Generalised Smooth Tests of Goodness of Fit Utilising L-Moments

O. Thas*[(1,2)], J.C.W. Rayner[(2,3)] and J. De Neve[1]

[1] Department of Mathematical Modelling, Statistics and Bioinformatics,
Ghent University, 9000 Gent, Belgium

[2] National Institute for Applied Statistics Research Australia (NIASRA),
School of Mathematics and Applied Statistics, University of Wollongong,
New South Wales 2522, Australia

[3] School of Mathematical and Physical Sciences, University of Newcastle,
New South Wales 2308, Australia

Olivier.Thas@UGent.be John.Rayner@newcastle.edu.au JanR.Deneve@UGent.be

* corresponding author. Full postal address: Department of Mathematical Modelling,
Statistics and Bioinformatics, Ghent University, Coupure Links 653,
B-9000 Gent, Belgium. Tel: +32-9-264 59 33

Abstract

We present semiparametric tests of goodness of fit which are based on the method of L-moments for the estimation of the nuisance parameters. This test is particularly useful for any distribution that has a convenient expression of its quantile function. The null hypothesis states equality of the first few L-moments of the true and the hypothesised distributions. We provide details and simulation studies for the logistic and the generalised Pareto distributions. Whereas for some distributions the method of L-moments estimator may be less efficient than the maximum likelihood estimator, the former has the advantage that it may be used in semiparametric settings and that it requires weaker existence conditions than a maximum likelihood estimator. The new tests often outperform competitor tests for the logistic and generalised Pareto distributions.

Keywords

generalised Pareto distribution; logistic distribution; order statistics; quantile function

1 INTRODUCTION

The initial motivation for the research presented in this paper is testing goodness of fit for the generalised Pareto distribution (GPD). Although many statistical hypothesis tests have been proposed for the GPD (Choulakian & Stephens, 2001; Radouane & Crétois, 2002; De Boeck et al., 2011), most of them cannot be applied in all circumstances. The reason is that the GPD parameter estimators do not always behave properly. For example, Hosking & Wallis (1987) found convergence problems with maximum likelihood estimation and showed that in certain parameter ranges the method of moments had small efficiency and large bias. Zhang (2007) proposed a likelihood moment estimation method, which is computationally easy and has high asymptotic efficiency. Hosking & Wallis (1987) preferred a method based on probability-weighted moments (Greenwood et al., 1979). Castillo & Hadi (1997) preferred their elemental percentile method, and they incidentally noted that the method of moments and probability-weighted moments both may result in sample observations falling outside the support of the fitted distribution, which is known as the "feasibility problem". See, for example, Chen & Balakrishnan (1995) for a discussion. Part of the problems related to the GPD is caused by the non-existence of some of the moments.

The non-existence of certain moments also affected the smooth tests and generalised smooth tests of goodness of fit for the GPD that were proposed by De Boeck et al. (2011). Smooth tests were first proposed by Neyman (1937), but a modern and comprehensive overview is given by Rayner et al. (2009). The general construction of a smooth test starts with embedding the probability density function of the hypothesised distribution in a larger family of distributions (*smooth family* *of alternatives*). This involves the introduction of extra parameters (*embedding parameters*), say θ, associated with user-defined score functions that should span the space of important alternatives. The term "smooth" comes from the fact that $\theta = 0$ corresponds to the probability density function of the hypothesised distribution, and moving θ away from 0 makes the alternative deviate smoothly from the hypothesised distribution. A smooth test is basically a score test for testing $\theta = 0$ in the smooth family of alternatives. An advantage of smooth tests that make use of polynomial score-functions is that at the rejection of the null hypothesis, the components of the test statistic may assist in diagnosing the type of deviation from the hypothesised distribution in terms of moment deviations. Rayner et al. (2009) provide an overview of smooth and generalised smooth tests, utilising both maximum likelihood estimators (MLE) and method of moments estimators (MME) of the parameters. However, when applied to the GPD both MLE and MME-based smooth tests may suffer from existence problems.

In this paper, another type of goodness of fit test is proposed. The test shares many properties with the traditional smooth tests. For example, the test statistic may be decomposed into components that are related to moment deviations, and the test may be related to a family of alternatives that varies smoothly from the hypothesised distribution. A contribution of the proposed test is that L-moments are used instead of the classical moments, and the smooth alternative is formulated in terms of the quantile function instead of the probability density function.

L-moments were first unified by Hosking (1990). There are two important properties of L-moments that make them suitable for goodness of fit testing. First, all L-moments of a random variable X exist if and only if the mean of X exists and is finite. Second, a distribution whose mean exists is uniquely characterised by its L-moments (Hosking, 1990, 2007). The new test therefore is referred to as an L-moments generalised smooth test (LGST). The LGST method is particularly convenient for distributions that have a simple definition in terms of their quantile function. Our method is thus not only applicable to the GPD.

L-moments have been used before in goodness of fit procedures and particularly with applications in hydrology and meteorology. Examples include Chowdhurry et al. (1991) for the generalised extreme value distribution and Harri & Coble (2011) for the normal distribution. In the discussion section (Section 6), we will relate our methods to existing methods.

Section 2 is a brief introduction to L-moments and the method of L-moments for parameter estimation. In Section 3 the LGST test is described in full generality. Section 4 reviews a simulation study in which the powers of the LGST are compared to powers of competitor tests for two selected parent distributions: the GPD and the logistic distribution. The test is also demonstrated on an example data set in Section 5. Conclusions are provided in Section 6.

2 L-MOMENTS

2.1 L-Moments and their Sample Estimators

If F denotes the cumulative distribution function of the random variable X with support $\mathcal{S} \subseteq \mathbb{R}$, then the quantile function of X is defined as $Q(u) = \inf\{x \in \mathcal{S} : F(x) \geq u\}$. The quantile function can be notationally simplified to $Q(u) =$

$F^{-1}(u)$ when F is continuous.

The theory of L-moments was unified by Hosking (1990). L-moment theory is parallel with the theory of conventional moments and in particular the sample L-moments, being linear functions of the sample observations, are more robust than their conventional sample moment counterparts. Hosking (1990) also demonstrated that probability weighted moments (Greenwood et al., 1979) can be expressed as linear combinations of L-moments. Consequently, several results for L-moments can be traced back to the probability weighted moments literature.

We limit the discussion here to continuous distributions. Let $X_1, \ldots, X_n$ denote n sample observations that are i.i.d. with cumulative distribution function F, and let $X_{i:n}$ denote the ith order statistic of the sample. The rth L-moment is defined as

$$\lambda_r = \frac{1}{r}\sum_{s=0}^{r-1}(-1)^s\binom{r-1}{s}\mathrm{E}\left(X_{r-s:r}\right) = (2r-1)^{-1/2}\int_0^1 Q(u)h_{r-1}(u)\mathrm{d}u, \quad (1)$$

in which

$$h_r(u) = (2r+1)^{1/2}\sum_{s=0}^{r}(-1)^{r-s}\binom{r}{s}\binom{r+s}{s}u^s$$

represents the rth order normalised shifted Legendre polynomial. These polynomials are orthonormal over $[0,1]$. Note that in many papers the defintion of L-moments (1) lacks the factor $(2r-1)^{-1/2} = (2(r-1)+1)^{-1/2}$ because they use the non-normalised polynomials. The reason for using normalised polynomials is that this is a convention in the construction of smooth tests and doing so will simplify expressions in Section 3.

The first four L-moments are given by

$$\lambda_1 = \mathrm{E}(X), \quad \lambda_2 = \frac{1}{2}\mathrm{E}(X_{2:2} - X_{1:2}),$$
$$\lambda_3 = \frac{1}{3}\mathrm{E}(X_{3:3} - 2X_{2:3} + X_{1:3}) \text{ and } \lambda_4 = \frac{1}{4}\mathrm{E}(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}).$$

The first L-moment thus coincides with the mean. The second order L-moment is a measure of scale, as is the classical second moment, and λ_3 and λ_4 are related to skewness and kurtosis, respectively. Note that all moments are linear in the expected order statistics, and consequently all L-moments are expressed in the same units as the original random variable X. For some applications and derivations the unitless L-moment ratios λ_3/λ_2 (L-skew) and λ_4/λ_2 (L-kurtosis) are preferred. The definition of L-moments in terms of expectations of order statistics implies that the L-moments can be estimated as averages of linear combinations of sample order statistics. For example, the sample estimator of λ_1 is the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$, and the sample estimator of λ_2 is $\frac{1}{2}\binom{n}{2}^{-1}\sum_{i>j}(X_{i:n} - X_{j:n})$, which is 0.5 times the Gini mean difference statistic. More generally, the sample estimator of λ_r involves an r-tuple summation. Fortunately, a computationally efficient formulation of these estimators was proposed by Hosking et al. (1985). In particular, the sample estimator of λ_r can also be calculated as $\hat{\lambda}_r = \sum_{k=1}^{r-1} p_{r-1,k} b_k$, where

$$b_k = n^{-1}\sum_{i=1}^{n} \frac{(i-1)(i-2)\cdots(i-k)}{(n-1)(n-2)\cdots(n-k)} X_{i:n}, \quad \text{and}$$
$$p_{r-1,k} = (-1)^{r-k-1}\binom{r-1}{k}\binom{r+k-1}{k}.$$

Theorem 3 in Hosking (1990) gives the asymptotic normality of the sample estimators $\hat{\lambda}_r$ provided that the first two moments are finite. The asymptotic covariance matrix of $n^{1/2}(\hat{\lambda}_1 - \lambda_1, \ldots, \hat{\lambda}_k - \lambda_k)$ is denoted by Λ_k.

