Stat Comput (2017) 27:1435–1451
DOI 10.1007/s11222-016-9697-3

Automated selection of *r* for the *r* largest order statistics approach with adjustment for sequential testing

Brian Bader[1] · **Jun Yan**[1] · **Xuebin Zhang**[2]

Received: 5 February 2016 / Accepted: 17 August 2016 / Published online: 5 October 2016

Abstract The r largest order statistics approach is widely used in extreme value analysis because it may use more information from the data than just the block maxima. In practice, the choice of r is critical. If r is too large, bias can occur; if too small, the variance of the estimator can be high. The limiting distribution of the r largest order statistics, denoted by GEV_r, extends that of the block maxima. Two specification tests are proposed to select r sequentially. The first is a score test for the GEV_r distribution. Due to the special characteristics of the GEV_r distribution, the classical chi-square asymptotics cannot be used. The simplest approach is to use the parametric bootstrap, which is straightforward to implement but computationally expensive. An alternative fast weighted bootstrap or multiplier procedure is developed for computational efficiency. The second test uses the difference in estimated entropy between the GEV_r and GEV_{r-1} models, applied to the r largest order statistics and the $r - 1$ largest order statistics, respectively. The asymptotic distribution of the difference statistic is derived. In a large scale simulation study, both tests held their size and had substantial power to detect various misspecification schemes. A new approach to address the issue of multiple, sequential hypotheses testing is adapted to this setting to control the false discovery rate or familywise error rate. The utility of the procedures is demonstrated with extreme sea level and precipitation data.

✉ Jun Yan
jun.yan@uconn.edu

Brian Bader
brian.bader@uconn.edu

Xuebin Zhang
xuebin.zhang@canada.ca

[1] University of Connecticut, Storrs, CT, USA

[2] Environment and Climate Change Canada, Toronto, ON, Canada

Keywords Entropy · Generalized extreme value · Goodness-of-fit · Multiplier bootstrap · Score test · Sequential testing

1 Introduction

The r largest order statistics is an extension of the block maxima approach that is often used in extreme value modeling. The focus of this paper is (Smith 1986, pp. 28–29): "Suppose we are given, not just the maximum value for each year, but the largest ten (say) values. How might we use this data to obtain better estimates than could be made just with annual maxima?" The r largest order statistics approach may use more information than just the block maxima in extreme value analysis, and is widely used in practice when such data are available for each block. The approach is based on the limiting distribution of the r largest order statistics which extends the generalized extreme value (GEV) distribution (e.g., Weissman 1978). This distribution, denoted by GEV_r, has the same parameters as the GEV distribution, which makes it useful to estimate the GEV parameters when the r largest values are available for each block. The approach was investigated by Smith (1986) for the limiting joint Gumbel distribution and extended to the more general limiting joint GEV_r distribution by Tawn (1988). Because of the potential gain in efficiency relative to the block maxima only, the method has found many applications in areas such as corrosion engineering (e.g., Scarf and Laycock 1996), hydrology (e.g., Dupuis 1997), coastal engineering (e.g., Guedes Soares and Scotto 2004), and wind engineering (e.g., An and Pandey 2007).

In practice, the choice of r is a critical issue in extreme value analysis with the r largest order statistics approach. In general r needs to be small relative to the block size B (not the number of blocks n) because as r increases, the rate of convergence to the limiting joint distribution decreases sharply (Smith 1986). There is a trade-off between the validity of the limiting result and the amount of information required for good estimation. If r is too large, bias can occur; if too small, the variance of the estimator can be high. Finding the optimal r should lead to more efficient estimates of the GEV parameters without introducing bias. A much related but different problem is the selection of threshold or fraction of a sample extreme value analysis (see Scarrott and MacDonald 2012 for a review). Our focus here is the selection of r for situations where a number of largest values are available each of n blocks. In contrast, the methods for threshold or fraction selection reviewed in Scarrott and MacDonald (2012) deal with a single block ($n = 1$) of a large size B.

The selection of r has not been as actively researched as the threshold selection problem in the one sample case. Smith (1986) and Tawn (1988) used probability (also known as PP) plots for the marginal distribution of the rth order statistic to assess its goodness-of-fit. The probability plot provides a visual diagnosis, but different viewers may reach different conclusions in the absence of a p value. Further, the probability plot is only checking the marginal distribution for a specific r as opposed to the joint distribution. Tawn (1988) suggested an alternative test of fit using a spacings results in Weissman (1978). Let $D_{n:i}$ be the spacing between the ith and $(i + 1)$th largest value in a sample of size B from a distribution in the domain of attraction of the Gumbel distribution. Then $\{iD_i : i = 1, \ldots, r - 1\}$ is approximately a set of independent and identically distributed exponential random variables as $B \to \infty$. The connections among the three limiting forms of the GEV distribution (e.g., Embrechts et al. 1997, p. 123) can be used to transform from the Fréchet and the Weibull distribution to the Gumbel distribution. Testing the exponentiality of the spacings on the Gumbel scale provides an approximate diagnosis of the joint distribution of the r largest order statistics when B is large. A limitation of this method, however, is that prior knowledge of the domain of attraction of the distribution is needed. Lastly, Dupuis (1997) proposed a robust estimation method, where the weights can be used to detect inconsistencies with the GEV$_r$ distribution and assess the fit of the data to the joint Gumbel model. The method can be extended to general GEV$_r$ distributions but the construction of the estimating equations is computing intensive with Monte Carlo integrations.

In this paper, two specification tests are proposed to select r through a sequence of hypothesis testing. The first is the score test (e.g., Rao 2005), but because of the nonstandard setting of the GEV$_r$ distribution the usual χ^2 asymptotic distribution is invalid. A parametric bootstrap can be used to assess the significance of the observed statistic, but is computationally demanding. A fast, large sample alternative to parametric bootstrap based on the multiplier approach (Kojadinovic and Yan 2012) is developed. The second test uses the difference in estimated entropy between the GEV$_r$ and GEV$_{r-1}$ models, applied to the r largest order statistics and the $r - 1$ largest order statistics, respectively. The asymptotic distribution is derived with the central limit theorem. Both tests are intuitive to understand, easy to implement, and have substantial power as shown in the simulation studies. Each of the two tests is carried out to test the adequacy of the GEV$_r$ model for a sequence of r values. The very recently developed stopping rules for ordered hypotheses in G'Sell et al. (2015) are adapted to control the false discovery rate (FDR), the expected proportion of incorrectly rejected null hypotheses among all rejections, or familywise error rate (FWER), the probability of at least one type I error in the whole family of tests. All the methods are available in the R package `eva` (Bader and Yan 2015.)

The rest of the article is organized as follows. The problem is set up in Sect. 2 with the GEV$_r$ distribution, observed data, and the hypothesis to be tested. The score test is proposed in Sect. 3 with two implementations: parametric bootstrap and multiplier bootstrap. The entropy difference (ED) test is proposed and the asymptotic distribution of the testing statistic is derived in Sect. 4. A large scale simulation study on the empirical size and power of the tests are reported in Sect. 5. In Sect. 6, the multiple, sequential testing problem is addressed by adapting recent developments on this application. The tests are applied to sea level and precipitation datasets in Sect. 7. A discussion concludes in Sect. 8. The Appendix contains the details of random number generation from the GEV$_r$ distribution and a sketch of the proof of the asymptotic distribution of the ED test statistic.

2 Model and data setup

The limit joint distribution of the r largest order statistics of a random sample of size B as $B \to \infty$ is the GEV$_r$ distribution with density function (Weissman 1978)

$$f_r(x_1, x_2, \ldots, x_r|\mu, \sigma, \xi) = \sigma^{-r} \exp\left\{ -(1+\xi z_r)^{-\frac{1}{\xi}} \right. \\ \left. - \left(\frac{1}{\xi}+1\right)\sum_{j=1}^{r}\log(1+\xi z_j)\right\} \tag{1}$$

for some location parameter μ, scale parameter $\sigma > 0$ and shape parameter ξ, where $x_1 > \cdots > x_r$, $z_j = (x_j - \mu)/\sigma$, and $1 + \xi z_j > 0$ for $j = 1, \ldots, r$. When $r = 1$, this distribution is exactly the GEV distribution. The parameters $\theta = (\mu, \sigma, \xi)^\top$ remain the same for $j = 1, \ldots, r$, $r \ll B$, but the convergence rate to the limit distribution reduces sharply as r increases. The conditional distribution of the rth component given the top $r - 1$ variables in (1) is the GEV distribution right truncated by x_{r-1}, which facilitates simulation from the GEV$_r$ distribution; see Appendix 1.

