Variations on Goodness-of-Fit Tests for the Generalized Pareto Distribution

Sean van der Merwe[a,*]

[a]University of the Free State, Box 339, Bloemfontein, 9300, South Africa;
vandermerwes@ufs.ac.za; +274013770
[*]Corresponding author

May 31, 2017

Abstract

We consider the problem of goodness-of-fit testing for whether a sample consists of independent Generalised Pareto observations above a known threshold. The shape of the density can vary markedly based on the value of the extreme value index (EVI). We comparing existing tests via simulation study. Then we investigate the effect of incorporating Bayesian parameter estimation methods and we note that different tests perform better for different EVI values. We use this knowledge to improve the existing tests through selective use of Bayesian parameter estimation methods and our simulations show that the methodology we propose is superior for the general problem.

Keywords: Bayes, Distribution, GPD, Hypothesis Testing, p-value, Simulation

1 Introduction

1.1 The Generalised Pareto Distribution (GPD)

The GPD is widely used to model excesses above a threshold. See Reiss et al. (2001) for applications across many fields. The use of the GPD is due in large part to the theorem of Pickands (1975), which says that if the threshold is chosen high enough the excesses should follow the GPD. In practice it is not easy to determine whether the conditions of the theorem are satisfied. Thus, it is worth testing whether a sample follows the GPD model, in the same way that one would test whether the residuals of a regression follow a Normal distribution.

1.2 The parametric bootstrap

Let $\mathbf{X}$ be an i.i.d. random sample of size n from an unknown distribution, and let $S = S(\mathbf{X}|m)$ be a test statistic for testing the null hypothesis that $\mathbf{X}$ follows a specific model (m), in this case the Generalized Pareto Distribution (GPD). In general, the test statistic S depends on the parameters $\boldsymbol{\theta}$ of the distribution to be tested, as is the case for most statistics based on the Empirical Distribution Function (see Darling, 1957 for a historical introduction). Thus, $S = S(\mathbf{X}|\boldsymbol{\theta}, m)$ is a function of $\boldsymbol{\theta}$ and S might be calculated as $S = S(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$, where $\hat{\boldsymbol{\theta}}$ is some estimate of $\boldsymbol{\theta}$.

The problem arises that, in the GPD case, the distribution of any currently used test statistic $S(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$ depends on the values of the unknown parameters $\boldsymbol{\theta}$, as well as the sample size. The distribution of the test statistic is also affected by the method of estimation of the unknown parameters. In order to obtain uniform p-values under the null hypothesis (sample is i.i.d. GPD with known threshold but unknown shape and scale) we need to estimate the distribution of the chosen test statistic accurately.

If the test statistic can be standardised in some way so that it is parameter invariant (its value and distribution do not depend on the parameters of the model) then simulating the distribution is trivial. In the case of the GPD, the extreme value index (EVI) parameter cannot be standardised and the distribution must be obtained by the method commonly referred to as the Parametric Bootstrap (Section 2.2). A good early reference to the theory is Beran (1988) who called it Pre-pivoting.

1.3 Existing tests for the GPD

Choulakian and Stephens (2001) said one should estimate the parameters using Maximum Likelihood (ML), then use the Cramér–von Mises statistic or the Anderson–Darling (AD) statistic. They apply the Parametric Bootstrap method to get a p-value. They gloss over any problems that might be induced by the ML estimation, such as that ML estimates don't always exist and don't work well for small samples (Zhang, 2010).

Meintanis and Bassiakos (2007) said one should use Probability Weighted Moments (PWM) or Method of Moments (MOM) when ML isn't the best. They constructed a new statistic that doesn't use the Empirical Distribution Function (EDF), which implies that one doesn't need to sort each sample and gives a small gain in speed.

They compared their test with the previous tests and showed that of all other statistics the Anderson-Darling (AD) statistic is the best, but theirs is slightly superior in the majority of situations tested. Our testing found it to be roughly equivalent to the AD statistic for most practical purposes (see simulations later in this paper).