2.2 Method of L-Moments Estimation

Suppose $f(x;\beta)$ denotes the probability density function of a distribution that is indexed by the p-dimensional parameter vector β. In goodness of fit testing, when the null hypothesis states that the sample data come from $f(x;\beta)$ without β being specified, β is referred to as a nuisance parameter. Although the MLE of β is usually an efficient estimator, it is not always the most meaningful choice as will be clarified in Section 3. The MME of β may not always be efficient, but in goodness of fit testing the MME may have advantages of interpretability; see, for example, Rayner et al. (2009). The MME of β, say $\tilde{\beta}$, makes the first p sample moments agree with those of $f(x;\tilde{\beta})$. If in the method of moments procedure the conventional moments are replaced by the corresponding L-moments, the method of L-moments is obtained. The method of L-moments estimator (MLME) of β will be denoted by $\hat{\beta}$. Because the MLME may be expressed as a function of the sample L-moments, the sampling distribution of the MLME may be derived from the sampling distributions of the sample L-moments. For many distributions the MLMEs are asymptotically normal; see, for example, Hosking (1986), Hosking et al. (1985), Hosking & Wallis (1987) and Hosking (2009). Despite the MLME generally not being asymptotically optimal, their efficiencies relative to the MLE are often quite large, even in small to moderately large samples (Hosking, 1990). More details are postponed to Section 3. Explicit formulae of MLMEs for the logistic distribution and the GPD are given in Supplementary Material Appendix 1.

3 L-MOMENTS GENERALISED SMOOTH TESTS

3.1 The Null Hypothesis

Let $f(x)$ denote the true density function of X, and suppose a sample of n i.i.d. observations is available. The one-sample goodness of fit null hypothesis is formulated as

$$H_0^{\text{P}} : f \in \mathcal{F}_P = \{f(\cdot;\beta) : \beta \in B \subseteq \mathbb{R}^p\}.$$

This null hypothesis is referred to as the *full parametric* (P) *null hypothesis*, because it fixes all moments of f to the corresponding moments of $f(\cdot;\beta)$. The alternative hypothesis is then usually formulated as $H_1 : f \notin \mathcal{F}_{\text{P}}$, but not all goodness of fit tests are consistent for each fixed alternative in H_1. It has been argued that traditional smooth and generalised smooth tests of fixed order k are basically tests for the *semiparametric* (SP) *null hypothesis* (Henze & Klar, 1996; Henze, 1997; Klar, 2000; Rayner et al., 2009; Thas et al., 2009; Thas, 2010). The SP null hypothesis states that the distribution of the sample observations agrees with the hypothesised distribution in the first k moments. It may be formulated as

$$H_0^{\text{SP}} : f \in \mathcal{F}_{\text{SP}} = \{g : m_r(g) = m_r(f(\cdot;\beta)), \beta \in B \subseteq \mathbb{R}^p, r = 1, \dots, k\}, \quad (2)$$

where $m_r(g)$ and $m_r(f(\cdot;\beta))$ denote the r-th moments of g and $f(\cdot;\beta)$, respectively. The smooth test statistic is basically a quadratic form in statistics that contrast the sample moments with the hypothesised moments, which may in turn depend on the estimated nuisance parameters. In this setting the MLE of β is not defined, because H_0^{SP} does not fully specify $f(\cdot;\beta)$. Moreover, because H_0^{SP}

is formulated in terms of agreement of moments, the MME and the MLME are natural choices, particularly because their use simplifies the interpretation of the test.

3.2 The L-Moment Generalised Smooth Test

We propose smooth tests for testing null hypothesis (2) with the moments m_r replaced by the corresponding L-moments. A motivation is the less restrictive existence conditions of L-moments compared to conventional moments. By doing so, it will be most natural to estimate the nuisance parameter by the method of L-moments.

If the sample estimator $\hat{\lambda}_r$ is denoted as $\lambda_r(X)$, with $X^t = (X_1, \dots, X_n)$ the vector of n i.i.d. sample observations, and if $\lambda_{0r}(\beta) = m_r(f(.;\beta))$ denotes the rth L-moment of the hypothesised distribution $f(.;\beta)$, then we define $\theta_r(X, \beta) = \lambda_r(X) - \lambda_{0r}(\beta)$, $r = 1, 2, \dots$. Thus, for $f \in \mathcal{F}_{SP}$, $\mathrm{E}\left(\theta_r(X, \beta)\right) = 0$ for $r = 1, 2, \dots, k$, and the MLME of β is the solution of $\theta_r(X;\beta) = 0$, $r = 1, \dots, p$.

Our test statistic is based on the statistics $\hat{\theta}_r = \theta_r(X, \hat{\beta}) = \lambda_r(X) - \lambda_{0r}(\hat{\beta})$ ($r = 1, 2, \dots, k > p$). The statistic $\hat{\theta}_r$ measures the deviation between the rth empirical L-moment and the rth L-moment of $f(.;\beta)$ with β replaced by its MLME. A direct consequence of the use of MLME is that the first p statistics $\hat{\theta}_r$ ($r = 1, \dots, p$) are zero due the to the estimation procedure and so they provide no information about testing goodness of fit.

Finally, the following vector notation is used: $\lambda_0^t = (\lambda_{0\beta}^t, \lambda_{0\theta}^t)$ where $\lambda_{0\beta}^t = (\lambda_{01}, \dots, \lambda_{0p})$, $\lambda_{0\theta}^t = (\lambda_{0p+1}, \dots, \lambda_{0k})$, $\hat{\lambda}_\beta^t = (\hat{\lambda}_1, \dots, \hat{\lambda}_p)$ and $\hat{\lambda}_\theta^t = (\hat{\lambda}_{p+1}, \dots, \hat{\lambda}_k)$.

To clarify, the subscript 0 refers to the null hypothesis and the subscripts β and θ refer to the first p orders and the orders $p+1$ up to k, respectively. We define the $p \times p$ matrix $C(\lambda^*)$ as the matrix with element (i,j) equal to $(i,j = 1,\ldots,p)$

$$\frac{\partial}{\partial \lambda_j}\, \beta_i(\lambda)|_{\lambda=\lambda^*}$$

(the existence of these derivatives will be guaranteed by assumptions made in the statements of the lemma and theorem following).

The asymptotic distribution of $\hat{\theta}^t = (\hat{\theta}_{p+1},\ldots,\hat{\theta}_k)$ is given in Theorem 1 following and the proof is given in Appendix A, but first a formal statement of the asymptotic normality of the MLME is given. Lemma 1 is a restatement of results by Hosking (2009, eq. (6.1)).

Lemma 1 *Assume that*

$f \in \mathcal{F}^{SP}$ has finite mean and variance (A1);

for all $\beta \in B$, $\lambda_{0\beta}(\beta)$ is continuously differentiable and its total derivative is invertible (A2).

Then, for all $f \in \mathcal{F}^{SP}$, as $n \to \infty$,

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$$

with $\Sigma = C\Lambda_p C^t$, where Λ_p is the asymptotic covariance matrix of $n^{1/2}\hat{\lambda}_\beta$.

The proof of Lemma 1 is direct upon using the Inverse Function Theorem, the asymptotic normality and consistency of the sample L-moments and the delta method, under assumptions (A1) and (A2).

Theorem 1 *For all $f \in \mathcal{F}_{SP}$, assume that*

f has finite mean and variance (A1);

for all $\beta \in B$, $\lambda_{0\beta}(\beta)$ is continuously differentiable and its total derivative is invertible (A2);

for all $\beta \in B$, $\lambda_{0\theta}(\beta)$ is differentiable (A3);

for all $\beta \in B$, the matrix $M(\beta)\Lambda_k M^t(\beta)$, with $M(\beta)$ as defined in Appendix A, is positive definite (A4).

Then, under H_0^{SP}, when $k > p$, as $n \to \infty$, $n^{1/2}\hat{\theta}$ converges weakly to a zero-mean normal distribution with covariance matrix $\Sigma_{\hat{\theta}} = M(\beta)\Lambda_k M^t(\beta)$, where Λ_k is the asymptotic covariance matrix of $\sqrt{n}(\hat{\lambda}_\beta^t, \hat{\lambda}_\theta^t)$.

The order k L-moments smooth test statistic is of the form $T_k = n\hat{\theta}^t\hat{\Sigma}_{\hat{\theta}}^{-1}\hat{\theta}$, in which $\hat{\Sigma}_{\hat{\theta}}$ is a consistent estimator of $\Sigma_{\hat{\theta}}$ under the semiparametric null hypothesis. The test based on T_k is referred to as the L-moments generalised smooth test (LGST).