The r largest order statistics approach is an extension of the block maxima approach in extreme value analysis when a number of largest order statistics are available for each one of a collection of independent blocks (Smith 1986; Tawn 1988). Specifically, let $(x_{i1}, \dots, x_{ir})$ be the observed r largest order statistics from block i for $i = 1, \dots, n$. Assuming independence across blocks, the GEV$_r$ distribution is used in place of the GEV distribution in the block maxima approach to make likelihood-based inference about θ. Let $l_i^{(r)}(\theta) = l^{(r)}(x_{i1}, \dots, x_{ir}|\theta)$, where

$$l^{(r)}(x_1, \dots, x_r|\theta) = -r \log \sigma - (1 + \xi z_r)^{-\frac{1}{\xi}} - \left(\frac{1}{\xi} + 1\right) \sum_{j=1}^{r} \log(1 + \xi z_j) \tag{2}$$

is the contribution to the log-likelihood from a single block $(x_1, \dots, x_r)$. The maximum likelihood estimator (MLE) of θ using the r largest order statistics is $\hat{\theta}_n^{(r)} = \arg\max \sum_{i=1}^{n} l_i^{(r)}(\theta)$.

Model checking is a necessary part of statistical analysis. The rationale of choosing a larger value of r is to use as much information as possible, but not set r too high so that the GEV$_r$ approximation becomes poor due to the decrease in convergence rate. Therefore, it is critical to test the goodness-of-fit of the GEV$_r$ distribution with a sequence of null hypotheses

$H_0^{(r)}$: the GEV$_r$ distribution fits the sample of the r largest order statistics well

for $r = 1, \dots, R$, where R is the maximum, predetermined number of top order statistics to test. Two test procedures for $H_0^{(r)}$ are developed for a fixed r first to help choose $r \geq 1$ such that the GEV$_r$ model still adequately describes the data. The sequential testing process and the multiple testing issue are investigated in Sect. 6.

3 Score test

A score statistic for testing the goodness-of-fit hypothesis $H_0^{(r)}$ is constructed in the usual way with the score function and the Fisher information matrix (e.g., Rao 2005). For ease of notation, the superscript (r) is dropped. Define the score function

$$S(\theta) = \sum_{i=1}^{n} S_i(\theta) = \sum_{i=1}^{n} \partial l_i(\theta)/\partial\theta$$

and Fisher information matrix $I(\theta)$, which have been derived in Tawn (1988). The behaviour of the maximum likelihood estimator is the same as that derived for the block maxima approach (Smith 1985; Tawn 1988), which requires $\xi > -0.5$. The score statistic is

$$V_n = \frac{1}{n} S^{\top}(\hat{\theta}_n) I^{-1}(\hat{\theta}_n) S(\hat{\theta}_n).$$

Under standard regularity conditions, V_n would asymptotically follow a χ^2 distribution with 3 degrees of freedom. The GEV$_r$ distribution, however, violates the regularity conditions for the score test (e.g., Casella and Berger 2002, pp. 516–517), as its support depends on the parameter values unless $\xi = 0$. For illustration, Fig. 1 presents a visual comparison of the empirical distribution of V_n with $n = 5000$ from 5000 replicates, overlaid with the $\chi^2(3)$ distribution, for $\xi \in \{-0.25, 0.25\}$ and $r \in \{1, 2, 5\}$. The sampling distribution of V_n appears to be much heavier tailed than $\chi^2(3)$, and the mismatch increases as r increases as a result of the reduced convergence rate.

Although the regularity conditions do not hold, the score statistic still provides a measure of goodness-of-fit since it is a quadratic form of the score, which has expectation zero under the null hypothesis. Extremely large values of V_n relative to its sampling distribution would suggest lack of fit, and, hence, possible misspecification of $H_0^{(r)}$. So the key to applying the score test is to get an approximation of the sampling distribution of V_n. Two approaches for the approximation are proposed.

3.1 Parametric bootstrap

The first solution is parametric bootstrap. For hypothesis $H_0^{(r)}$, the test procedure goes as follows:

1. Compute $\hat{\theta}_n$ under H_0 with the observed data.
2. Compute the testing statistic V_n.
3. For every $k \in \{1, ..., L\}$ with a large number L, repeat:
   (a) Generate a bootstrap sample of size n for the r largest statistics from GEV$_r$ with parameter vector $\hat{\theta}_n$.
   (b) Compute the $\hat{\theta}_n^{(k)}$ under H_0 with the bootstrap sample.
   (c) Compute the score test statistic $V_n^{(k)}$.
4. Return an approximate p value of V_n as $L^{-1} \sum_{k=1}^{L} 1(V_n^{(k)} > V_n)$.

Straightforward as it is, the parametric bootstrap approach involves sampling from the null distribution and computing the MLE for each bootstrap sample, which can be very computationally expensive. This is especially true as the sample size n and/or the number of order statistics r included in the model increases.

3.2 Multiplier bootstrap

Multiplier bootstrap is a fast, large sample alternative to parametric bootstrap in goodness-of-fit testing (e.g., Kojadinovic and Yan 2012). The idea is to approximate the asymptotic

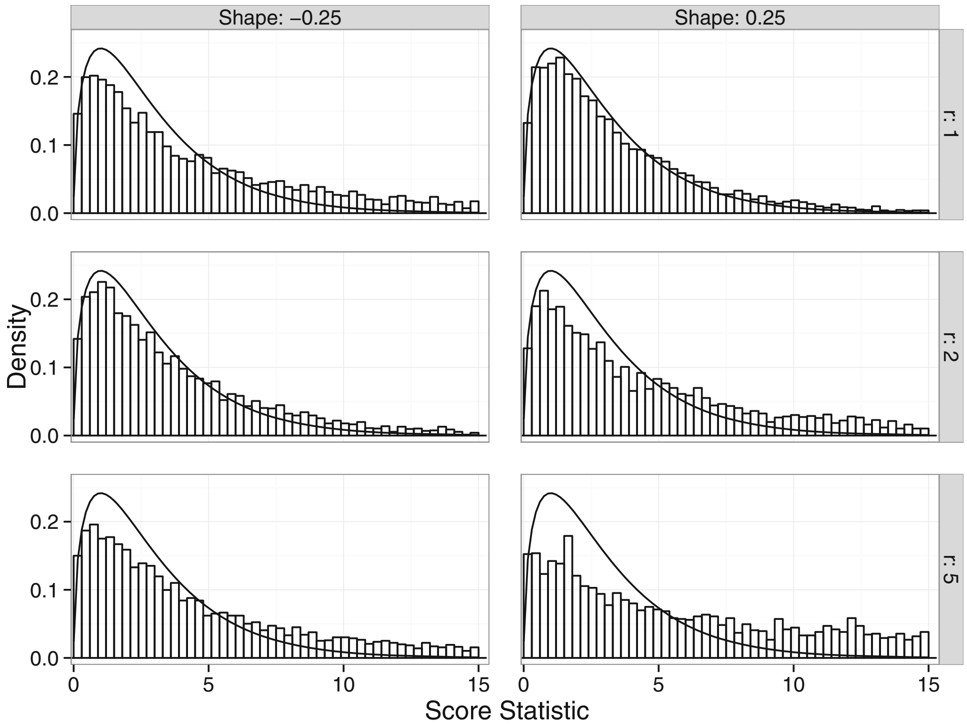

Fig. 1 Comparisons of the empirical distribution based on 5000 replicates of the score test statistic and the $\chi^2(3)$ distribution (*red solid curve*). The number of blocks used is $n = 5000$ with parameters $\mu = 0$, $\sigma = 1$, and $\xi \in (-0.25, 0.25)$

distribution of $n^{-1/2}I^{-1/2}(\theta)S(\theta)$ using its asymptotic representation

$$n^{-1/2}I^{-1/2}(\theta)S(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi_i(\theta),$$

where $\phi_i(\theta) = I^{-1/2}(\theta)S_i(\theta)$. Its asymptotic distribution is the same as the asymptotic distribution of

$$W_n(\mathbf{Z}, \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})\phi_i(\theta),$$

conditioning on the observed data, where $\mathbf{Z} = (Z_1, ..., Z_n)$ is a set of independent and identically distributed multipliers (independent of the data), with expectation 0 and variance 1, and $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. The multipliers must satisfy $\int_0^\infty \{\Pr(|Z_1| > x)\}^{\frac{1}{2}}\mathrm{d}x < \infty$. An example of a possible multiplier distribution is $N(0, 1)$.