We use the following form of the GPD throughout this paper. It is equivalent to the forms used in the goodness-of-fit literature of the GPD.

$$f(x|\xi,\sigma) = \frac{1}{\sigma}\left[1+\frac{\xi x}{\sigma}\right]^{-\frac{1}{\xi}-1}, \begin{cases} 0 < x < -\dfrac{\sigma}{\xi},\ \xi < 0 \\ \quad\quad x > 0,\ \xi > 0 \end{cases} \tag{1}$$

Consider Figure 1 and note the different forms of the GPD as the EVI crosses the boundaries of -1 and 0 especially. If the EVI is less than -1 the density is increasing. At -1 we have a Uniform density. Between -1 and 0 we see a bounded density which falls to a limiting value at an increasing rate. At 0 we have an Exponential density. Finally, if the EVI is positive we have the heavy tailed density most commonly associated with the GPD (heavier tails than the Exponential).

José A. Villaseñor-Alva and Elizabeth González-Estrada developed a new test for the GPD in 2009 based on the principle of testing negative and positive EVI cases separately and then combining the outcomes. They cite the book of Casella and Berger (1990) for the validity of this approach. This splitting idea has the advantage that if the test fails to reject we can immediately draw a conclusion on the sign of the EVI without further calculations.

For the positive test they estimated the parameters using Asymptotic Maximum Likelihood. Their sample size was not large — this was for speed and convenience, and we will show later that their power suffers greatly as a result. For the negative test they combine a Method of Moments (MM) step with a Maximum Likelihood (ML) step. This is a new approach and they tested it in the article because it's key to the success of their test. It performs well in our simulations.

For their statistic they used the correlation of the sorted observations with a transformation of the empirical distribution function (EDF). They had different transformations for different parameter values.

It is worth noting that they did not directly compare their test to previous tests themselves, and this comparison forms the starting point for this paper.

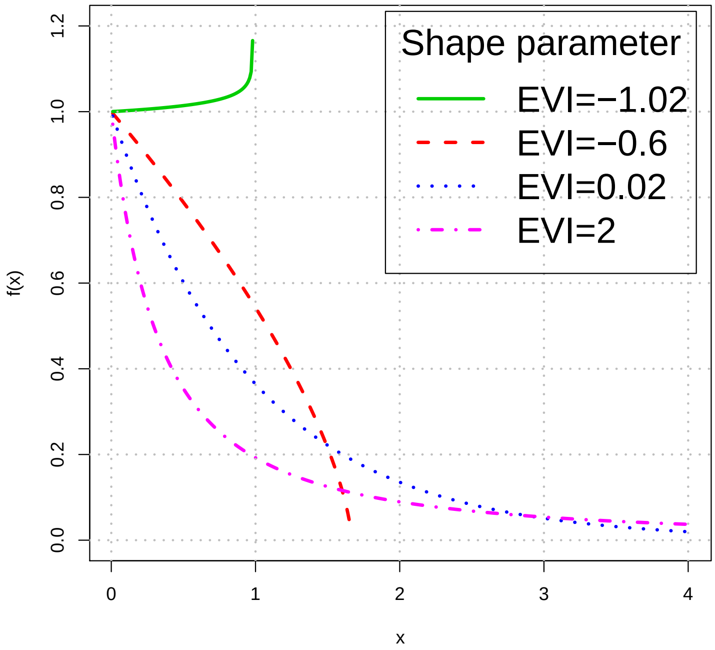

Figure 1: Illustration of various forms of the Generalized Pareto Distribution.

1.4 Objectives and outline of the present paper

We begin by directly comparing the test of Villaseñor-Alva and González-Estrada (2009) to the tests published prior to that, for the purpose of deducing the strengths and weaknesses of each (Section 2). We then investigate various modifications of the existing tests, ultimately arriving at new proposed methodology which appears to simultaneously capture the strengths of existing tests and reduce their weaknesses (Section 3). We end with a conclusion (Section 4).

2 Initial Experiments

2.1 Implementation of existing tests

At this stage we assume that the threshold is a known value. The threshold (μ) can then be subtracted from every observation, resulting in an otherwise identical GPD but with a fixed threshold of zero. The scale parameter (σ) does not affect the performance of any of the tests considered, but cannot be eliminated. It is a minor nuisance parameter.