Elamir & Seheult (2004) give a simple unbiased estimator of the covariance matrix of the first k sample L-moments, say $n^{-1}\hat{\Lambda}_k$. A consistent estimator of $\Sigma_{\hat{\theta}}$ is obtained by replacing β with its consistent MLME $\hat{\beta}$, or, equivalently, by replacing $\lambda_{0\beta}$ with $\hat{\lambda}_{0\beta}$, and replacing Λ_k with the exact estimator of Elamir & Seheult (2004), resulting in $\hat{\Sigma}_{\hat{\theta}} = M(\hat{\beta})\hat{\Lambda}_k M^t(\hat{\beta})$.

Tests for individual L-moments may be constructed using the test statistics $V_j = \hat{\theta}_j/\hat{\sigma}_j$ ($j = p+1, \ldots, k$), where $\hat{\sigma}_j^2$ is the appropriate diagonal element of the estimated covariance matrix $\hat{\Sigma}_{\hat{\theta}}$. Tests based on V_j are referred to as the component tests in the smooth test literature. The asymptotic null distributions of the test statistics T_k and V_j are direct consequences of Theorem 1. In particular, under the null hypothesis $f \in \mathcal{F}_{SP}$, T_k converges weakly to χ^2_{k-p} ($k > p$), and V_j converges weakly to $N(0, 1)$ ($j = p + 1, \ldots, k$), provided that assumptions (A1) to (A4) of Theorem 1 hold.

3.3 Relationship with Smooth Tests, the Wasserstein Distance and the QQ-Plot

The construction of classical smooth tests starts by defining an order k smooth alternative to the hypothesised density function $f(.; \beta)$. In this section, however, we show that the test is related to an order k alternative that varies smoothly from the hypothesised quantile function, say $Q_0(.; \beta)$. This enables the examination of the relationship of the new test to tests based on the Wasserstein distance and the QQ-plot.

We consider the following kth order alternative to the hypothesised quantile function Q_0,

$$Q_k(u; \beta, \theta) = Q_0(u; \beta) + \sum_{j=1}^{k} \theta_j (2j-1)^{1/2} h_{j-1}(u), \tag{3}$$

where the h_{j-1} are as before. Using (1), it is apparent that the θ parameters have interpretations in terms of the L-moments. In particular, the rth L-moment of Q_k

is (suppressing the dependence on β)

$$
\begin{aligned}
\lambda_{kr} &= (2r-1)^{-1/2}\int_0^1 Q_k(u)h_{r-1}(u)\mathrm{d}u \\
&= (2r-1)^{-1/2}\int_0^1 \left(Q_0(u)+\sum_{j=1}^{k}\theta_j(2j-1)^{1/2}h_{j-1}(u)\right)h_{r-1}(u)\mathrm{d}u \\
&= (2r-1)^{-1/2}\int_0^1 Q_0(u)h_{r-1}(u)\mathrm{d}u \\
&\quad +(2r-1)^{-1/2}\sum_{j=1}^{k}\theta_j(2j-1)^{1/2}\int_0^1 h_{j-1}(u)h_{r-1}(u)\mathrm{d}u \\
&= \lambda_{0r}+\theta_r,
\end{aligned}
$$

where λ_{0r} is the rth L-moment of the hypothesised distribution. This demonstrates that $\theta_r = \lambda_{kr} - \lambda_{0r}$ measures the deviation from the rth L-moment of the hypothesised distribution. Note that (3) also arises from the approximations of the quantile functions $Q_k(u;\beta,\theta)$ and $Q_0(u;\beta)$, using the truncated expansions of Sillitto (1969).

Under the regularity conditions, note that $\hat{\theta}_r = \theta_r(X,\hat{\beta})$ is a consistent estimator of θ_r. Thus, at the rejection of the null hypothesis, the expansion of (3), with θ_j replaced with $\hat{\theta}_j$, may be used as a diagnostic tool. This gives an estimate of the true quantile function, which can be considered as an improvement over the hypothesised distribution. Because $\hat{\beta}$ is the MLME and $\theta_j(X,\hat{\beta})$ only depends on X through the sample L-moments $\hat{\lambda}$, we write $Q_k(u;\hat{\lambda})$ for $Q_k(u;\hat{\beta},\hat{\theta})$. Thas et al. (2009) studied density estimators that were built on the same idea, except that the density estimators were related to smooth alternatives to the hypothesised probability density function. We refer to $Q_k(u;\hat{\lambda})$ as an *improved quantile estimator*. Along the lines of Efron & Tibshirani (1996), $Q_k(u;\hat{\lambda})$ may also be interpreted as a nonparametric quantile function estimator that has the special property that its first k L-moments agree exactly with the corresponding sample L-moments.

LaRiccia (1991) proposed smooth tests of goodness of fit for distributions in a location-scale family, starting from an expansion of $Q(u)$ similar to (3), but he considered functions other than $h_j(u)$. However note that LaRiccia's work does not involve L-moments.

We now relate our test with goodness of fit tests based on the *empirical quantile function* (EQF), which is defined for $u \in [0,1]$, by

$$\hat{Q}(u) = X_{i:n} \text{ if } \frac{i-1}{n} \leq u < \frac{i}{n} \text{ for some } 1 \leq i \leq n.$$

While not referring to L-moments, del Barrio et al. (1999), del Bario et al. (2000) and de Wet (2002) studied EQF tests. In particular, the unweighted test statistic is of the form

$$W = n \int_0^1 \left(\hat{Q}(u) - Q_0(u; \tilde{\beta}) \right)^2 \mathrm{d}u, \tag{4}$$

in which $\tilde{\beta}$ is an estimator of β, different from the estimators discussed previously in this paper. The integral in (4) is the Wasserstein distance between the EQF and the hypothesised quantile function. If the MLME $\hat{\beta}$ is used instead of $\tilde{\beta}$ and the

EQF is replaced with $Q_k(u;\hat{\lambda})$, we find a new statistic,

$$\begin{aligned} W_k &= n\int_0^1 \left(Q_k(u;\hat{\lambda}) - Q_0(u;\hat{\beta})\right)^2 \mathrm{d}u = n\int_0^1 \left(\sum_{j=p+1}^{k} \hat{\theta}_j (2j-1)^{1/2} h_{j-1}(u)\right)^2 \mathrm{d}u \\ &= n \sum_{i=p+1}^{k} \sum_{j=p+1}^{k} \hat{\theta}_i \hat{\theta}_j (2i-1)^{1/2}(2j-1)^{1/2} \int_0^1 h_{i-1}(u) h_{j-1}(u) \mathrm{d}u \\ &= n \sum_{j=p+1}^{k} (2j-1)\hat{\theta}_j^2. \end{aligned}$$

The components $\hat{\theta}_j$ of our test statistic are thus also the components in the decomposition of the EQF statistic when MLME is used for nuisance parameter estimation and when $Q_k(u;\hat{\lambda})$ is used as the nonparametric quantile function estimate. Although the final form of W_k suggests that the terms $\hat{\theta}_j^2$ receive weights $(2j-1)$, we note that the asymptotic variance of $\hat{\theta}_j$ is proportional to $1/(2j-1)$. Thus, with σ_j^2 denoting the asymptotic variance of $n^{1/2}\hat{\theta}_j$ under H_0^{SP}, $V_j = \hat{\theta}_j/\sigma_j$ the jth component test statistic of Section 3.2, and $c_j^2 = (2j-1)\sigma_j^2$, the test statistic W_k can be expressed as

$$W_k = n \sum_{j=p+1}^{k} c_j^2 V_j^2,$$

which is a weighted sum of the component test statistics. When k is finite, W_k has a proper limiting null distribution under the same assumptions required for the convergence of the $n^{1/2}\hat{\theta}_j$ (under H_0). However, in the limit as $k \to \infty$, convergence of W_k depends on the c_j. The study of this convergence is beyond the scope of this paper, and the interested readers are referred to the papers of del Barrio et al. (1999), del Bario et al. (2000), and de Wet (2002) in which similar convergence issues as discussed.

The asymptotic null distribution of W_k, as $n \to \infty$, is stated in Corollary 1 in Appendix 2 of the Supplementary Material. The proof is a direct consequence of the asymptotic null distribution of $\hat{\theta}$ and the application of the singular value decomposition.

Finally, we note that the EQF statistic (4) is directly related to the area under the curve of the QQ-plot. This suggests that the use of a QQ-plot as a graphical diagnostic tool for goodness of fit can benefit from a formal test based on the $\hat{\theta}_i$ components, and vice versa. We refer to chapters 3 and 5 of Thas (2010) for a detailed discussion on the interplay between QQ and PP-plots with smooth, EDF, and EQF tests.

4 SIMULATION STUDY

In this section the behavior of the new class of tests is empirically investigated in a simulation study. Two distributions have been selected for the null hypothesis: the logistic distribution and the GPD. The GPD is included because it formed the motivation for this research; see Section 1. The logistic is chosen because its quantile function has a simple expression and because many competitor tests are available for inclusion in the simulation study. For both distributions, the quantile functions, MLMEs of the nuisance parameters and the elements of the Δ and C matrices required for the construction of the GLST, are presented in Supplementary Material Appendix 1.