The multiplier bootstrap test procedure is summarized as follows:

1. Compute $\hat{\theta}_n$ under H_0 with the observed data.
2. Compute the testing statistic V_n.
3. For every $k \in \{1, ..., L\}$ with a large number L, repeat:
   (a) Generate $\mathbf{Z}^{(k)} = (Z_1^{(k)}, \ldots, Z_n^{(k)})$ from $N(0, 1)$.
   (b) Compute a realization from the approximate distribution of $W_n(\mathbf{Z}, \theta)$ with $W_n(\mathbf{Z}^{(k)}, \hat{\theta}_n)$.
   (c) Compute $V_n^{(k)}(\hat{\theta}_n) = W_n^\top(\mathbf{Z}^{(k)}, \hat{\theta}_n)W_n(\mathbf{Z}^{(k)}, \hat{\theta}_n)$.
4. Return an approximate p value of V_n as $L^{-1}\sum_{k=1}^{L} 1(V_n^{(k)} > V_n)$.

This multiplier bootstrap procedure is much faster than parametric bootstrap procedure because, for each sample, it only needs to generate $\mathbf{Z}$ and compute $W_n(\mathbf{Z}, \hat{\theta}_n)$. The MLE only needs to be obtained once from the observed data.

4 Entropy difference test

Another specification test for the GEV$_r$ model is derived based on the difference in entropy for the GEV$_r$ and GEV$_{r-1}$ models. The entropy for a continuous random variable with density f is (e.g., Singh 2013)

$$E[-\ln f(y)] = -\int_{-\infty}^{\infty} f(y)\log f(y)\mathrm{d}y.$$

It is essentially the expectation of negative log-likelihood. The expectation can be approximated with the sample average of the contribution to the log-likelihood from the observed data, or simply the log-likelihood scaled by the sample size n. Assuming that the $r - 1$ top order statistics fit the GEV$_{r-1}$ distribution, the difference in the log-likelihood between GEV$_{r-1}$ and GEV$_r$ provides a measure of deviation from $H_0^{(r)}$. Its asymptotic distribution can be derived. Large deviation from the expected difference under $H_0^{(r)}$ suggests a possible misspecification of $H_0^{(r)}$.

From the log-likelihood contribution in (2), the difference in log-likelihood for the ith block, $Y_{ir}(\theta) = l_i^{(r)} - l_i^{(r-1)}$, is

$$Y_{ir}(\theta) = -\log\sigma - (1+\xi z_{ir})^{-\frac{1}{\xi}} + (1+\xi z_{ir-1})^{-\frac{1}{\xi}} - \left(\frac{1}{\xi}+1\right)\log(1+\xi z_{ir}). \tag{3}$$

Let $\bar{Y}_r = \frac{1}{n}\sum_{i=1}^n Y_{ir}$ and $S_{Y_r}^2 = \sum_{i=1}^n (Y_{ir} - \bar{Y}_r)^2/(n-1)$ be the sample mean and sample variance, respectively. Consider a standardized version of $\bar{Y}_r$ as

$$T_n^{(r)}(\theta) = \sqrt{n}(\bar{Y}_r - \eta_r)/S_{Y_r}, \tag{4}$$

where $\eta_r = -\log\sigma - 1 + (1+\xi)\psi(r)$, and $\psi(x) = \mathrm{d}\log\Gamma(x)/\mathrm{d}x$ is the digamma function. The asymptotic distribution of $T_n^{(r)}$ is summarized by Theorem 1 whose proof is relegated to Appendix 2.

Theorem 1 *Let $T_n^{(r)}(\theta)$ be the quantity computed based on a random sample of size n from the GEV$_r$ distribution with parameters θ and assume that $H_0^{(r-1)}$ is true. Then $T_n^{(r)}$ converges in distribution to $N(0,1)$ as $n \to \infty$.*

Note that in Theorem 1, $T_n^{(r)}$ is computed from a random sample of size n from a GEV$_r$ distribution. If the random sample were from a distribution in the domain of attraction of a GEV distribution, the quality of the approximation of the GEV$_r$ distribution to the r largest order statistics depends on the size of each block $B \to \infty$ with $r \ll B$. The block size B is not to be confused with the sample size n. Assuming $\xi > -0.5$, the proposed ED statistic for $H_0^{(r)}$ is $T_n^{(r)}(\hat{\theta}_n)$, where $\hat{\theta}_n$ is the MLE of θ with the r largest order statistics for the GEV$_r$ distribution. Since $\hat{\theta}_n$ is consistent for θ with $\xi > -0.5$, $T_n^{(r)}(\hat{\theta}_n)$ has the same limiting distribution as $T_n^{(r)}(\theta)$ under $H_0^{(r)}$.

To assess the convergence of $T_n^{(r)}(\hat{\theta}_n)$ to $N(0,1)$, 1000 GEV$_r$ replicates were simulated under configurations of $r \in \{2, 5, 10\}$, $\xi \in \{-0.25, 0, 0.25\}$, and $n \in \{50, 100\}$. Their quantiles are compared with those of $N(0,1)$ via quantile-quantile plots (not presented). It appears that a larger sample size is needed for the normal approximation to be good for larger r and negative ξ. This is expected because larger r means higher dimension of the data, and because the MLE only exists for $\xi > -0.5$ (Smith 1985). For r less than 5 and $\xi \geq 0$, the normal approximation is quite good; it appears satisfactory for sample size as small as 50. For r up to 10, sample size 100 seems to be sufficient.

5 Simulation results

5.1 Size

The empirical sizes of the tests are investigated first. For the score test, the parametric bootstrap version and the multiplier bootstrap version are equivalent asymptotically, but may behave differently for finite samples. It is of interest to know how large a sample size is needed for the two versions of the score test to hold their levels. Random samples of size n were generated from the GEV$_r$ distribution with $r \in \{1, 2, 3, 4, 5, 10\}$, $\mu = 0$, $\sigma = 1$, and $\xi \in \{-0.25, 0, 0.25\}$. All three parameters (μ, σ, ξ) were estimated.

When the sample size is small, there can be numerical difficulty in obtaining the MLE. For the multiplier bootstrap score and ED test, the MLE only needs to obtained once, for the dataset being tested. However, in addition, the parametric bootstrap score test must obtain a new sample and obtain the MLE for each bootstrap replicate. To assess the severity of this issue, 10,000 datasets were simulated for $\xi \in \{-0.25, 0, 0.25\}$, $r \in \{1, 2, 3, 4, 5, 10\}$, $n \in \{25, 50\}$, and the MLE was attempted for each dataset. Failure never occurred for $\xi \geq 0$. With $\xi = -0.25$ and sample size 25, the highest failure rate of 0.69 % occurred for $r = 10$. When the sample size is 50, failures only occurred when $r = 10$, at a rate of 0.04 %.

For the parametric bootstrap score test with sample size $n \in \{25, 50, 100\}$, Table 1 summarizes the empirical size of the test at nominal levels 1, 5, and 10 % obtained from 1000 replicates, each carried out with bootstrap sample size $L = 1000$. Included only are the cases that converged successfully. Otherwise, the results show that the agreement between the empirical levels and the nominal level is quite good for samples as small as 25, which may appear in practice when long record data is not available.

For the multiplier bootstrap score test, the results for sample sizes $n \in \{25, 50, 100, 200, 300, 400\}$ are summarized in Table 2. When the sample size is less than 100, it appears that there is a large discrepancy between the empirical and nominal level. For $\xi \in \{0, 0.25\}$, there is reasonable agreement between the empirical level and the nominal levels for sample size at least 100. For $\xi = -0.25$ and sample size at least 100, the agreement is good except for $r = 1$, in which case, the empirical level is noticeably larger than the nominal level. This may be due to different rates of convergence for various ξ values as ξ moves away from -0.5. It is also interesting to note that, everything else being held, the agreement becomes better as r increases. This may be explained by the more information provided by larger r for the same sample size n, as can be seen directly in the fisher information matrix (Tawn 1988, pp. 247–249). For the most difficult case with $\xi = -0.25$ and $r = 1$, the agreement gets better as sample size increases and becomes acceptable when sample size was 1000 (not reported).

