` notifies that the constraint (14) is violated for all window sizes, which suggests that data are not likely to be drawn from the log-normal distribution; see Section 2. Turning `relax` to `TRUE` yields the following result.

```
set.seed(1)
vs.test(x = toluene, densfun = 'dlnorm', extend = TRUE, relax = TRUE)


Vasicek-Song GOF test for the log-normal distribution

data:  toluene
Test statistic = -2.4984, Optimal window = 11, p-value = 0.7308
sample estimates:
Location    Scale
4.651002 3.579041
```

Again, this last result looks spurious because the test statistic is negative – resulting from (14) not being satisfied by setting `relax = TRUE`. An alternative is to test normality of the log-transformed sample as follows.

```
set.seed(1)
vs.test(x = log(toluene), densfun ='dnorm', extend = TRUE)


Vasicek-Song GOF test for the normal distribution

data:  log(toluene)
Test statistic = 0.6536, Optimal window = 11, p-value = 2e-04
sample estimates:
    Mean St. dev.
4.651002 3.579041
```

The log-normal hypothesis is not rejected for `aluminium1` and `manganese` while it is rejected for `aluminium2` and `toluene`. These results are consistent with those obtained by Singh *et al.* (1997). Further, the goodness-of-fit to the Pareto distributions is performed for `aluminium2` and `toluene`. Log-normal and Pareto distributions usually compete with closely related generating processes and hard to distinguish tail properties; see for example Malevergne *et al.* (2011). Goodness-of-fit of Pareto distribution is rejected for `aluminium2`.

```
set.seed(1)
vs.test(x = aluminium2, densfun = 'dpareto')


Vasicek-Song GOF test for the Pareto distribution

data:  aluminium2
Test statistic = 1.3676, Optimal window = 2, p-value < 2.2e-16
sample estimates:
         mu           c
  0.3288148 360.0000000
```

Applying `vs.test` to `toluene` with default settings yields no result because of numerous ties and the violation of (14). Uniformity of the sample transformed by the cumulative density function of the Pareto distribution can be tested as follows. Goodness-of-fit of the Pareto distribution is not rejected for `toluene`.

```
#Compute the MLE of parameters of Pareto dist.
res.test <- vs.test(x = toluene,
                    densfun = 'dpareto',
                    extend = TRUE, relax = TRUE)
#Test uniformity of transformed data
set.seed(5)
vs.test(x = ppareto(toluene,
                    mu = res.test$estimate[1],
                    c = res.test$estimate[2]),
        densfun ='dunif', param = c(0,1), extend = TRUE)


Vasicek-Song GOF test for the uniform distribution with Min=0,
Max=1

data:  ppareto(toluene, mu = res.test$estimate[1], c = res.test$estimate[2])
Test statistic = 0.25383, Optimal window = 10, p-value = 0.2496
```

Conclusion

Vasicek-Song tests constitute powerful GOF tests for classical parametric families of distributions, relying on an information theoretical framework. They can be easily performed by using the **vsgoftest** package for R. Default and optional settings of the functions provided by the package make the procedure both intuitive and flexible. Its application to real datasets manages to illustrate its practical usage.

The package allows for testing GOF of a significant list of parametric models; this list could be extended in further releases. New entropy-based GOF tests could also be considered by

using Rényi entropy and divergence – see Lequesne (2015a), thus extending even more the class of possible distributions, e.g., Student distributions.

References

Alizadeh Noughabi R, Alizadeh Noughabi H, Ebrahimi Moghaddam Behabadi A (2014). "An entropy test for the Rayleigh distribution and power comparison." *Journal of Statistical Computation and Simulation*, **84**(1), 151–158.

Arizono I, Ohta H (1989). "A test for normality based on Kullback–Leibler information." *The American Statistician*, **43**(1), 20–22.

Choi B, Kim K (2006). "Testing goodness-of-fit for laplace distribution based on maximum entropy." *Statistics*, **40**(6), 517–531.

Csiszár I (1975). "I-divergence geometry of probability distributions and minimization problems." *The Annals of Probability*, pp. 146–158.

Delignette-Muller ML, Dutang C (2015). "fitdistrplus: An R Package for Fitting Distributions." *Journal of Statistical Software*, **64**(4), 1–34. URL http://www.jstatsoft.org/v64/i04/.

Dudewicz EJ, Van Der Meulen EC (1981). "Entropy-based tests of uniformity." *Journal of the American Statistical Association*, **76**(376), 967–974.

Ebrahimi N, Habibullah M, Soofi ES (1992). "Testing exponentiality based on Kullback-Leibler information." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 739–748.

Eddelbuettel D, Francois R (2011). "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software, Articles*, **40**(8), 1–18. ISSN 1548-7660. doi:10.18637/jss.v040.i08. URL https://www.jstatsoft.org/v040/i08.

Faraway J, Marsaglia G, Marsaglia J, Baddeley A (2015). *goftest: Classical Goodness-of-Fit Tests for Univariate Distributions*. R package version 1.0-3, URL https://CRAN.R-project.org/package=goftest.