The EVI (ξ) is the shape parameter and the key parameter to estimate when fitting a GPD. In general, better estimation of the EVI results in a test with greater power. Unfortunately, there is no estimator that is universally better than all other estimators, and so we see that each existing test performs better in specific situations.

That said, the EVI is not explicitly part of the hypothesis being tested and the effect of its estimation (along with σ) must be accommodated. This is done using the parametric bootstrap method. We define $\boldsymbol{\theta}$ as the vector $(\xi,\ \sigma)$ throughout.

2.2 The parametric bootstrap method

The parametric bootstrap is implemented as follows:

1. Obtain base parameter estimates $\hat{\boldsymbol{\theta}}$.
2. Calculate the base statistic $S(\mathbf{x}|\hat{\boldsymbol{\theta}}(\mathbf{x}), m)$.
3. Draw N new samples $\mathbf{x}_i,\ i = 1, \ldots, N$ from the model $f(\mathbf{X}|\hat{\boldsymbol{\theta}}, m)$ given the parameter estimates from Step 1.
4. Calculate N new statistics $S(\mathbf{x}_i|\hat{\boldsymbol{\theta}}(\mathbf{x}_i), m)$ corresponding to each new sample drawn in Step 3. The parameter estimation procedure must be repeated for each new sample.
5. Calculate the proportion of the new statistics (from Step 4) that exceed the base statistic (from Step 2) and report this result as a p-value.

2.3 Test statistics

Given an i.i.d. sample $\mathbf{x} = (x_1, \ldots, x_n)$ with CDF $F_X(\boldsymbol{\theta})$, and corresponding order statistics $x_{(1)}, \ldots, x_{(n)}$, the AD statistic A^2 is defined as:

$$A^2 = s(\mathbf{x}|\boldsymbol{\theta}) = -n - \sum_{k=1}^{n} \frac{2k-1}{n}[\log F_X(x_{(k)}|\boldsymbol{\theta}) + \log(1 - F_X(x_{(n+1-k)}|\boldsymbol{\theta}))] \tag{2}$$

Meintanis and Bassiakos (2007) calculate their statistic in 2 steps:

1. Let $y_j = \hat{\xi}^{-1} \log\left(1 + \frac{\hat{\xi} x_j}{\hat{\sigma}}\right)$.
2. Given a hyperparameter a that you choose (say 0.25),

$$T_{n,a} = \frac{1}{n} \sum_{j,k=1}^{n} \frac{1 + (y_j + y_k + a + 1)^2}{(y_j + y_k + a)^3} - 2 \sum_{j=1}^{n} \frac{1 + y_j + a}{(y_j + a)^2}$$

We use ‘Tn’ to denote the above statistic and ‘Rr’ to denote the correlation statistic used by Villaseñor-Alva and González-Estrada (2009).

The Rr statistic is the sample correlation between $\mathbf{X}^*$ and $\mathbf{Y}^*$, which are defined as follows:

$$X_i^* = \begin{cases} X_i \ , \ \hat{\xi} < 0.5 \\ \log(X_i) \ , \ \hat{\xi} \geq 0.5 \end{cases} \quad \text{where } X_i \text{ denotes the } i^{th} \text{ sorted observation.}$$

$$Y_i^* = \begin{cases} \left(\bar{F}_n(X_i^*)\right)^{-\hat{\xi}} \ , \ \hat{\xi} < 0.5 \\ \log\left[\left(\bar{F}_n(X_i^*)\right)^{-\hat{\xi}} - 1\right] \ , \ \hat{\xi} \geq 0.5 \end{cases} \quad \text{where } \bar{F}_n \text{ denotes the empirical survival function.}$$

2.4 Type 1 error and power of proposed tests

We calculate, through simulation, the type 1 error and power of the proposed tests over 5000 samples from each of various distribution scenarios. We use 491 replicates for the parametric bootstrap in all cases. The software used is R (R Core Team, 2013), with add-on packages ‘evir’ (Pfaff and McNeil, 2012) and ‘gPdtest’ (Estrada and Alva, 2012). The built-in ‘parallel’ package was used extensively to perform the simulations in parallel over multiple servers.