To assess the convergence of $T_n^{(r)}(\hat{\theta}_n)$ to $N(0,1)$, 10,000 replicates of the GEV$_r$ distribution were simulated with $\mu = 0$ and $\sigma = 1$ for each configuration of $r \in \{2, 5, 10\}$, $\xi \in \{-0.25, 0, 0.25\}$, and $n \in \{50, 100\}$. A rejection for nominal level α, is denoted if $|T_n^{(r)}(\hat{\theta}_n)| > |Z_{\frac{\alpha}{2}}|$, where $Z_{\frac{\alpha}{2}}$ is the

Table 1 Empirical size (in %) for the parametric bootstrap score test under the null distribution GEV$_r$, with $\mu = 0$ and $\sigma = 1$ based on 1000 samples, each with bootstrap sample size $L = 1000$

Sample size	r	25			50			100		
Nominal size		1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
$\xi = -0.25$	1	0.4	2.8	6.0	1.1	4.8	9.3	0.6	4.1	8.0
	2	0.1	2.6	6.0	0.8	3.4	6.5	0.6	3.6	8.1
	3	0.3	2.5	5.0	0.8	4.3	7.7	1.1	4.8	8.1
	4	0.3	1.8	5.4	0.6	3.1	6.9	1.1	5.1	8.8
	5	0.4	2.4	6.7	0.4	3.3	8.3	0.6	3.1	6.5
	10	2.7	5.3	8.7	0.5	3.9	8.4	0.7	4.2	7.6
$\xi = 0$	1	1.3	5.2	8.9	1.6	5.3	9.0	0.8	4.7	9.3
	2	1.4	5.1	9.4	2.0	4.9	10.0	1.0	4.3	9.9
	3	1.7	6.2	10.9	2.1	6.0	10.2	0.8	4.9	9.8
	4	1.5	4.5	8.5	1.3	6.0	10.2	1.0	4.4	9.8
	5	1.6	5.8	10.4	2.4	6.2	9.9	1.2	5.0	9.7
	10	1.5	4.0	7.3	1.5	4.3	8.9	0.7	4.6	8.2
$\xi = 0.25$	1	1.7	4.5	9.7	2.6	7.1	11.5	1.1	4.6	9.1
	2	1.8	5.1	8.7	1.8	4.4	8.5	0.5	2.9	7.5
	3	1.5	4.4	9.4	1.5	3.7	8.1	1.0	4.2	9.4
	4	1.2	3.3	8.1	1.1	4.6	9.7	1.1	4.3	9.6
	5	1.7	4.4	9.4	1.1	4.2	8.6	0.6	4.8	9.6
	10	1.1	4.6	8.3	1.5	6.1	10.7	1.0	3.9	8.5

$\alpha/2$ percentile of the $N(0, 1)$ distribution. Using this result, the empirical size of the ED test can be summarized, and the results are presented in Table 3.

For sample size 50, the empirical size is above the nominal level for all configurations of r and ξ. As the sample size increases from 50 to 100, the empirical size stays the same or decreases in every setting. For sample size 100, the agreement between nominal and observed size appears to be satisfactory for all configurations of r and ξ. For sample size 50, the empirical size is slightly higher than the nominal size, but may be acceptable to some practitioners. For example, the empirical size for nominal size 10 % is never above 12 %, and for nominal size 5 %, empirical size is never above 7 %.

In summary, the multiplier bootstrap procedure of the score test can be used as a fast, reliable alternative to the parametric bootstrap procedure for sample size 100 or more when $\xi \geq 0$. When only small samples are available (less than 50 observations), the parametric bootstrap procedure is most appropriate since the multiplier version does not hold its size and the ED test relies upon samples of size 50 or more for the central limit theorem to take effect.

5.2 Power

The powers of the score tests and the ED test are studied with two data generating schemes under the alternative hypothesis. In the first scheme, 4 largest order statistics were generated from the GEV$_4$ distribution with $\mu = 0$, $\sigma = 1$, and $\xi \in \{-0.25, 0, 0.25\}$, and the 5th one was generated from a KumGEV distribution right truncated by the 4th largest order statistic. The KumGEV distribution is a generalization of the GEV distribution (Eljabri 2013) with two additional parameters a and b which alter skewness and kurtosis. Defining $G_r(\mathbf{x})$ to be the distribution function of the GEV$_r(\mu, \sigma, \xi)$ distribution, the distribution function of the KumGEV$_r(\mu, \sigma, \xi, a, b)$ is given by $F_r(\mathbf{x}) = 1 - \{1 - [G_r(\mathbf{x})]^a\}^b$ for $a > 0, b > 0$. The score test and the ED test were applied to the top 5 order statistics with sample size $n \in \{100, 200\}$. When $a = b = 1$, the null hypothesis of GEV$_5$ is true. Larger difference from 1 of parameters a and b means larger deviation from the null hypothesis of GEV$_5$.

Table 4 summarizes the empirical rejection percentages obtained with nominal size 5 %, for a sequence value of $a = b$ from 0.4 to 2.0, with increment 0.2. Both tests hold their sizes when $a = b = 1$ and have substantial power in rejecting the null hypothesis for other values of $a = b$. Between the two tests, the ED test demonstrated much higher power than the score test in the more difficult cases where the deviation from the null hypothesis is small; for example, the ED test's power almost doubled the score test's power for $a = b \in \{0.8, 1.2\}$. As expected, the powers of both tests increase as $a = b$ moves away from 1 or as the sample sizes increases.

In the second scheme, top 6 order statistics were generated from the GEV$_6$ distribution with $\mu = 0$, $\sigma = 1$, and $\xi \in \{-0.25, 0, 0.25\}$, and then the 5th order statistic was replaced from a mixture of the 5th and 6th order

Table 2 Empirical size (in %) for multiplier bootstrap score test under the null distribution GEV$_r$, with $\mu = 0$ and $\sigma = 1$

Sample size	r	25			50			100		
Nominal size		1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
$\xi = -0.25$	1	7.0	13.4	18.9	6.3	13.8	19.6	5.4	11.4	16.3
	2	2.0	6.9	13.4	1.3	6.4	12.4	1.6	6.9	13.6
	3	2.1	5.8	11.7	1.1	5.9	11.1	1.1	5.0	10.8
	4	3.3	7.2	12.3	1.1	4.9	10.8	1.0	5.2	11.9
	5	3.6	9.0	14.0	2.3	6.8	11.2	1.1	6.2	10.6
	10	2.0	7.0	10.3	2.6	7.4	12.8	2.1	6.4	10.1
$\xi = 0$	1	3.3	8.4	15.3	2.2	7.0	12.5	1.1	4.6	9.2
	2	2.8	8.7	14.4	1.8	7.5	13.0	0.9	5.7	10.3
	3	6.1	12.1	16.5	3.0	7.2	12.2	1.5	6.0	10.4
	4	5.1	10.4	14.5	3.6	10.1	14.9	1.0	5.6	10.3
	5	4.2	9.0	14.5	2.2	8.2	12.5	1.7	6.5	12.0
	10	3.1	9.2	14.4	2.4	6.4	9.8	0.6	4.6	9.0
$\xi = 0.25$	1	1.8	6.7	13.7	1.3	4.7	10.4	0.8	4.4	11.5
	2	5.7	12.7	17.1	4.7	9.9	14.9	3.5	7.4	11.6
	3	7.1	12.2	16.5	5.3	9.4	14.8	4.2	8.4	12.5
	4	5.4	9.8	16.8	3.7	9.0	13.4	2.6	6.0	11.4
	5	4.4	10.1	15.8	3.5	8.2	13.6	2.4	7.4	11.4
	10	3.3	8.9	15.3	2.4	6.6	12.3	1.6	5.8	10.9
Sample size	r	200			300			400		
Nominal size		1.0	5.0	10.0	1.0	5.0	10.0	1.0	5.0	10.0
$\xi = -0.25$	1	5.4	10.5	15.2	3.6	8.2	12.6	2.8	7.1	12.5
	2	1.4	6.7	12.8	1.4	6.4	11.4	1.4	5.1	10.9
	3	1.5	5.9	11.8	1.1	5.4	10.8	1.2	6.6	11.9
	4	1.1	5.6	10.6	1.0	5.6	11.5	1.0	4.7	9.0
	5	1.1	4.5	9.3	1.2	5.7	11.7	1.2	4.7	10.2
	10	1.4	6.4	11.6	1.7	6.2	11.3	0.8	5.0	10.1
$\xi = 0$	1	1.3	6.1	11.2	0.8	5.2	10.0	1.0	5.1	11.4
	2	0.5	5.0	10.6	1.2	5.7	11.8	1.0	5.9	11.0
	3	1.4	4.5	9.8	1.3	6.0	9.6	0.9	4.4	8.3
	4	1.1	5.4	10.6	1.3	5.2	9.9	0.9	5.0	9.1
	5	1.8	6.2	12.5	0.9	4.6	9.8	1.2	4.6	9.0
	10	1.1	3.8	9.3	0.9	5.2	12.6	1.2	4.9	9.7
$\xi = 0.25$	1	0.9	4.9	11.4	0.9	5.0	10.8	0.7	5.2	9.2
	2	3.2	7.9	11.7	2.3	7.1	11.2	2.5	6.6	12.1
	3	1.8	6.1	10.7	2.6	7.0	11.2	1.0	4.8	10.6
	4	1.2	4.9	11.2	1.2	6.0	9.9	1.2	5.8	11.8
	5	1.6	5.9	10.0	1.3	7.3	11.8	1.2	3.9	8.4
	10	1.7	6.6	12.4	0.9	4.4	9.8	1.6	5.7	10.4