Girardin V, Lequesne J (2017). "Entropy-based goodness-of-fit tests – a unifying framework. Application to DNA replication." *Communications in Statistics-Theory and Methods*, pp. 1–13. doi:10.1080/03610926.2017.1401084. URL https://doi.org/10.1080/03610926.2017.1401084.

Gonzalez-Estrada E, Villasenor-Alva JA (2016). *goft: Tests of Fit for some Probability Distributions*. R package version 1.3.1, URL https://CRAN.R-project.org/package=goft.

Gurevich G, Davidson A (2008). "Standardized forms of Kullback-Leibler information based statistics for normality and exponentiality." *Computer Modelling and New Technologies*, **12**(1), 14–25.

Hennig C (2012). *smoothmest: Smoothed M-estimators for 1-dimensional location*. R package version 0.1-2, URL https://CRAN.R-project.org/package=smoothmest.

Jaynes ET (1957). "Information theory and statistical mechanics." *Physical review*, **106**(4), 620.

Krit M (2015). *EWGoF: Goodness-of-Fit Tests for the Exponential and Two-Parameter Weibull Distributions*. R package version 2.1, URL https://CRAN.R-project.org/package=EWGoF.

Kullback S, Leibler RA (1951). "On information and sufficiency." *The annals of mathematical statistics*, **22**(1), 79–86.

Lequesne J (2013). "Entropy-based goodness-of-fit test: Application to the Pareto distribution." In *AIP Conference Proceedings*, volume 1553, pp. 155–162. AIP.

Lequesne J (2015a). "A goodness-of-fit test of Student distributions based on Rényi entropy." In *AIP Conference Proceedings*, volume 1641, pp. 487–494. AIP.

Lequesne J (2015b). *Tests statistiques basés sur la théorie de l'information, applications en biologie et en démographie*. Ph.D. thesis, Université de Caen Normandie, France.

Lund U, Rao Jammalamadaka S (2000). "An entropy-based test for goodness of fit of the von Mises distribution." *Journal of statistical computation and simulation*, **67**(4), 319–332.

Malevergne Y, Pisarenko V, Sornette D (2011). "Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities." *Physical Review E*, **83**(3), 036111.

Mergel V (1999). "Test of goodness-of-fit for the inverse-gaussian distribution." *Mathematical Communications*, **4**(2), 191–195.

Miecznikowski JC, Vexler A, Shepherd L (2013). "dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two-Sample Comparisons Based on Sample Entropy." *Journal of Statistical Software*, **54**(3), 1–19. URL http://www.jstatsoft.org/v54/i03/.

Mudholkar GS, Tian L (2002). "An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test." *Journal of statistical planning and inference*, **102**(2), 211–221.

Novack-Gottshall P, Wang SC (2016). *KScorrect: Lilliefors-Corrected Kolmogorov-Smirnoff Goodness-of-Fit Tests*. R package version 1.2.0, URL https://CRAN.R-project.org/package=KScorrect.

Perez-Rodriguez P, Vaquera-Huerta H, Villaseñor-Alva JA (2009). "A goodness-of-fit test for the gumbel distribution based on Kullback–Leibler information." *Communications in StatisticsTheory and Methods*, **38**(6), 842–855.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ribatet M, Dutang C (2016). *POT: Generalized Pareto Distribution and Peaks Over Threshold*. R package version 1.1-6, URL https://CRAN.R-project.org/package=POT.

Shannon CE (1948). "A mathematical theory of communication." *Bell System Technical Journal*, **27**(3), 379–423.

Singh AK, Singh A, Engelhardt M (1997). "The lognormal distribution in environmental applications." In *Technology Support Center Issue Paper*. Citeseer.

Song KS (2002). "Goodness-of-fit tests based on Kullback-Leibler discrimination information." *IEEE Transactions on Information Theory*, **48**(5), 1103–1117.

Stephens MA (1974). "EDF statistics for goodness of fit and some comparisons." *Journal of the American statistical Association*, **69**(347), 730–737.

Tsujitani M, Ohta H, Kase S (1980). "Goodness-of-fit test for extreme-value distribution." *IEEE Transactions on Reliability*, **29**(2), 151–153.

Vasicek O (1976). "A test for normality based on sample entropy." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 54–59.

Vexler A, Gurevich G (2010). "Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy." *Computational Statistics & Data Analysis*, **54**(2), 531–545.

Yee TW (2010). "The VGAM Package for Categorical Data Analysis." *Journal of Statistical Software*, **32**(10), 1–34. URL http://www.jstatsoft.org/v32/i10/.

Affiliation:

Justine Lequesne, Centre Henri Becquerel, Unité de Recherche Clinique, Rue d'Amiens, CS 11516, 76038 Rouen cedex 1, France,
E-mail: justine.lequesne@chb.unicancer.fr

Philippe Regnault, Laboratoire de Mathématiques de Reims, FRE 2011, Université de Reims Champagne-Ardenne, Campus Moulin de la Housse, BP 1039, 51687 Reims cedex 2, France.
E-mail: philippe.regnault@univ-reims.fr