Test Procedure	GPD(-1.4)	GPD(-1.1)	GPD(-0.6)	GPD(-0.2)	GPD(0.2)	GPD(0.6)	GPD(1.2)
1. Villaseñor-Alva	0.032	0.025	0.013	0.008	0.01	0.02	0.027
2. ADML	0.032	0.034	0.037	0.024	0.03	0.038	0.042
3. TnML	0.046	0.052	0.054	0.043	0.038	0.044	0.04
4. RrML	0.023	0.04	0.038	0.051	0.067	0.055	0.042
5. ADPWM	0.042	0.036	0.046	0.065	0.071	0.048	0.049
6. ADZhang	0.144	0.092	0.063	0.056	0.051	0.05	0.048
7. TnZhang	0.088	0.07	0.061	0.052	0.052	0.05	0.047
8. RrZhang	0.127	0.081	0.036	0.03	0.058	0.051	0.047
9. MixZhang	0.018	0.027	0.065	0.057	0.051	0.049	0.047
10. ADvdMerwe	0.03	0.025	0.024	0.017	0.028	0.05	0.039
11. TnvdMerwe	0.006	0.006	0.006	0.01	0.033	0.049	0.043
12. RrvdMerwe	0.035	0.03	0.012	0.004	0.012	0.015	0.027
13. MixvdMerwe	0.039	0.042	0.037	0.021	0.013	0.027	0.027

Table 1: Rejection rates for sample size 25 under the null. 5000 samples for each distribution. Closer to 0.05 is better.

Test Procedure	GPD(-1.4)	GPD(-1.1)	GPD(-0.6)	GPD(-0.2)	GPD(0.2)	GPD(0.6)	GPD(1.2)
1. Villaseñor-Alva	0.047	0.046	0.041	0.023	0.027	0.033	0.044
2. ADML	0.034	0.043	0.028	0.038	0.045	0.049	0.047
3. TnML	0.049	0.048	0.045	0.045	0.046	0.05	0.05
4. RrML	0.033	0.049	0.043	0.056	0.059	0.039	0.047
5. ADPWM	0.048	0.044	0.049	0.07	0.07	0.035	0.058
6. ADZhang	0.112	0.08	0.06	0.048	0.054	0.051	0.052
7. TnZhang	0.072	0.067	0.057	0.051	0.05	0.055	0.053
8. RrZhang	0.093	0.074	0.047	0.028	0.048	0.043	0.047
9. MixZhang	0.02	0.045	0.062	0.048	0.052	0.049	0.051
10. ADvdMerwe	0.045	0.048	0.036	0.015	0.034	0.049	0.048
11. TnvdMerwe	0.015	0.011	0.008	0.01	0.037	0.048	0.049
12. RrvdMerwe	0.046	0.047	0.028	0.003	0.02	0.018	0.043
13. MixvdMerwe	0.046	0.047	0.052	0.039	0.024	0.033	0.043

Table 2: Rejection rates for sample size 50 under the null. 5000 samples for each distribution. Closer to 0.05 is better.

2.4.1 Type 1 error

First we consider various EVI values in the GPD case (null hypothesis is true) and determine to what extent each test works, as expected from a classical hypothesis test. Specifically, we expect that if a significance level is chosen as α (say) then the test will falsely reject the null hypothesis proportion α of the time. This is equivalent to the p-value being uniformly distributed. We choose $\alpha = 0.05$ throughout. The results for $\alpha = 0.10$ follow the same pattern very closely.

A rejection rate above α is generally considered worse than the same discrepancy below α but both types of discrepancy can lead to poor decisions. A low rejection rate can result in false confidence in the null, while a rejection rate above α results in excessive Type 1 errors. We use the term ‘accuracy’ to refer to how close a test procedure comes to achieving the desired significance level.