1000 samples, each with bootstrap sample size $L = 1000$ were used. Although not shown, the empirical size for $r = 1$ and $\xi = -0.25$ becomes acceptable when sample size is 1000

statistics. The tests were applied to the sample of first 5 order statistics with sample sizes $n \in \{100, 200\}$. The mixing rate p of the 5th order statistic took values in $\{0.00, 0.10, 0.25, 0.50, 0.75, 0.90, 1.00\}$. When $p = 1$ the null hypothesis of GEV$_5$ is true. Smaller values of p indicate larger deviations from the null. Again, both tests hold their sizes when $p = 1$ and have substantial power for other values of p, which increases as p decreases or as the sample sizes increases. The ED test again outperforms the score test with almost doubled power in the most difficult cases with

Table 3 Empirical size (in %) for the entropy difference (ED) test under the null distribution GEV$_r$, with $\mu = 0$ and $\sigma = 1$ based on 10,000 samples

Sample size	r	50			100		
Nominal Size		1.0	5.0	10.0	1.0	5.0	10.0
$\xi = -0.25$	2	1.5	5.7	10.8	1.3	5.5	10.1
	5	2.4	6.8	11.9	1.6	5.9	10.6
	10	2.3	6.8	11.7	1.9	6.0	11.1
$\xi = 0$	2	1.3	5.6	11.0	1.2	5.3	10.4
	5	1.6	5.9	11.2	1.5	5.7	10.6
	10	2.3	6.5	11.8	1.6	5.9	10.7
$\xi = 0.25$	2	1.3	5.7	10.7	1.3	5.4	10.5
	5	1.6	5.8	11.5	1.3	5.6	10.2
	10	2.0	6.6	11.9	1.4	5.5	10.4

$p \in \{0.75, 0.90\}$. For sample size 100 with $p = 0.50$, for instance, the ED test has power above 93 % while the score test only has power above 69 % (Table 5).

6 Automated sequential testing procedure

As there are R hypotheses $H_0^{(r)}$, $r = 1, \ldots, R$, to be tested in a sequence in the methods proposed, the sequential, multiple testing issue needs to be addressed. Most methods for error control assume that all the tests can be run first and then a subset of tests are chosen to be rejected (e.g., Benjamini 2010a, b). The errors to be controlled are either the FWER (Shaffer 1995), or the FDR (Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001). In contrast to the usual multiple testing procedures, however, a unique feature in this setting is that the hypotheses must be rejected in an ordered fashion: if $H_0^{(r)}$ is rejected, $r < R$, then $H_0^{(k)}$ will be rejected for all $r < k \leq R$. Despite the extensive literature on multiple testing and the more recent developments on FDR control and its variants, no definitive procedure has been available for error control in ordered tests until the recent work of G'Sell et al. (2015).

Consider a sequence of null hypotheses $H_1, \ldots, H_m$. An ordered test procedure must reject $H_1, \ldots, H_k$ for some $k \in \{0, 1, \ldots, m\}$, which rules out the classical methods for FDR control Benjamini and Hochberg 1995. Let $p_1, \ldots, p_m \in [0, 1]$ be the corresponding p values of the m hypotheses such that p_j is uniformly distributed over [0, 1] when H_j is true. The methods of G'Sell et al. (2015) transform the sequence of p values to a monotone sequence and then apply the original Benjamini–Hochberg procedure on the monotone sequence. They proposed two rejections rules, each returning a cutoff $\hat{k}$ such that $H_1, \ldots, H_{\hat{k}}$ are rejected. The first is called ForwardStop,

$$\hat{k}_{\mathrm{F}} = \max\left\{k \in \{1, \ldots, m\} : -\frac{1}{k}\sum_{i=1}^{k}\log(1 - p_i) \leq \alpha\right\},$$

and the second is called StrongStop,

$$\hat{k}_{\mathrm{S}} = \max\left\{k \in \{1, \ldots, m\} : \exp\left(\sum_{j=k}^{m}\frac{\log p_j}{j}\right) \leq \frac{\alpha k}{m}\right\},$$

where α is a pre-specified level. Both rules were shown to control the FDR at level α under the assumption of independent p values. ForwardStop sets the rejection threshold at the largest k at which the average of first k transformed p values is small enough. As it does not depend on those p values with later indices, this rule is robust to potential misspecification at later indices. StrongStop offers a stronger guarantee than ForwardStop. If the non-null p values indeed precede the null

Table 4 Empirical rejection rate (in %) of the multiplier score test and the ED test in the first data generating scheme from 1000 replicates

Sample size	ξ	Test	Value of $a = b$								
			0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
100	−0.25	Score	99.9	84.8	20.4	5.4	21.0	41.2	62.3	79.0	83.0
		ED	100.0	99.0	46.5	4.6	48.7	89.5	99.2	100.0	99.8
	0	Score	100.0	87.0	21.6	7.4	24.2	48.9	67.8	79.6	89.4
		ED	100.0	98.8	40.0	5.2	40.6	87.2	98.5	100.0	99.7
	0.25	Score	100.0	87.7	20.3	6.2	25.8	54.2	74.2	82.9	89.5
		ED	100.0	97.5	37.7	4.8	34.8	78.1	96.1	99.5	99.7
200	−0.25	Score	100.0	98.6	40.7	5.2	29.8	64.7	86.4	95.9	97.5
		ED	100.0	100.0	78.4	6.2	70.0	99.2	100.0	100.0	100.0
	0	Score	100.0	99.4	44.6	6.1	34.9	75.0	92.4	97.3	98.6
		ED	100.0	99.9	75.0	5.5	64.6	98.1	99.8	100.0	100.0
	0.25	Score	100.0	99.3	44.5	6.3	37.0	73.4	91.8	97.0	98.9
		ED	100.0	100.0	71.0	5.2	57.2	95.9	100.0	100.0	100.0

Table 5 Empirical rejection rate (in %) of the multiplier score test and the ED tests in the second data generating scheme from 1000 replicates

Sample size	ξ	Test	Mixing rate p						
			0.00	0.10	0.25	0.50	0.75	0.90	1.00
100	−0.25	Score	99.7	99.5	95.5	69.4	24.1	7.8	5.8
		ED	100.0	100.0	100.0	97.7	51.8	10.9	6.2
	0	Score	100.0	99.7	97.8	72.4	22.7	6.2	6.8
		ED	100.0	100.0	100.0	96.0	47.6	10.3	5.6
	0.25	Score	99.9	99.7	96.6	70.8	24.7	5.8	5.3
		ED	100.0	100.0	99.9	93.6	43.4	9.8	5.2
200	−0.25	Score	99.9	100.0	99.7	95.6	43.4	11.4	5.1
		ED	100.0	100.0	100.0	100.0	83.6	20.0	5.8
	0	Score	100.0	100.0	100.0	96.5	44.4	11.2	5.4
		ED	100.0	100.0	100.0	100.0	79.5	20.0	5.5
	0.25	Score	100.0	100.0	100.0	97.2	46.9	9.2	5.5
		ED	100.0	100.0	100.0	99.7	72.5	17.9	4.2

p values, it controls the FWER at level α in addition to the FDR. Thus, for ForwardStop, this α refers to the FDR and for StrongStop, α refers to the FWER. As the decision to stop at k depends on all the p values after k, its power may be harmed if, for example, the very last p values are slightly higher than expected under the null hypotheses.