Finally, we mention that the logistic distribution forms a location-scale family with two nuisance parameters. It is a symmetric distribution, which is often visually hard to distinguish from a normal distribution. The GPD included in this study has both a scale and a shape parameter. It is a skew distribution and its support depends the parameters.

4.1 Data-Driven Tests

The simulation studies in this section also include data-driven versions of the LGST for which the order k is selected from the data. This class of tests is well known for classical smooth tests in combination with MME or MLE; see Thas (2010, section 4.3) for a detailed overview and references. Although empirical power studies published in the statistical literature generally indicate that data-driven tests have good overall power, the simulation studies will illustrate that this behaviour does not apply to the data-driven versions of the LGST studied here. Therefore, we have chosen not to present the theory of the data-driven LGST in the body of this paper, but still include the empirical powers in the tables following. The description of the data-driven LGST can be found in Appendix 2 of the Supplementary Material.

4.2 The Logistic Distribution

The new L-moments generalised smooth test (LGST) and its component tests (LV3 and LV4 for $r = 3$ and $r = 4$, respectively) are compared to selected competitor tests. In particular we have included the generalised smooth test of Thas et al. (2009) (GST), its component tests (V3 and V4 for $r = 3$ and $r = 4$, respectively), the Anderson-Darling (AD) and Cramér-von Mises tests (CvM), both described by Stephens (1979), the empirical characteristic function test (Epps, 2005; Meintanis, 2004) based on the MME (ECF1) and the MLE (ECF2) of the nuisance parameters and the empirical moment generating function tests (Meintanis, 2004) based on MME (EMGF1) and MLE (EMGF2). We choose the GST and LGST tests to have $k = 4$. We have also included a data-driven LGST (DD), based on the modified BIC criterion, BIC_m, with $\mathcal{M}$ containing increasing index sets up to order $k = 6$. Finally, the test related to the Wasserstein distance (Wk) is included with $k = 6$. The alternatives considered here, are similar to those studied by Thas et al. (2009).

As alternatives to the logistic distribution, the exponential, gamma and normal distributions are considered.

Because the logistic distribution is from a location-scale family, the null distributions of all test statistics, except those for the DD and Wk tests, were approximated using 1,000 Monte Carlo simulation runs. For the DD and Wk tests, which are not location-scale invariant, the parametric bootstrap (1,000 runs) is used for p-value calculations. All tests are performed at the 5% level of significance and the empirical powers are obtained based on 1,000 simulation runs. Table 1 shows the emprical levels and powers.

The levels of all tests are very close to the nominal level of 5%. For the exponential alternatives many tests (AD, CvM, LGST, LV3, ECF1, ECF2, DD and Wk) have very large powers and do not differ by more than 5% from one another. The MME-based GST test and its component tests V3 and V4 have much smaller powers. This holds also for the EMGF tests. Note that the LV4 fourth order component test has small power too. This may be explained by the fact that for both the logistic and the exponential distribution $\lambda_2/\lambda_4 = 6$. Because the use of MLME makes the second L-moments of the fitted logistic distribution and the sample of the exponential distribution coincide, their 4th order L-moments will also be equal. The reason for the power of the LV4 test not being exactly equal to the nominal significance level of 5% is that the variance estimator used in LV4 depends on L-moments up to order 8 and not all these L-moments agree between the fitted logistic and the exponential sample observations. See Henze (1997) for a theoretical treatment of this issue.

For the gamma alternatives the new component test LV3 outperforms all the others. The ECF1 test is the second best. Overall the LGST test has also very good power. The data-driven and Wasserstein-type L-moments based tests (DD and Wk) are less powerful, which may be explained by the very good power of LV3 as compared to LV4. In this case, the DD and Wk tests exhibit a dilution effect, i.e. they are sensitive to L-moment deviations up to order six, but because a strong deviation is present in the third order L-moment (skewness), they loose power by also assessing the higher orders.

Finally, the powers of all tests for detecting a normal distribution, are very poor. Only the V4 and LV4 tests show some power.

Table 1: Empirical levels (%) (for the logistic distribution) and powers (%) (for the exponential, gamma and normal distributions) of tests for the logistic distribution at the 5% level of significance, for a sample size of $n = 50$. The parameters p_1 and p_2 refer to the mean and scale parameter for the logistic distribution, the rate and shape parameter for the gamma distribution and the mean and standard deviation for the normal distributions. For the exponential distribution p_1 gives the rate parameter. Results are based on 1,000 Monte Carlo simulation runs.

Parameter		Test													
p_1	p_2	LGST	LV3	LV4	AD	CvM	GST	V3	V4	ECF1	ECF2	EMGF1	EMGF2	DD	Wk
logistic															
0.00	0.70	4	4	4	5	4	6	5	5	5	4	5	5	7	4
0.00	1.00	4	4	4	5	5	5	5	5	5	5	6	5	5	3
0.00	1.30	3	4	4	5	5	5	5	5	5	5	5	5	5	3
exponential															
0.40		100	99	9	99	96	71	76	20	100	99	40	55	99	99
0.55		100	99	9	99	96	71	76	20	99	99	40	55	100	99
0.80		100	99	9	99	95	71	76	20	99	99	40	55	99	99
gamma															
1.00	1.50	95	96	8	94	78	70	74	14	95	91	29	38	42	89
1.00	3.00	63	71	10	61	41	56	56	11	67	58	17	17	55	44
1.00	5.00	37	46	10	38	25	40	40	12	43	36	14	9	32	22
1.00	7.50	23	30	10	26	17	30	29	12	30	25	12	5	22	12
normal															
0.00	0.55	5	2	11	6	7	5	3	15	7	6	9	0	6	1
0.00	2.50	5	2	9	6	7	5	3	14	7	6	9	0	6	1

4.3 The Generalised Pareto Distribution

For the GPD we have included four alternatives: the gamma, Weibull, lognormal, and inverse normal distributions. Each of these alternatives has also been considered in many other studies on goodness of fit testing for the GPD and they cover a large subset of the GDP nuisance parameter space. We refer to De Boeck et al. (2011) and the references therein for a more detailed discussion on the alternatives and the competitor tests. The LGST test and its component tests (LV3 and LV4) are compared with the Anderson-Darling (AD) and Cramér-von Mises (CvM) tests of Choulakian & Stephens (2001), the Neyman smooth test of Radouane & Crétois (2002) based on MME (NS1) and MLE (NS2) and the generalised smooth test of order 4 (GST) and its components tests (V3 and V4) (De Boeck et al., 2011). The data-driven test (DD) and the Wasserstein-type test (Wk) are also included ($k = 6$ and p-value calculation as for the logistic distribution). All tests are performed at the 5% level of significance and are implemented as parametric bootstrap tests using 750 and 500 bootstrap samples for sample size $n < 100$ and $n = 100$, respectively. Empirical levels and powers are computed based on 1,000 simulation runs. The results are presented in Tables 2 (levels) and 3 (power).

All tests have empirical sizes close to the nominal level of 5%, except for the GST tests, which are conservative for GPDs with negative shape parameter. The latter may be related to the non-existence of the MME when the shape parameter κ is smaller than $-1/(2k)$, with $k = 4$, the order of GST. The same is observed for the MME-based NS1 test. Although the MLE does not exist for $\kappa \geq 1$ (Smith, 1984), we were able to apply MLE-based tests, because the profile likelihood estimator of Davidson (1984) was possible. Choulakian & Stephens (2001) advocated this estimator for their AD and CvM tests. De Boeck et al. (2011) gave a more detailed discussion on the existence problems of the MME and MLE and their use in goodness of fit tests for the GPD. The MLME does not suffer from these problems.

For the gamma alternatives the largest powers are always obtained with the third order component test based on MME (V3), which are slightly larger than the powers of the Wk and/or LV3 tests. For the Weibull alternatives the V3 test has again the highest powers, except when the shape parameter equals 0.75, in which case it shows a complete power break-down and the LV3 test is the best. At this point we should remark that the powers of the MME-based tests (e.g. V3 and NS1) cannot be trusted for the following reasons. Table 4 shows the expected values of the shape parameter estimators for $n = 100$, using MLE, MME and MLME for the four distributions used in this simulations study. For the Weibull with shape parameter 0.75 and the log-normal distributions the estimators estimate the GPD parameter to be negative, for which it is known that MME is not reliable (Hosking & Wallis, 1987). The levels of the MME-based NS1 and V3 tests are indeed not controlled for negative shape parameters, as illustrated in Table 2.

Finally, we note that the Wk test has overall rather good powers for the Weibull alternatives. For the log-normal and inverse normal alternatives LV3 is generally the best, while the MME-based tests under-perform and the MLE-based AD test is rather good.