At this stage let us only consider the first three rows of Table 1, Table 2 and Table 3. The three tables correspond to sample sizes of 25, 50 and 100 observations above the threshold, which we consider to be small, medium and large sample sizes.

We note that the test of Villaseñor-Alva and González-Estrada (2009) has low rejection rates for sample size 25 and high rejection rates for sample size 100. The AD ML approach suggested by Choulakian and Stephens (2001) appears adequate, but inferior to the approach of Meintanis and Bassiakos (2007) across all forms of GPD and sample sizes considered.

Test Procedure	GPD(-1.4)	GPD(-1.1)	GPD(-0.6)	GPD(-0.2)	GPD(0.2)	GPD(0.6)	GPD(1.2)
1. Villaseñor-Alva	0.053	0.042	0.053	0.109	0.104	0.053	0.05
2. ADML	0.07	0.055	0.04	0.049	0.046	0.046	0.051
3. TnML	0.045	0.049	0.048	0.045	0.048	0.043	0.052
4. RrML	0.058	0.048	0.046	0.053	0.059	0.034	0.049
5. ADPWM	0.047	0.043	0.045	0.069	0.072	0.028	0.074
6. ADZhang	0.092	0.063	0.055	0.053	0.047	0.048	0.051
7. TnZhang	0.07	0.056	0.049	0.051	0.048	0.045	0.055
8. RrZhang	0.086	0.056	0.052	0.036	0.046	0.034	0.05
9. MixZhang	0.029	0.038	0.056	0.052	0.047	0.047	0.053
10. ADvdMerwe	0.045	0.046	0.054	0.012	0.039	0.04	0.052
11. TnvdMerwe	0.043	0.019	0.009	0.005	0.036	0.041	0.05
12. RrvdMerwe	0.054	0.044	0.048	0.005	0.031	0.023	0.049
13. MixvdMerwe	0.054	0.044	0.052	0.055	0.033	0.036	0.052

Table 3: Rejection rates for sample size 100 under the null. 5000 samples for each distribution. Closer to 0.05 is better.

Test Procedure	Beta(2,1)	LN(0,1)	Weibull(3,1)	Gamma(5,1)	Gamma(8,1)	Chisq(6)
1. Villaseñor-Alva	0.162	0.004	0.195	0.061	0.121	0.025
2. ADML	0.017	0.12	NA	0.543	0.812	0.181
3. TnML	0.039	0.123	NA	0.384	0.434	0.212
4. RrML	0.024	0.194	NA	0.73	0.862	0.488
5. ADPWM	0.311	0.246	0.87	0.947	0.991	0.744
6. ADZhang	0.848	0.231	0.983	0.98	0.999	0.746
7. TnZhang	0.759	0.345	0.987	0.995	1	0.852
8. RrZhang	0.197	0.155	0.239	0.321	0.359	0.245
9. MixZhang	0.174	0.232	0.406	0.754	0.6	0.724
10. ADvdMerwe	0.366	0.077	0.898	0.896	0.985	0.492
11. TnvdMerwe	0.243	0.002	0.761	0.788	0.976	0.247
12. RrvdMerwe	0.181	0.03	0.173	0.038	0.094	0.013
13. MixvdMerwe	0.188	0.086	0.42	0.442	0.622	0.285

Table 4: Rejection rates for sample size 25 under alternatives. 5000 samples for each distribution. Closer to 1 is better

2.4.2 Power

Second we consider data from alternative distributions. We would like to see that each test will correctly reject the null hypothesis more often than proportion α (significance level) in all cases. Ideally, a test should reject as close as possible to 100% of cases where the null hypothesis does not hold true. We refer to the rejection rate under alternative distributions as the 'power' of a test procedure.

When comparing the rejection rates of tests, it is important to bare in mind the effect of failing to achieve the correct significance level on power comparisons. A test can spuriously appear to have higher power simply because it just rejects more often in all cases, which is rarely useful in practice.

Considering the first three rows of Table 4, Table 5 and Table 6, we note that the ML based approach appears to be superior to the approach of Villaseñor-Alva and González-Estrada (2009) in all cases except the Weibull(3,1) alternative, where the ML method consistently fails to obtain valid parameter estimates.