To apply the two rules to our setting, note that our objective is to give a threshold $\hat{r}$ such that the first $\hat{r}$ of $m = R$ hypotheses are accepted instead of rejected. Therefore, we put the p values in reverse order: let the ordered set of p values $\{p_1, \ldots, p_R\}$ correspond to hypotheses $\{H_0^{(R)}, \ldots, H_0^{(1)}\}$. The two rules give a cutoff $\hat{k} \in \{1, \ldots, R\}$ such that the hypotheses $H_0^{(R)}, \ldots, H_0^{(R-\hat{k}+1)}$ are rejected. If no $\hat{k} \in \{1, \ldots, R\}$ exists, then no rejection is made.

A caveat is that, unlike the setting of G'Sell et al. (2015), the p values of the sequential tests are dependent. Nonetheless, the ForwardStop and StrongStop procedures may still provide some error control. For example, in the non-sequential multiple testing scenario Benjamini and Yekutieli (2001) show that their procedure controls the FDR under certain positive dependency conditions, while Blanchard and Roquain (2009) implement adaptive versions of step-up procedures that provably control the FDR under unspecified dependence among p values.

The empirical properties of the two rules for the tests in this paper are investigated in simulation studies. To check the empirical FWER of the StrongStop rule, only data under the null hypotheses are needed. With $R = 10$, $\xi \in \{-0.25, 0.25\}$, $n \in \{30, 50, 100, 200\}$, $\mu = 0$, and $\sigma = 1$, 1000 GEV$_{10}$ samples were generated. For the ED, multiplier bootstrap score, and parametric bootstrap score test, the observed FWER is compared to the expected rates at various nominal α control levels. The StrongStop procedure is used, as well as no error control (i.e. a rejection occurs any time the raw p value is below the nominal level). The results of this simulation are presented in Fig. 2.

It is clear that the StrongStop reasonably controls the FWER for the ED test and the agreement between the observed and expected rate increases as the sample size increases. For both the parametric and multiplier bootstrap versions of the score test however, the observed FWER is above the expected rate, at times 10 % higher. Regardless, it is apparent that using no error control results in an inflated FWER, and this inflation can only increase as the number of tests increase.

To check the empirical FDR of the ForwardStop rule, data need to be generated from a non-null model. To achieve this, consider the sequence of specification tests of GEV$_r$ distribution with $r \in \{1, \ldots, 6\}$, where the 5th and 6th order statistics are misspecified. Specifically, data from the GEV$_7$ distribution with $\mu = 0$ and $\sigma = 1$ were generated for n blocks; then the 5th order statistic is replaced with a 50/50 mixture of the 5th and 6th order statistics, and the 6th order statistic is replaced with a 50/50 mixture of the 6th and 7th order statistics. This is replicated 1000 times for each value of $\xi \in \{-0.25, 0.25\}$ and $n \in \{30, 50, 100, 200\}$. For nominal level α, the observed FDR is defined as the number of false rejections (i.e. any rejection of $r \leq 4$) divided by the number of total rejections.

The results are presented in Fig. 3. The plots show that the ForwardStop procedure controls the FDR for the ED test, while for both versions of the score test, the observed FDR is slightly higher than the expected at most nominal rates. Here, sample size does not appear to effect the observed rates.

Similarly, the observed FWER rate in this particular simulation setting can be found by taking the number of simulations with at least one false rejection (here, any rejection of $r \leq 4$) and dividing that number by the total number of simulations. This calculation is performed for a variety

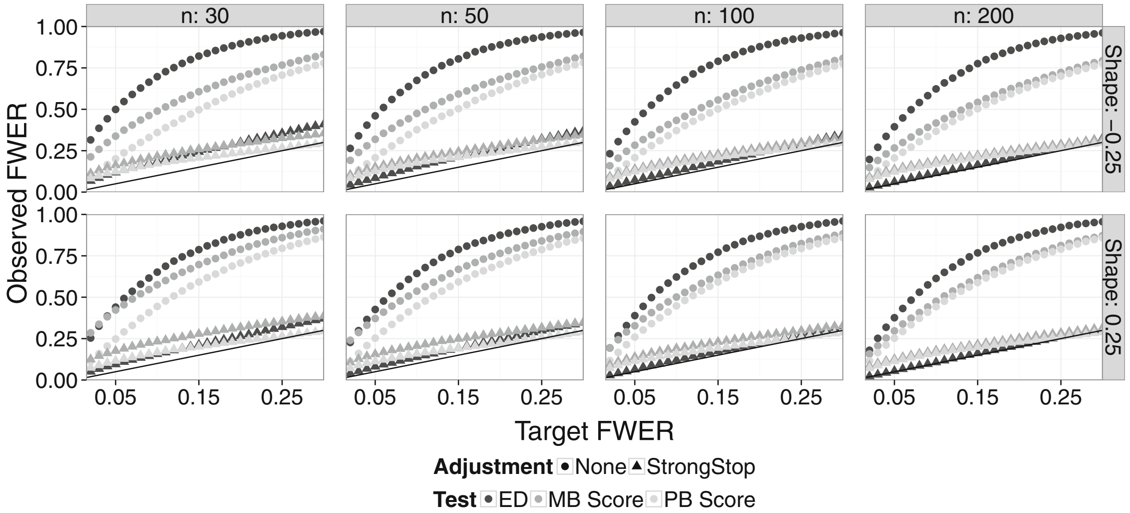

Fig. 2 Observed FWER for the ED, parametric bootstrap (PB) score, and multiplier bootstrap (MB) score tests (using No Adjustment and StrongStop) versus expected FWER at various nominal levels. The 45° *line* indicates agreement between the observed and expected rates under H_0

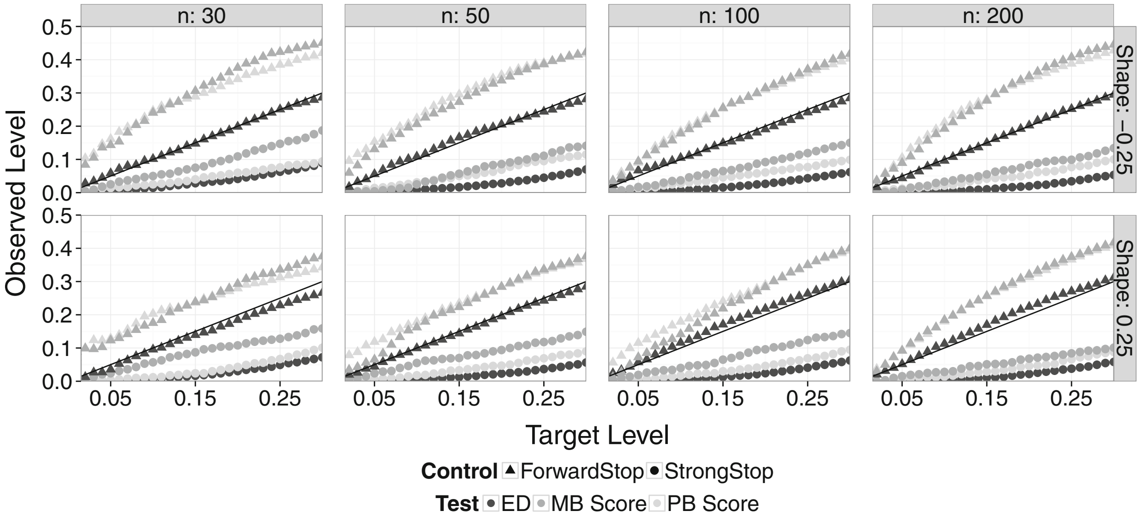

Fig. 3 Observed FDR (from ForwardStop) and observed FWER (from StrongStop) versus expected FDR and FWER, respectively, at various nominal levels. This is for the simulation setting described in Sect. 6, using the ED, parametric bootstrap (PB) score, and multiplier bootstrap (MB) score tests. The 45° *line* indicates agreement between the observed and expected rates

of nominal levels α, using the StrongStop procedure. The results are presented in Fig. 3. In this particular simulation setting, the StrongStop procedure controls the FWER for the ED test and both versions of the score test at all sample sizes investigated.