From the simulation study, we may conclude that tests focussing on the skewness (V3 and LV3) are often the best. However, the MME-based V3 test may suffer from existence issues when the estimated shape parameter is negative. The L-moments based LV3 test does not suffer from this problem. We do not recommend the tests based of GST and LGST, and the higher order tests based on V4 and LV4 for the alternatives considered. Overall the Wasserstein test (Wk) generally performs better than the data-driven DD test. Finally, we note that the MLE-based tests (e.g. AD) is almost never most powerful.

Table 2: Empirical levels (%) of tests for the GPD performed at the 5% level of significance; p_1 and p_2 are the scale and shape parameters, respectively. Results are based on 1,000 Monte Carlo simulation runs.

Parameter			Test											
p_1	p_2	n	LGST	LV3	LV4	AD	CvM	NS2	NS1	GST	V3	V4	DD	Wk
1	-0.75	20	6	3	6	3	3	4	12	1	2	2	4	3
1	-0.75	50	4	6	5	5	5	5	17	3	0	7	5	5
1	-0.35	20	5	4	4	3	3	5	6	1	1	2	4	2
1	-0.35	50	5	5	5	5	5	5	5	1	0	1	5	2
1	-0.15	20	4	5	6	3	2	4	5	2	2	3	5	4
1	-0.15	50	5	4	4	4	3	4	4	1	1	2	5	2
1	0.25	20	5	5	4	6	6	6	5	5	4	5	5	4
1	0.25	50	6	5	5	4	4	5	5	5	4	5	5	4
1	0.75	20	5	5	4	7	7	6	4	6	6	7	6	2
1	0.75	50	7	6	6	6	6	7	7	6	5	6	5	4
1	1.25	20	6	5	5	6	6	6	4	7	6	6	5	2
1	1.25	50	4	4	5	4	4	4	8	5	5	6	4	3

Table 3: Empirical powers (%) of tests for the GPD performed at the 5% level of significance, for sample sizes $n = 30, 50, 60$ and $n = 100$; p_1 and p_2 are the rate and shape parameters of the gamma and Weibull distributions, the mean and standard deviation (log scale) of the log-normal distribution and the location and scale parameter of the inverse normal distribution. Results are based on 1,000 Monte Carlo simulation runs.

Parameter			Test											
p_1	p_2	n	LGST	LV3	LV4	AD	CvM	NS2	NS1	GST	V3	V4	DD	Wk
gamma														
1.5	2	60	49	71	52	57	53	53	49	66	80	46	43	73
1.5	2	100	84	94	80	88	82	83	76	88	96	65	86	92
1	2	30	10	28	17	15	17	19	20	38	51	30	15	43
1	2	50	33	63	40	47	44	42	43	59	75	40	29	66
1	2	100	87	95	82	92	85	88	85	91	97	68	89	94
Weibull														
2	1.50	60	27	46	41	41	39	35	34	60	72	49	30	60
1	1.25	100	20	30	27	26	26	21	23	34	43	20	27	37
1	0.75	30	10	18	11	16	8	11	19	3	0	6	11	3
2	1.50	100	60	77	66	72	66	63	53	80	91	63	68	82
1	1.25	30	5	7	8	5	6	5	7	10	13	8	6	14
1	0.75	50	23	29	20	28	16	20	29	3	0	9	21	8
1	1.25	50	7	11	12	11	12	11	12	19	25	14	9	22
1	0.75	100	45	54	40	51	33	37	49	4	0	9	47	19
log-normal														
0	1	30	14	24	13	18	19	17	15	6	7	2	11	15
0	1	50	37	48	22	41	35	36	35	11	8	1	23	22
0	1	100	86	82	43	84	72	80	78	12	9	1	69	41
inverse normal														
1	1	30	31	57	18	36	35	36	37	19	23	5	27	33
1	1	50	82	85	40	76	65	75	74	34	32	6	59	53
1	1	100	100	100	75	100	98	100	99	60	48	5	98	84

5 EXAMPLE

Dupuis & Tsao (1998) fit a generalised extreme value distribution to 22 annual maximum December temperatures at Fair Isle Weather Station, UK for 1974 to 1995. The data in degrees Celsius are presented in Table 5.

We proceed as Choulakian & Stephens (2001) by testing for goodness of fit for the GPD on the complete data set, setting the threshold (i.e. lower limit of the GPD) at the smallest observed data point while removing this observation from the data, and later progressively increasing the threshold by removing the smallest observations one by one until the test gives an insignificant result. At most three consecutive observations are deleted. We do not recommend the approach they adopted, but for comparison purposes proceed along the same lines. We analyse the data with all tests for the GPD that were included in the simulation study of Section 4.3, except for the NS test which does not exist for this dataset (it involves the logarithm of a negative result).

Table 6 reports the parameter estimates with MLE, MME and MLME, and the results of the hypothesis tests, based on $10{,}000$ bootstrap runs. The data have also been analysed by De Boeck et al. (2011). At the 5% level of significance, their MME-based GST tests gave significant results for all three thresholds, whereas the MLE-based tests give p-values larger than 5% for the two larger thresholds (9.0 and 9.2). These inconsistent results might have been caused by the parameter estimation: the parameter estimates indicate that the GPD shape parameter is larger then 0, or even larger than 1, for which the MME behaves better than the MLE. Therefore, De Boeck et al. (2011) argued that for this data set the MME-based

Table 4: Approximate expected values of the estimators (MLE, MME and MLME) of the GPD shape parameter for several distributions (gamma, Weibull, log-normal and inverse normal) and for sample size $n = 100$. The meaning of p_1 and p_2 is as for Table 3. Approximations are based on 10,000 simulation runs.

Parameter		Estimator		
p_1	p_2	MLE	MME	MLME
		gamma		
1.5	2	0.32	0.54	0.67
1	2	0.32	0.55	0.68
		Weibull		
2	1.50	0.39	0.62	0.71
1	1.25	0.24	0.30	0.36
1	0.75	-0.38	-0.20	-0.33
		log-normal		
0	1	-0.11	-0.13	-0.06
		inverse normal		
1	1	0.04	0.06	0.21

Table 5: Annual maximum December temperatures at the Fair Isle weather station, UK, for 1974 to 1995.

10.3	10.2	9.0	10.9	7.5	10.7	10.0	11.5	9.2	10.2	10.4
10.6	9.6	10.4	9.7	9.4	9.4	9.3	10.4	10.3	10.4	10.1

GST may be trusted more than the MLE-based tests. The current results with the MLME based LGST tests, however, rather confirm the results of the MLE-based tests: no significant deviation from a GPD distribution for the two larger thresholds 9.0 and 9.2. For the smallest threshold of 7.5 all tests agree that the sample cannot be described by a GPD distribution.

Figure 1 shows the QQ-plots of the data with thresholds 7.5 and 9.0; the plot for the threshold of 9.2 is similar to the latter and is not shown here. The plots show both the QQ-plot of the observed quantiles versus the expected, and the QQ-plot of the observed versus the expected quantiles from the improved quantile function estimate of (3) with $k = 4$. These plots demonstrate that the QQ-plot for the threshold of 7.5 can be improved by including two additional terms. For the threshold of 9.2 not much improvement can be achieved, for the GPD fits well. The convenient agreement of the conclusions from the QQ-plot and the formal LGST test is a consequence of the direct link between these two goodness of fit tools (Section 3.3).

The advantage of using an MLME-based test is that the existence of the parameter estimators is no longer an issue when interpreting test results. Moreover, the simulation study of Section 4 demonstrated that the LGST tests have good powers in general.

6 DISCUSSION

We have proposed a class of smooth tests of goodness of fit. The tests make use of the method of L-moment estimation for nuisance parameter estimation, and

Table 6: Results for the Fair Isle data for the three thresholds: the first four sample L-moments, parameter estimates with MME, MLE, and MLME, and the results of several hypothesis tests (observed test statistics are given, with p-values in parentheses).

	threshold		
	7.5	9.0	9.2
sample L-moments			
$\hat{\lambda}_1$	10.0952	10.1500	10.2000
$\hat{\lambda}_2$	0.3543	0.3311	0.3123
$\hat{\lambda}_3$	-0.0114	-0.0046	-0.0017
$\hat{\lambda}_4$	0.0481	0.0515	0.0610
parameter estimates			
$\tilde{\kappa}$ (MME)	8.65	1.54	1.22
$\tilde{\sigma}$ (MME)	25.04	2.93	2.22
$\tilde{\kappa}$ (MLE)	1.27	0.77	0.70
$\tilde{\sigma}$ (MLE)	5.09	1.97	1.65
$\hat{\kappa}$ (MLME)	5.33	1.47	1.20
$\hat{\sigma}$ (MLME)	6.63	2.94	2.20
test results			
AD	2.97 (0.002)	0.946 (0.176)	0.80 (0.286)
CvM	0.57 (0.001)	0.165 (0.134)	0.15 (0.192)
LGST	21.61 (0.027)	2.64 (0.280)	-3.98 (0.813)
LV3	3.86 (0.011)	1.01 (0.168)	0.35 (0.314)
LV4	-0.41 (0.674)	1.18 (0.090)	1.33 (0.073)
GST	110.06 (< 0.001)	28.11 (0.013)	27.60 (0.011)
V3	3.44 (< 0.001)	1.73 (0.174)	1.21 (0.326)
V4	9.91 (< 0.001)	5.01 (0.012)	5.11 (0.020)