2.5 Expanding on initial comparison

The investigation can be expanded by considering the effect of the correlation statistic (Rr) on its own. This is given in row 4 of all six tables. If we directly compare the statistic to the competing statistics under ML estimation we note that the rejection rates often differ greatly from α under the null, and the power is markedly lower under most alternatives except for small sample sizes. One possible explanation is that the Rr statistic does not make use of σ from the parameter estimation method, causing it to over-fit the transformed empirical distribution function.

Test Procedure	Beta(2,1)	LN(0,1)	Weibull(3,1)	Gamma(5,1)	Gamma(8,1)	Chisq(6)
1. Villaseñor-Alva	0.358	0.008	0.668	0.214	0.363	0.1
2. ADML	0.339	0.432	NA	0.966	0.992	0.787
3. TnML	0.203	0.542	NA	0.959	0.978	0.82
4. RrML	0.074	0.215	NA	0.839	0.954	0.698
5. ADPWM	0.638	0.329	0.998	1	1	0.973
6. ADZhang	0.968	0.53	1	1	1	0.973
7. TnZhang	0.957	0.707	1	1	1	0.993
8. RrZhang	0.277	0.181	0.574	0.672	0.767	0.533
9. MixZhang	0.357	0.531	0.907	0.994	0.973	0.973
10. ADvdMerwe	0.82	0.212	1	1	1	0.907
11. TnvdMerwe	0.706	0.009	0.999	1	1	0.911
12. RrvdMerwe	0.359	0.095	0.626	0.118	0.273	0.03
13. MixvdMerwe	0.359	0.273	0.756	0.712	0.873	0.484

Table 5: Rejection rates for sample size 50 under alternatives. 5000 samples for each distribution. Closer to 1 is better

Test Procedure	Beta(2,1)	LN(0,1)	Weibull(3,1)	Gamma(5,1)	Gamma(8,1)	Chisq(6)
1. Villaseñor-Alva	0.661	0.06	0.946	0.334	0.42	0.224
2. ADML	0.63	0.839	NA	1	1	0.998
3. TnML	0.739	0.931	NA	1	1	0.974
4. RrML	0.304	0.223	NA	0.958	0.991	0.872
5. ADPWM	0.938	0.406	1	1	1	1
6. ADZhang	0.999	0.878	1	1	1	1
7. TnZhang	1	0.963	1	1	1	1
8. RrZhang	0.479	0.191	0.927	0.949	0.984	0.846
9. MixZhang	0.671	0.877	0.999	1	1	1
10. ADvdMerwe	0.994	0.52	1	1	1	0.999
11. TnvdMerwe	0.98	0.036	1	1	1	1
12. RrvdMerwe	0.659	0.166	0.971	0.358	0.64	0.091
13. MixvdMerwe	0.659	0.654	0.976	0.904	0.978	0.676

Table 6: Rejection rates for sample size 100 under alternatives. 5000 samples for each distribution. Closer to 1 is better

It is also worth looking at using the PWM estimator instead of ML. Directly comparing rows 5 and 2 of all the tables we note that the PWM method appears to reject too often when the EVI is near zero (common in practice). It does, however, have higher power under all the alternatives considered except the Log-Normal. This is a clear indication that it may be possible to achieve higher power through alternative parameter estimation, which will be the focus going forward.

Villaseñor-Alva and González-Estrada (2009) used the Hill estimator for positive ξ, so as a first attempt to improve on their test procedure we replaced the Hill estimator with a bias-reduced Hill estimator (Gomes et al., 2016). This yielded little to no improvement. The difference in results were negligible across all scenarios and is not worth reporting. This led to the testing of more aggressive alterations to the test procedures.

3 New suggested methodology

3.1 Bayesian parameter estimation

In theory, under the Bayesian paradigm, it is possible to arrive at more accurate parameter estimates for most models through the posterior distribution, given a specific loss function. For example, under squared error loss the posterior mean is the optimal estimator in this framework. In practice, however, the choice of prior distribution may affect the practical performance of a test based on such a Bayesian parameter estimation method, especially in small samples. We thus investigate the impact of specific Bayesian approaches by directly comparing them to the previous approaches in a simulation study.