It is of interest to investigate the performance of the ForwardStop and StrongStop in selecting r for the r largest order statistics method. In the simulation setting described in the last paragraph, the correct choice of r should be 4, and a good testing procedure should provide a selection $\hat{r}$ close to 4. The choice $\hat{r} \in \{0, \ldots, 6\}$ is recorded using the ED test and bootstrap score tests with both ForwardStop and StrongStop. Due to space constraints, we choose to present one setting, where $\xi = 0.25$ and $n = 100$. The non-adjusted sequential procedure is also included, testing in an ascending manner from $r = 1$ and $\hat{r}$ is chosen by the first rejection found (if any). The results are summarized in Table 6.

In general, larger choices of α lead to a higher percentage of $\hat{r} = 4$ being correctly chosen with ForwardStop or StrongStop. Intuitively, this is not surprising since a smaller α makes it more difficult to reject the 'bad' hypotheses of $r \in \{5, 6\}$. A larger choice of α also leads to a higher probability of rejecting too many tests; i.e. choosing r too small. From the perspective of model specification, this is more desirable than accepting true negatives. A choice of 6, 5, or 0 is problematic, but choosing 1, 2, or 3 is acceptable, although

Table 6 Percentage of choice of r using the ForwardStop and StrongStop rules at various significance levels or FDRs, under ED, parametric bootstrap (PB) score, and multiplier bootstrap (MB) score tests, with $n = 100$ and $\xi = 0.25$ for the simulation setting described in Sect. 6

Test	r	Unadjusted						ForwardStop						StrongStop					
	Significance:	0.01	0.05	0.1	0.2	0.3	0.4	0.01	0.05	0.1	0.2	0.3	0.4	0.01	0.05	0.1	0.2	0.3	0.4
ED	6	19.0	3.4	1.5	0.5	0.0	0.0	86.6	69.8	58.5	43.0	30.0	22.4	52.4	22.2	13.1	5.4	1.7	1.0
	5	1.9	2.1	1.1	0.7	0.2	0.1	1.3	1.5	1.1	0.9	0.2	0.0	25.0	18.9	13.6	6.7	2.9	1.4
	4	76.2	79.9	70.2	50.5	35.2	22.3	12.0	25.1	31.7	33.4	31.6	26.3	22.6	58.9	72.9	84.9	89.0	85.7
	3	0.8	4.1	7.6	11.8	15.3	15.1	0.1	3.4	6.2	12.4	17.0	18.5	0.0	0.0	0.3	2.2	4.5	6.9
	2	1.1	5.3	9.5	16.2	19.9	22.8	0.0	0.1	1.9	5.4	9.4	11.5	0.0	0.0	0.1	0.7	1.8	4.4
	1	1.0	5.2	10.1	20.3	29.4	39.7	0.0	0.1	0.6	4.9	11.8	21.3	0.0	0.0	0.0	0.1	0.1	0.6
	0	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
PB Score	6	35.8	16.1	8.5	1.8	0.6	0.2	53.5	33.0	23.4	13.9	8.6	5.9	40.6	25.1	18.8	12.1	7.6	5.6
	5	2.5	1.8	0.8	0.5	0.2	0.1	1.9	1.3	0.7	0.8	0.4	0.4	29.8	37.7	29.2	17.5	10.5	6.6
	4	58.4	68.1	63.9	46.1	31.3	19.8	42.6	57.8	58.9	51.1	42.0	31.1	29.3	36.5	50.5	65.7	73.5	74.6
	3	0.5	2.3	4.5	6.8	7.1	8.0	1.4	4.2	9.0	15.5	16.0	17.1	0.0	0.4	1.1	3.3	4.8	5.6
	2	0.6	3.0	4.9	10.3	11.7	12.7	0.5	1.7	3.0	7.3	10.2	12.3	0.1	0.1	0.2	1.0	2.4	4.5
	1	0.8	3.7	6.9	13.9	18.3	18.5	0.1	1.3	2.3	4.6	8.3	9.6	0.2	0.2	0.2	0.4	1.2	2.7
	0	1.4	5.0	10.5	20.6	30.8	40.7	0.0	0.7	2.7	6.8	14.5	23.6	0.0	0.0	0.0	0.0	0.0	0.4
MB Score	6	49.9	16.9	6.9	1.3	0.2	0.0	71.7	40.3	24.7	12.6	7.8	5.5	51.6	27.3	16.9	10.4	6.2	4.3
	5	2.5	2.3	0.7	0.3	0.1	0.0	1.3	2.0	0.8	0.5	0.3	0.4	39.4	50.7	42.9	25.5	15.8	9.4
	4	38.3	59.6	59.1	44.0	31.2	18.4	26.6	53.3	59.3	49.9	40.1	28.0	6.2	18.5	35.0	55.4	62.3	64.8
	3	1.6	2.8	4.0	6.6	7.5	6.0	0.3	2.8	7.4	15.7	16.0	17.9	0.6	1.2	2.5	3.6	7.0	7.8
	2	2.7	4.4	7.1	11.0	10.0	10.6	0.1	0.6	3.4	8.6	9.7	9.5	0.7	0.8	1.2	2.8	5.3	7.0
	1	4.2	8.3	10.6	14.0	19.5	20.0	0.0	0.9	2.7	4.7	7.9	7.8	1.5	1.5	1.5	2.3	3.4	6.4
	0	0.8	5.7	11.6	22.8	31.5	45.0	0.0	0.1	1.7	8.0	18.2	30.9	0.0	0.0	0.0	0.0	0.0	0.3

Correct choice is $r = 4$

some information is lost. When no adjustment is used and an ascending sequential procedure is used, both tests have reasonable classification rates. When $\alpha = 0.05$, the ED test achieves the correct choice of r 79.9 % of the time, with the parametric bootstrap and multiplier bootstrap score tests achieving 68.1 and 59.6 % respectively. Of course, as the number of tests (i.e., R) increase, with no adjustment the correct classification rates will go down and the ForwardStop/StrongStop procedures will achieve better rates. This may not be too big an issue here as R is typically small. In the case where rich data are available and R is big, the ForwardStop and StrongStop become more useful as they are designed to handle a large number of ordered hypothesis.

7 Illustrations

7.1 Lowestoft sea levels

Sea level readings in 60 and 15 minute intervals from a gauge at Lowestoft off the east coast of Britain during the years 1964–2014 are available from the UK Tide Gauge Network website. The readings are hourly from 1964–1992 and in fifteen minute intervals from 1993 to present. Accurate estimates of extreme sea levels are of great interest. The current data are of better quality and with longer record than those used in Tawn (1988)—annual maxima during 1953–1983 and hourly data during 1970–1978 and 1980–1982.

Justification of the statistical model was considered in detail by Tawn (1988). The three main assumptions needed to justify use of the GEV$_r$ model are: (1) The block size B is large compared to the choice of r; (2) Observations within each block and across blocks are approximately independent; and (3) The distribution of the block maxima follows GEV$_1$. The first assumption is satisfied, by letting $R = 125$, and noting that the block size for each year is $B = 365 \times 24 = 8760$ from 1964–1992 and $B = 365 \times 96 = 35040$ from 1993–2014. This ensures that $r \ll B$. The third assumption is implicitly addressed in the testing procedure; if the goodness-of-fit test for the block maxima rejects, all subsequent tests for $r > 1$ are rejected as well.

The second assumption can be addressed in this setting by the concept of independent storms (Tawn 1988). The idea is to consider each storm as a separate event, with each storm having some storm length, say τ. Thus, when selecting the r largest values from each block, only a single contribution can be obtained from each storm, which can be considered the r largest independent annual events. By choosing τ large enough, this ensures both approximate independence of observations within each block and across blocks. The procedure to extract the independent r largest annual events is as follows:

1. Pick out the largest remaining value from the year (block) of interest.
2. Remove observations within a lag of $\tau/2$ from both sides of the value chosen in step 1.
3. Repeat (within each year) until the r largest are extracted.