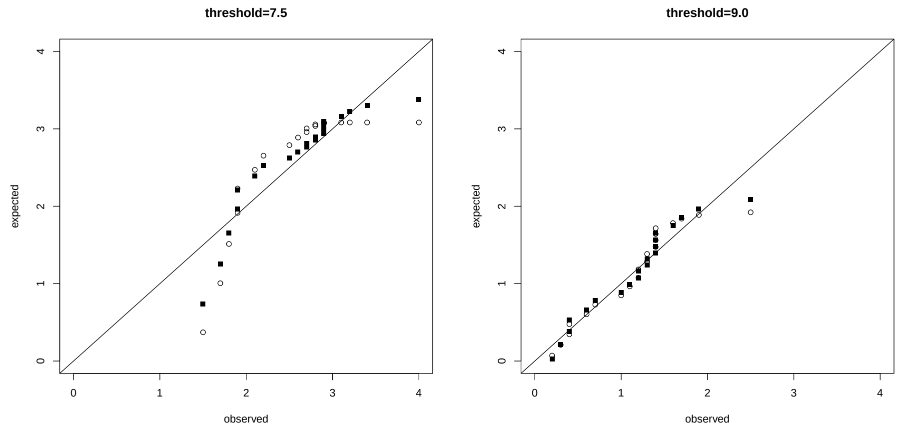

Figure 1: QQ-plots for the Fair Isle data with a threshold of 7.5 (left) and 9.0 (right). The circles represent the QQ-plots for the hypothesised GPD and the black squares represent the QQ-plots for the improved quantile function with $k = 4$.

the components of the test statistic may be interpreted in terms of L-moments. The test is basically a semiparametric test for testing equality of the first few L-moments of the true distribution with those of the hypothesised distribution. Our tests are particularly useful when the hypothesised distribution has a convenient quantile function, or when other types of estimators (e.g. maximum likelihood or method of moments) suffer from existence problems. The logistic, generalised lambda, g-and-h, g-and-k and the generalised Pareto distributions (GPD) are examples. In this paper we have demonstrated that the new test has good power for testing for the GPD and the logistic distributions. The new test often outperforms many of its competitor tests in a simulation study. Moreover, the new tests do not suffer from existence problems; this is particularly useful for the GPD distribution.

We have outlined our methods for continuous distributions, but as the definition of L-moments applies equally well to discrete random variables, our methods can be applied for testing goodness of fit for discrete distributions too.

The tests proposed in this paper resemble the generalised smooth tests Rayner et al. (2009), but with MLE or MME replaced with MLME, and with smooth alternatives to the quantile function rather than to the density function.

L-moments have been used before in methods for assessing the goodness of fit to hypothesised distributions. Harri & Coble (2011) constructed tests for normality, using estimators of the L-skew (λ_3/λ_2) and L-kurtosis (λ_4/λ_2). These tests may to some extent be considered analogous to our LV3 and LV4 tests, but our theory is more generally applicable.

Chowdhurry et al. (1991) also used estimators of the L-moment ratios for constructing tests for *regional* generalised extreme value (GEV) distributions. A *regional* distribution is a term that is used in hydrology in settings where one has data series from several sites in a region (e.g. water level time series measured at many location along a coast line). The *regional* distribution is a distribution that holds for all sites in that region. The null hypotheses is that all data series from sites in a region can be described by a hypothesised distribution; see e.g. Chowdhurry et al. (1991); Peel et al. (2001); Vogel et al. (2009) and the references therein for methods relying on L-moments. Many authors also recommend the use of L-moment diagrams as a visual diagnostic tool for assessing the fit of a regional distribution (Hosking, 1990; Vogel & Fennessey, 1993; Hosking & Wallis, 1997, among others). For each data set the estimated L-kurtosis is plotted against the estimated L-skew. Each potential regional distribution is added as a line or a region. This is possible because for many distributions an exact or approximate relation between L-skew and L-kurtosis can be obtained. If the points are closely scattered about one of the lines, the corresponding distribution is chosen as an appropriate regional distribution. For formal hypothesis testing for a single data set, we believe that L-moment diagrams add little value.

Many goodness of fit methods start from QQ-plots, which are also known as probability plots. For example, the Shapiro-Wilk (Shapiro & Wilk, 1965) and the Filliben (Filliben, 1975) tests for normality are related to the correlation coefficient in a particularly constructed QQ-plot. See D'Agostino & Stephens (1986) and Thas (2010) for detailed discussions of this type of test for distributions fitted with the method of moments or with maximum likelihood. Such *Probability Plot Correlation Coefficient* (PPCC) tests have also been studied by several authors for distributions fitted with the method of L-moments (Chowdhurry et al., 1991; Vogel et al., 2009) and good power has been claimed. However, disadvantages of PPCC tests are that for many distributions no solid theory is available, choices of plotting positions should be made, and null distribution approximation is computationally intensive or one has to rely on published regression equations for percentile approximations; see for example Heo et al. (2008). In this paper, we have also recommended the use of QQ-plots when nuisance parameters are estimated by the method of L-moments, but we have demonstrated that the component test statistics V_j^2 (or the unscaled $\hat{\theta}_j$) are related to the sample version of the area under the curve of the QQ-plot. In this sense, when MLME is used, we advocate the use of the graphical QQ-plot tool in combination with our tests. This procedure will minimise the chance of conflicting conclusions from the graphical and the formal testing assessments.

The idea of using L-moments for the construction of a goodness of fit test may also be employed for the two-sample and K-sample cases in which the null hypothesis is expressed as the equality of the first few L-moments of the two or K distributions. A test statistic may then be constructed based on the contrasts of the L-moments sample estimators. For example, $\hat{\theta}_r = \hat{\lambda}_{1r} - \hat{\lambda}_{2r}$ with $\hat{\lambda}_{1r}$ and $\hat{\lambda}_{2r}$ the estimators of the r-th L-moments of the two distributions to be compared.

Acknowledgements

This work was carried out using the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI. Olivier Thas and Jan De Neve acknowledge financial support from the IAP research network P7/06 of the Belgian Government (Belgian Science Policy).

The authors also want to thank the associate editor and the two referees whose comments considerably improved the paper.

A Proof of Theorem 1

The inverse of the function $\lambda_{0\beta}(\beta)$ is expressed as $\beta(\lambda_{0\beta})$; its existence is guaranteed by (A2) and the Inverse Function Theorem.

The $(k-p)\times p$ matrix $\Delta(\beta^*)$ has element (i,j) equal to $(i=1,\ldots,k-p; j=1,\ldots,p)$

$$\frac{\partial}{\partial\beta_j}\,\lambda_{0(i+p)}(\beta)\big|_{\beta=\beta^*}$$

(these derivatives exist due to (A3)). Let $M(\beta)$ denote the $(k-p)\times k$ matrix with first p columns given by $-\Delta(\beta)C(\lambda_{0\beta})$ and with the last $(k-p)$ columns given by the $(k-p)\times(k-p)$ identity matrix.

Consider now the nested Taylor expansion of $\hat\theta = \theta(X,\hat\beta) = \hat\lambda_\theta - \lambda_{0\theta}(\hat\beta)$ (using (A2) and (A3)),

$$\begin{aligned}
\theta(X,\hat\beta) &= \hat\lambda_\theta - \Big(\lambda_{0\theta}(\beta) + \Delta(\beta)(\hat\beta-\beta) + o_P(\hat\beta-\beta)\Big)\\
&= \hat\lambda_\theta - \lambda_{0\theta}(\beta) - \Delta(\beta)\Big(\beta(\hat\lambda)-\beta(\lambda)\Big) + o_P(\hat\beta-\beta)\\
&= \hat\lambda_\theta - \lambda_{0\theta}(\beta) - \Delta(\beta)\Big(\beta + C(\lambda_{0\beta})(\hat\lambda_{0\beta}-\lambda_{0\beta}) + o_P(\hat\lambda_{0\beta}-\lambda_{0\beta}) - \beta\Big)\\
&\quad + o_P(\hat\beta-\beta)\\
&= \hat\lambda_\theta - \lambda_{0\theta}(\beta) - \Delta(\beta)C(\lambda_{0\beta})(\hat\lambda_{0\beta}-\lambda_{0\beta}) + o_P(\hat\beta-\beta) + o_P(\hat\lambda_{0\beta}-\lambda_{0\beta}).
\end{aligned}$$

With $M(\beta)$ as defined previously, we may write $\theta(\hat\beta) = M(\beta)\Big(\hat\lambda - \lambda_0(\beta)\Big) + o_P(n^{-1/2})$, from which we find the asymptotic covariance matrix of $n^{1/2}\theta(\hat\beta)$ under H_0^{SP}. Zero-mean normality follows from the asymptotic normality and consistency of the sample L-moments and the MLME $\hat\beta$ under assumptions (A1) and (A2) and the semiparametric null hypothesis.

References

Castillo, E., & Hadi, A. (1997). Fitting the generalized Pareto distribution to data. *Journal of the American Statistical Association*, *92*, 1609-1620.