3.2 New Test 1

To begin the new approach, we keep the classical parametric bootstrap framework; but replace the parameter estimation with the Empirical Bayes estimation of Zhang (2010).

In all the tables, rows 6 to 8 can be compared to rows 2 to 4 respectively to assess the impact of this replacement. Under the null hypothesis (first three tables) there is a slight improvement in accuracy in most cases, except where $\xi < -0.5$. In that situation we note that the test rejects too often. For the alternatives (last three tables) we note a dramatic improvement in power when using the AD and Tn statistics, but less so for the Rr statistic. It is not yet clear, however, what proportion of the improvement is genuine, as opposed to the test merely rejecting more often in general.

In order to address the concerns of excessive rejection under the null when $\xi < -0.5$, let us consider a mixed test procedure. The new procedure is to estimate the parameters using the code of Zhang (2010), then check whether they lie in the problem area. If the EVI estimate is small enough, the parameters are replaced with the estimates of Villaseñor-Alva and González-Estrada (2009) and the test proceeds using the Rr statistic. The same check is done for the replicate samples in the parametric bootstrap procedure.

The results of the suggested procedure are given in row 9 of the tables. While the excessive rejection under the null is reduced, the power suffers, especially in small samples.

In conclusion, replacing the parameter estimation with the Empirical Bayes estimation of Zhang (2010) and use of the Tn statistic is recommended whenever one has prior knowledge restricting the parameter space to cases where $\xi > -0.5$. It should be noted that almost all practical applications of the GPD do meet this requirement.

3.3 New Test 2

The second new approach considered is based on the split approach of Villaseñor-Alva and González-Estrada (2009). The split approach involves performing three hypothesis tests. The first test is H_0 :

$\mathbf{X}\ GPD \mid \xi < 0$. The second test is H_0 : $\mathbf{X}\ GPD \mid \xi > 0$. The third test is the combination of these (covering the entire parameter space) and is performed by taking the maximum of the two p-values produced by the first two tests.

For the first test we use the parameter estimation introduced in Villaseñor-Alva and González-Estrada (2009) and apply all three test statistics. For the second test we simulate from the conditional posterior distribution of the parameters and take the posterior mean. A detailed explanation is given below.

3.4 Objective prior distribution

The maximal data information (MDI) prior (Zellner, 1997, pp. 112–116) is used as an objective prior. To quote Zellner, the MDI prior provides "maximal prior average data information relative to the information in the prior distribution". The MDI prior is defined as $\exp\{E[\log f(x)]\}$, which works out to $\frac{e^{-\xi}}{\sigma}$ for the GPD.

Alternative priors such as the Jeffreys prior (Jeffreys, 1998) can be used, but one must be careful of additional restrictions placed on the parameters. The test procedure may malfunction or fail when the parameters estimated from the sample fall inside or near the restricted area of their domain. The MDI prior does not create such restrictions, which further motivates its use.

3.5 Posterior distribution

We will simulate from the posterior using the Metropolis algorithm applied using a bivariate Normal proposal density on the log scale of the parameters. See Robert and Casella (2004, pp. 267–301) for an in-depth general discussion of this algorithm. By working on the log scale the log parameters can vary freely, while the parameters produced on the original scale are naturally limited to positive values, as desired, with no loss of accuracy near the boundary. We used 491 posterior simulations following a short burn-in period in each case.

3.6 Discussion

Rows 10, 11 and 12 of all tables show the new test procedure using the AD, Tn and Rr statistics respectively, with the same statistic used on both sides of the test. Row 13 is the result of using the Rr statistic on the negative side and the AD statistic on the positive side.

Looking at the first three tables (the null) we note that none of these tests seriously over-reject under any scenario. They are the only tests besides the TnML test of Meintanis and Bassiakos (2007) to meet this important criterion. The tests using the Tn and Rr statistics do under-reject rather badly in some cases, which relates to them having lower power under the alternative, and false confidence in the null.

Looking at the last three tables (the alternative) we note that the AD statistic produces the highest power in most cases, and higher power than the TnML test. It is thus the test procedure with the highest power of all tests that reliably have observed significance levels below the chosen α.

4 Conclusion

It is clear from the results of the simulations that the second new test constructed in this paper (row 10 in the tables) performs very well and helps address the problem of testing for the GPD with known threshold autonomously. The new test is based on a split approach where two conditional tests are performed ($\xi < 0$ and $\xi > 0$) and the maximum of the two p-values is reported. We suggest using the AD statistic throughout. For the negative side test we used the MMML parameter estimation and for the positive side test we used the mean of the posterior simulations.

If the researcher is convinced that the $\xi > -0.5$ for the sample being considered, then the superior approach appears to be the Empirical Bayes parameter estimation combined with the Tn statistic in the standard parametric bootstrap framework (row 6 in the tables).

All tests considered in this paper do still assumed a known threshold. If the threshold is chosen, especially if the threshold is picked so that the sample is most GPD-like, then the 'Do not reject' side of the test will not work as expected. On the other hand, a rejection is then an even stronger result.

References

Beran, Rudolf (1988), "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements", *Journal of the American Statistical Association* 83.403, pp. 687–697, URL: `http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1988.10478649`.

Casella, George and Berger, Roger L (1990), *Statistical Inference.*

Choulakian, V and Stephens, MA (2001), "Goodness-of-fit tests for the generalized Pareto distribution", *Technometrics* 43.4, pp. 478–484.

Darling, D. A. (1957), "The Kolmogorov-Smirnov, Cramer-von Mises Tests", English, *The Annals of Mathematical Statistics* 28.4, pp. 823–838, ISSN: 00034851, URL: `http://www.jstor.org/stable/2237048`.

Estrada, Elizabeth Gonzalez and Alva, Jose A. Villasenor (2012), *gPdtest: Bootstrap goodness-of-fit test for the generalized Pareto distribution*, R package version 0.4, URL: `http://CRAN.R-project.org/package=gPdtest`.

Gomes, M Ivette, Brilhante, M Fátima, and Pestana, Dinis (2016), "New reduced-bias estimators of a positive extreme value index", *Communications in Statistics-Simulation and Computation* 45, pp. 1–30.

Jeffreys, H. (1998), *The Theory of Probability*, 3rd ed., OUP Oxford, ISBN: 9780191589676, URL: `http://books.google.co.za/books?id=vh9Act9rtzQC`.

Meintanis, Simos G and Bassiakos, Yiannis (2007), "Data-transformation and test of fit for the generalized Pareto Hypothesis", *Communications in Statistics—Theory and Methods* 36.4, pp. 833–849.

Pfaff, Bernhard and McNeil, Alexander (2012), *evir: Extreme Values in R*, R package version 1.7-3, URL: `http://CRAN.R-project.org/package=evir`.

Pickands, James (1975), "Statistical inference using extreme order statistics", *the Annals of Statistics*, pp. 119–131.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL: `http://www.R-project.org/`.

Reiss, Rolf-Dieter, Thomas, Michael, and Reiss, RD (2001), *Statistical analysis of extreme values*, Springer.

Robert, Christian and Casella, George (2004), *Monte Carlo Statistical Methods*, 2nd ed., Springer, ISBN: 978-0387212395.

Villaseñor-Alva, José A and González-Estrada, Elizabeth (2009), "A bootstrap goodness of fit test for the Generalized Pareto Distribution", *Computational Statistics & Data Analysis* 53.11, pp. 3835–3841.

Zellner, A. (1997), *Bayesian Analysis in Econometrics and Statistics: The Zellner View and Papers*, Economists of the Twentieth Century Series, Edward Elgar Pub, ISBN: 9781858982205, URL: `http://books.google.co.za/books?id=ICW7AAAAIAAJ`.

Zhang, Jin (2010), "Improving on Estimation for the Generalized Pareto Distribution", *Technometrics* 52.3, pp. 335–339, URL: `http://dx.doi.org/10.1198/TECH.2010.09206`.