Chen, G., & Balakrishnan, N. (1995). The infeasibility of probability weighted moments estimation of some generalized distributions. In N. Balakrishnan (Ed.), *Recent advances in life-testing and reliability.*

Choulakian, V., & Stephens, M. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, *43*, 478-484.

Chowdhurry, J., Stedinger, J., & Lu, L. (1991). Goodness-of-fit tests for regional the generalized extreme value flood distributions. *Water Resources Research*, *27*, 1765-1776.

Claeskens, G., & Hjort, N. (2004). Goodness of fit via non-parametric likelihood ratios. *Scandinavian Journal of Statistics*, *31*, 487-513.

D'Agostino, R., & Stephens, M. (1986). *Goodness-of-fit techniques*. New York, USA: Marcel Dekker.

Davidson, A. (1984). Modeling excesses over high thresholds, with an application. In J. Tiage de Oliveira (Ed.), *Statistical extremes and applications* (p. 621-635). Dordrecht: Reidel.

de Wet, T. (2002). Goodness-of-fit tests for location and scale families based on a weighted L_2-Wasserstein distance measure. *Test*, *11*, 89-107.

De Boeck, B., Thas, O., Rayner, J., & Best, D. (2011). Generalised smooth tests for the generalised pareto distribution. *Journal of Statistical Theory and Practice*, *4*, 737-750.

del Bario, E., Cuesta-Albertos, J., & Matrán, C. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, *9*, 1-96.

del Barrio, E., Cuesta Albertos, J., & Matrán Y J. Rodríguez Rodríguez, C. (1999). Tests of goodness of fit based on the L2-Wasserstein distance. *Annals of Statistics*, *27*, 1230-1239.

Dupuis, D., & Tsao, M. (1998). A hybrid estimator for generalized Pareto and extreme-value distributions. *Communications in Statistics - Theory and Methods*, *27*, 925-941.

Efron, B., & Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Annals of Statistics*, *24*, 2431-2461.

Elamir, E., & Seheult, A. (2004). Exact variance structure of sample L-moments. *Journal of Statistical Planning and Inference*, *124*, 337-359.

Epps, T. (2005). Tests for location-scale families based on the empirical characteristic function. *Metrika*, *62*, 99-114.

Filliben, J. (1975). The probability plot coefficient test for normality. *Technometrics*, *17*, 111-117.

Greenwood, J., Landwehr, J., Matalas, N., & Wallis, J. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, *15*, 1049-1054.

Harri, A., & Coble, K. (2011). Normality testing: Two new tests using L-moments. *Journal of Applied Statistics*, *38*, 1369-1379.

Henze, N. (1997). Do components of smooth tests of fit have diagnostic properties? *Metrika*, *45*, 121-130.

Henze, N., & Klar, B. (1996). Properly rescaled components of smooth tests of fit are diagnostic. *Australian Journal of Statistics*, *38*, 61-74.

Heo, J., Kho, Y., Shin, H., & Kim, S. (2008). Regression equations of probability plot correlation coefficient test statistics from several probability distributions. *Journal of Hydrology*, *355*, 1-15.

Hosking, J. (1986). *The theory of probability weighted moments* (Tech. Rep. No. 12210). IBM Research, Yorktown Heights.

Hosking, J. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B*, *52*, 105-124.

Hosking, J. (2007). Distributions with maximum entropy subject to constraints on their l-moments or expected order statistics. *Journal of Statistical Planning and Inference*, *137*, 2870-2891.

Hosking, J. (2009). *The theory of probability weighted moments* (Tech. Rep. No. 12210 (2nd edition)). IBM Research, Yorktown Heights.

Hosking, J., & Wallis, J. (1987). Parameter and quantile estimation for the generalised Pareto distribution. *Technometrics*, *29*, 339-349.

Hosking, J., & Wallis, J. (1997). *Regional frequency analysis: An approach based on L-moments*. Cambridge, UK: Cambridge University Press.

Hosking, J., Wallis, J., & Wood, E. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, *27*, 251-261.

Inglot, T., Kallenberg, W., & Ledwina, T. (1997). Data driven smooth tests for composite hypotheses. *Annals of Statistics*, *25*, 1222-1250.

Janic, A., & Ledwina, T. (2009). Data-driven smooth tests for a location-scale family revisited. *Journal of Statistical Theory and Practice*, *3*(3), 645–664.

Kallenberg, W., & Ledwina, T. (1997). Data-driven smooth tests when the hypothesis is composite. *Journal of the American Statistical Association*, *92*, 1094-1104.

Klar, B. (2000). Diagnostic smooth tests of fit. *Metrika*, *52*, 237-252.

LaRiccia, V. (1991). Smooth goodness-of-fit tests: A quantile function approach. *Journal of the American Statistical Association*, *86*, 427-431.

Ledwina, T. (1994). Data-driven version of Neyman's smooth test of fit. *Journal of the American Statistical Association*, *89*, 1000-1005.

Meintanis, S. (2004). Goodness-of-fit tests for the logistic distribution based on empirical transforms. *Sankhya, Series B*, *66*, 306-326.

Neyman, J. (1937). Smooth test for goodness of fit. *Skandinavisk Aktuarietidskrift*, *20*, 149-199.

Peel, M., Wang, Q., Vogel, R., & McMahon, T. (2001). The utility of L-moment ratio diagrams for selecting a regional probability distribution. *Hydrological Sciences Journal*, *46*, 147-155.

Radouane, O., & Crétois, E. (2002). Neyman smooth test for the generalized Pareto distribution. *Communications in Statistics - Theory and Methods*, *31*, 1067-1078.

Rayner, J., Thas, O., & Best, D. (2009). *Smooth tests of goodness of fit: Using R*. Singapore: Wiley.

Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611.

Sillitto, G. (1969). Derivations of approximations of the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika*, *56*, 641-650.

Smith, R. (1984). Threshold methods for sample extremes. In J. Tiage de Oliveira (Ed.), *Statistical extremes and applications* (p. 621-635). Dordrecht: Reidel.

Stephens, M. (1979). Tests of fit for the logistic distribution based on the empirical disrtibution function. *Biometrika*, *66*, 591-595.

Thas, O. (2010). *Comparing distributions*. New York, USA: Springer.

Thas, O., Rayner, J., Best, D., & De Boeck, B. (2009). Informative statistical analyses using smooth goodness-of-fit tests. *Journmal of Statistical Theory and Practice*, *3*, 705-733.

Vogel, R., & Fennessey, N. (1993). L moment diagrams should replace product moment diagrams. *Water Resources Research*, *29*, 1745-1752.

Vogel, R., Hosking, J., Elphick, C., Roberts, D., & Reed, J. (2009). Goodness of fit of probability distributions for sightings of species approach extinction. *Bulletin of Mathematical Biology*, *71*, 701-719.

Zhang, J. (2007). Likelihood moment estimation for the generalized Pareto distribution. *Australian and New Zealand Journal of Statistics*, *49*, 69-77.

Supplementary Material

Supl. Mat. App. 1 The Logistic and the Generalised Pareto Distributions

The Logistic Distribution

The logistic disrtibution has quantile function

$$Q_0(u) = \mu + \sigma \log \frac{u}{1-u}.$$

The first four L-moments are listed below:

$$\lambda_1 = \mu \qquad \lambda_2 = \sigma \qquad \lambda_3 = 0 \text{ and } \lambda_4 = \sigma/6.$$

The estimating equations of the L-estimators of the nuisance parameters σ and σ are given by

$$\hat{\mu} = \hat{\lambda}_1 \text{ and } \hat{\sigma} = \hat{\lambda}_2.$$

The elements of the Δ matrix:

$$\frac{\partial}{\partial \mu}\lambda_3 = \frac{\partial}{\partial \sigma}\lambda_3 = \frac{\partial}{\partial \mu}\lambda_4 = 0 \text{ and } \frac{\partial}{\partial \sigma}\lambda_4 = 1/6.$$

The C matrix is the 2×2 identity, because the nuisance parameter estimators are equal to the two first sample L-moments.

The Generalised Pareto Distribution

The generalised Pareto distribution (GPD) has quantile function

$$Q_0(u) = \mu + \frac{\sigma}{\kappa}\left[1 - (1-u)^{\kappa}\right].$$

We assume that the location parameter μ is known and we set $\mu = 0$.

The first four L-moments are listed below.

$$\begin{aligned}
\lambda_1 &= \frac{\sigma}{1+\kappa} \qquad \lambda_2 = \frac{\sigma}{(1+\kappa)(2+\kappa)} \\
\lambda_3 &= \frac{\sigma(1-\kappa)}{(1+\kappa)(2+\kappa)(3+\kappa)} \text{ and } \lambda_4 = \frac{\sigma(1-\kappa)(2-\kappa)}{(1+\kappa)(2+\kappa)(3+\kappa)(4+\kappa)}.
\end{aligned}$$

The estimation equations of the L-estimators of the nuisance parameters σ and κ are given by

$$\hat{\sigma} = (1+\hat{\kappa})\hat{\lambda}_1 \text{ and } \hat{\kappa} = \hat{\lambda}_1/\hat{\lambda}_2 - 2.$$

The elements of the Δ matrix:

$$\begin{aligned}
\frac{\partial}{\partial\sigma}\lambda_3 &= \frac{1-\kappa}{(1+\kappa)(2+\kappa)(3+\kappa)} \qquad \frac{\partial}{\partial\sigma}\lambda_4 = \frac{(1-\kappa)(2-\kappa)}{(1+\kappa)(2+\kappa)(3+\kappa)(4+\kappa)} \\
\frac{\partial}{\partial\kappa}\lambda_3 &= \sigma\frac{2\kappa^3+3\kappa^2-12\kappa-17}{\left[(1+\kappa)(2+\kappa)(3+\kappa)\right]^2} \text{ and } \frac{\partial}{\partial\kappa}\lambda_4 = -\sigma\frac{2\kappa^5+\kappa^4-52\kappa^3-95\kappa^2+92\kappa+172}{\left[(1+\kappa)(2+\kappa)(3+\kappa)(4+\kappa)\right]^2}.
\end{aligned}$$

The elements of the C matrix:

$$\begin{aligned}\frac{\partial}{\partial \lambda_1}\sigma &= 2\lambda_1/\lambda_2 - 1 \quad \frac{\partial}{\partial \lambda_1}\kappa = 1/\lambda_2 \\ \frac{\partial}{\partial \lambda_2}\sigma &= -\lambda_1^2/\lambda_2^2 \text{ and } \frac{\partial}{\partial \lambda_2}\kappa = -\lambda_1/\lambda_2^2.\end{aligned}$$

Supl. Mat. App. 2. The Data-Driven LGST and a Wasserstein-Type Test

Data-driven smooth goodness of fit tests were first proposed by Ledwina (1994) and later further developed by Kallenberg & Ledwina (1997) and Inglot et al. (1997), among others. The core of data-driven smooth tests exists in selecting the order k by using the observed data, and using the selected order as the order of the (data-driven) smooth test statistic. In this way the data-driven order k becomes a random variable. Hence, the null distribution of the data-driven test statistic depends on the data-driven order selection procedure. The order selection criterium should be chosen so as to increase the power of the data-driven test for a large class of alternatives.

We proceed along the lines of Claeskens & Hjort (2004) to extend the test to allow the order k to be selected by the data. More generally, a data-driven test allows the selection of even individual $\hat{\theta}_i$'s in the test statistic. First we need some extra notation. Let S denote an index set which indicates which $\hat{\theta}_j$'s are to be included in the LGST test statistic. For a given S, the LGST test statistic becomes $T_S = n\hat{\theta}_S^t \hat{\Sigma}_{\hat{\theta}_S}^{-1}\hat{\theta}_S$ with $\hat{\theta}_S$ a vector with elements $\hat{\theta}_i$ for which $i \in S$, and $\Sigma_{\hat{\theta}_S}$ is constructed from $\Sigma_{\hat{\theta}}$ by selecting the rows and columns specified by S. The component vector V_S is defined in a similar fashion. We restrict S so that $1, \ldots, p \notin S$ and $p+1 \leq \#S < \infty$, where $\#S$ denotes the number of elements in the set S. The results of the previous section immediately show that the asymptotic null distribution of T_S is chi-squared with $\#S - p$ degrees of freedom. A data-driven test allows the "model" S to be selected by the data. As in Claeskens & Hjort (2004) we restrict the process to select from a finite number of models. Let $\mathcal{M}$ denote the nonempty set of index sets S that may be selected. Let $C(S)$ denote a model selection criterion so that $C(S_1) > C(S_2)$ indicates that S_1 is a better model than S_2. Examples include the modified AIC and BIC criteria, defined as (Janic & Ledwina, 2009; Claeskens & Hjort, 2004)

$$AIC(S) = T_S - 2\#S \quad AIC_m(S) = V_S^t V_S - 2\#S$$

$$BIC(S) = T_S - \ln(n)\#S \text{ and } BIC_m(S) = V_S^t V_S - \ln(n)\#S.$$

The selection rule, which gives the data-driven choice of an index set to be used for testing, may now be formulated as

$$\hat{S} = \{R \in \mathcal{M} : R \neq \phi \text{ and } C(R) \geq C(Q), \forall Q \in \mathcal{M}\}. \tag{5}$$

If this results in more than one index set R with the same $C(R)$, the one with the smallest cardinality is selected. This description includes both order selection and subset selection. The next theorem is basically Theorem 4.8 of Thas (2010), based on section 3.1 of Claeskens & Hjort (2004). It requires the joint asymptotic null distribution of all T_S for $S \in \mathcal{M}$. Since these T_S statistics are quadratic forms of

$\hat{\theta}_S$, it is sufficient to know the limiting null distribution of $\hat{\theta}_R$ with R the smallest index set so that $R \cap S \neq \phi$ for all $S \in \mathcal{M}$. Theorem 2 gives this distribution.

Theorem 2 *Suppose that for all nonempty S with $p + 1 \leq \#S \leq \infty$, the asymptotic joint null distribution of the test statistics T_S ($S \in \mathcal{M}$) is known and stochastically represented by the vector $(\check{T}_S)_{S\in\mathcal{M}}$. Also suppose that $1 \leq \#\mathcal{M} < \infty$. Then, under assumptions (A1) up to (A4) of Theorem1, as $n \to \infty$, under H_0,*

$$T_{\hat{S}} \xrightarrow{d} \sum_{R\in\mathcal{M}} I\left(R = \tilde{S}\right) \check{T}_R, \tag{6}$$

where $\tilde{S}$ is given by (5) with $C(.)$ replaced by $\check{C}(.)$, which is the limiting null distribution of $C(.)$.

The asymptotic null distribution provided in (6) may be approximated by means of Monte Carlo simulations. In particular, by simulating from the $(\check{T}_S)_{S\in\mathcal{M}}$ distribution, the limiting distribution $(\check{C}(S))_{S\in\mathcal{M}}$ can be generated, and hence also the limiting distribution of (5). With these simulated statistics, the right hand side of (6) can be simulated, resulting in an approximation of the asymptotic null distribution of the data-driven test statistic.

Kallenberg & Ledwina (1997), Inglot et al. (1997) and Janic & Ledwina (2009), among others, constructed data-driven tests in a different way. They advocate an order selection procedure that asymptotically allows the selected order, say k_n, to grow with the sample size n. When this happens at an appropriate rate, their tests possess an omnibus consistency property. Their theory is more complicated and depends on the particular choice of the model selection criterion, which is typically of the BIC-type. However, since in practice the implementation of the test will always involve a horizon $\mathcal{M}$ with only a finite number of models, we prefer to work in the framework of Claeskens & Hjort (2004), which, moreover, allows for a wider range of model selection criteria. We refer to section 4.3 of Thas (2010) for a more detailed discussion on data-driven tests in a goodness of fit setting.

Finally we present a corollary with the limiting distribution of the Wasserstein-type test of Section 3.3.

Corollary 1 *Assume that the regularity conditions (A1) up to (A4) of Theorem 1 hold true. Let J denote a diagonal matrix with elements $(2j-1)$ ($j = p+1, \ldots, k$), and let Z and Z^* be independent random vectors with $(k-p)$ i.i.d. standard normal variates. If $k > p$ and the null hypothesis is true, then, as $n \to \infty$, W_k converges weakly to $Z^t \Sigma_{\hat{\theta}}^{1/2} J \Sigma_{\hat{\theta}}^{1/2} Z$. The limiting distribution may also be represented by $\sum_{j=p+1}^{k} \gamma_j Z^{*t} Z^*$, where the γ_j are the eigenvalues of $\Sigma_{\hat{\theta}}^{1/2} J \Sigma_{\hat{\theta}}^{1/2}$.*

The representation of the limiting distribution in Corollary 1 allows for a parametric bootstrap procedure with $\Sigma_{\hat{\theta}}$ replaced by a consistent estimator.

Table Suppl. Mat. Table 1: Abbreviations

AD	Anderson-Darling test
CvM	Cramér-von Mises test of Stephens (1979)
DD	data-driven LGST
ECF1	empirical characteristic function test (Epps, 2005; Meintanis, 2004) based on MME
ECF2	empirical characteristic function test (Epps, 2005; Meintanis, 2004) based on MLE
EMGF1	empirical moment generating function tests (Meintanis, 2004) based on MME
EMGF2	empirical moment generating function tests (Meintanis, 2004) based on MLE
GPD	generalised Pareto distribution
GST	generalised smooth test of Thas et al. (2009)
LGST	L-moments generalised smooth test
LV3	third component test of the LGST
LV4	fourth component test of the LGST
MLE	maximum likelihood estimator
MLME	method of L-moments estimator
MME	method of moments estimator
NS1	Neyman smooth test of Radouane & Crétois (2002) based on MME
NS2	Neyman smooth test of Radouane & Crétois (2002) based on MLE
V3	third component test of the GST
V4	fourth component test of the GST
Wk	Wasserstein-type test