A full analysis is performed on the Lowestoft sea level data using $\tau = 60$ as the estimated storm length (Tawn 1988). Using $R = 125$, both the parametric bootstrap score (with bootstrap sample size $L = 10{,}000$) and ED test are applied sequentially on the data. The p values of the sequen-

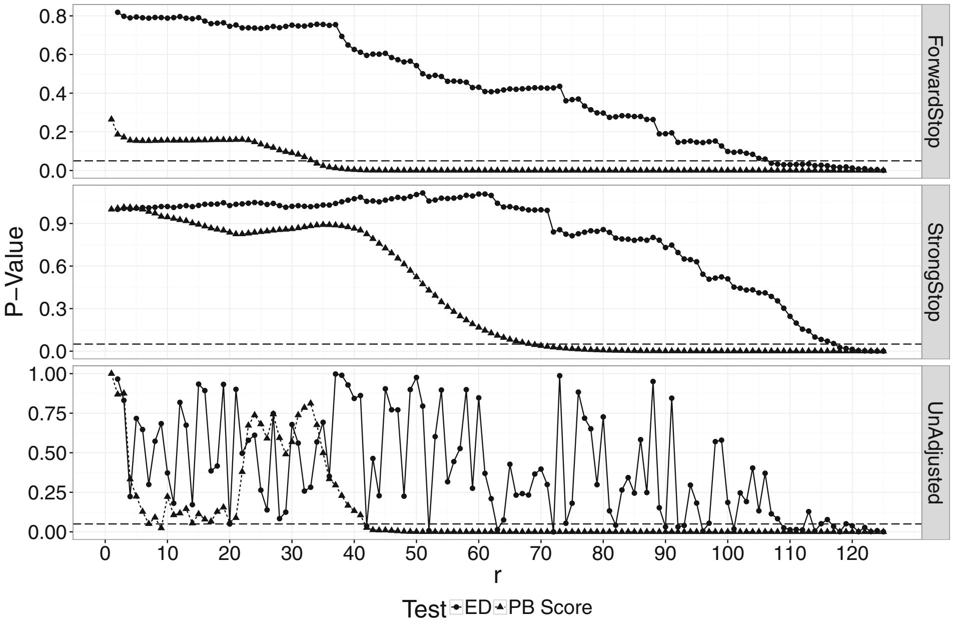

Fig. 4 P Values using ForwardStop, StrongStop, and no adjustment for the ED and PB Score tests applied to the Lowestoft sea level data. The *horizontal dashed line* represents the 0.05 possible cutoff value

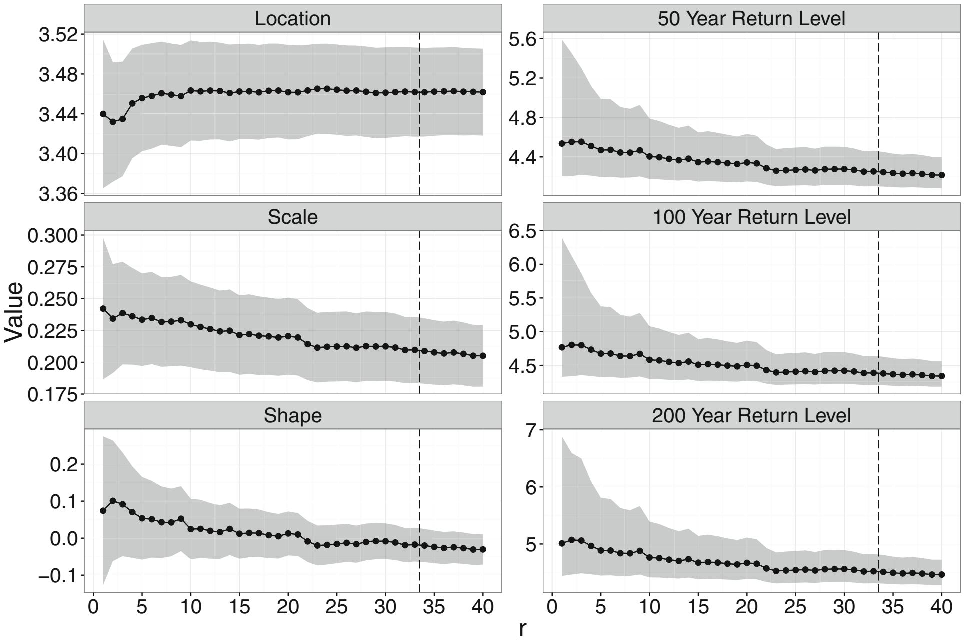

Fig. 5 Location, scale, and shape parameter estimates, with 95 % profile confidence intervals for $r = 1, \ldots, 40$ for the Lowestoft sea level data. Also included are the estimates and 95 % profile likelihood confidence intervals for the 50 and 100 year return levels

tial tests (adjusted and unadjusted) can be seen in Fig. 4. Due to the large number of tests, the adjustment for multiplicity is desired and thus, ForwardStop is used to choose r. For this dataset, the score test is more powerful than the ED test. With ForwardStop and the score test, Fig. 4 suggests that $r = 33$. The remainder of this analysis proceeds with the choice of $r = 33$. The estimated parameters and corresponding 95 % profile confidence intervals for $r = 1$ through $r = 40$ are shown in Fig. 5.

When $r = 33$, the parameters are estimated as $\hat{\mu} = 3.462$ (0.023), $\hat{\sigma} = 0.210$ (0.013), and $\hat{\xi} = -0.017$ (0.023), with standard errors in parenthesis. An important risk measure is the t-year return level z_t (e.g., Hosking 1990; Ribereau et al. 2008; Singo et al. 2012). It can be thought of here as the sea level that is exceeded once every t years on average. Specifically, the t-year return level is the $1 - 1/t$ quantile of the GEV distribution

$$z_t = \begin{cases} \mu - \frac{\sigma}{\xi}\{1 - [-\log(1 - \frac{1}{t})]^{-\xi}\}, & \xi \neq 0, \\ \mu - \sigma \log[-\log(1 - \frac{1}{t})], & \xi = 0. \end{cases}$$

The return levels can be estimated with parameter values replaced with their estimates, and confidence intervals can be constructed using profile likelihood (e.g., Coles 2001, p. 57).

The 95 % profile likelihood confidence intervals for the 50, 100, and 200 year return levels (i.e. z_{50}, z_{100}, z_{200}) are given by (4.102, 4.461), (4.210, 4.641) and (4.312, 4.824), respectively. The benefit of using $r = 1$ versus $r = 33$ can be seen in the return level confidence intervals in Fig. 5. For example, the point estimate of the 100 year return level decreases slightly as r increases and the width of the 95 % confidence interval decreases drastically from 2.061 ($r = 1$) to 0.432 ($r = 33$), as more information is used. The lower bound of the interval however remains quite stable, shifting from 4.330 to 4.210—less than a 3 % change. Similarly, the standard error of the shape parameter estimate decreases by over two-thirds when using $r = 33$ versus $r = 1$.

7.2 Annual maximum precipitation: Atlantic City, NJ

The top 10 annual precipitation events (in centimeters) were taken from the daily records of a rain gauge station in Atlantic City, NJ from 1874–2015. The year 1989 is missing, while the remaining records are greater than 98 % complete. This provides a total record length of 141 years. The raw data is a part of the Global Historical Climatology Network (GHCN-Daily), with an overview given by Menne et al. (2012). The specific station identification in the dataset is USW00013724.

Unlike for the Lowestoft sea level data, a rather small value is set for R at $R = 10$ because of the much lower frequency of the daily data. Borrowing ideas from Sect. 7.1, a storm length of $\tau = 2$ is used to ensure approximate independence of observations. Both the parametric bootstrap score (with $L =$ 10,000) and ED test are applied sequentially on the data. The p values of the sequential tests (ForwardStop, StrongStop, and unadjusted) are shown in Fig. 6. The score test does not pick up anything. The ED test obtains raw p values 0.002 and 0.016, respectively, for $r = 9$ and $r = 10$, which translates into a rejection using ForwardStop. Thus, Fig. 6 suggests that $r = 8$ be used for the analysis.

With $r = 8$, the estimated parameters are given as $\hat{\mu} =$ 6.118 (0.139), $\hat{\sigma} = 2.031$ (0.118), and $\hat{\xi} = 0.219$ (0.032).

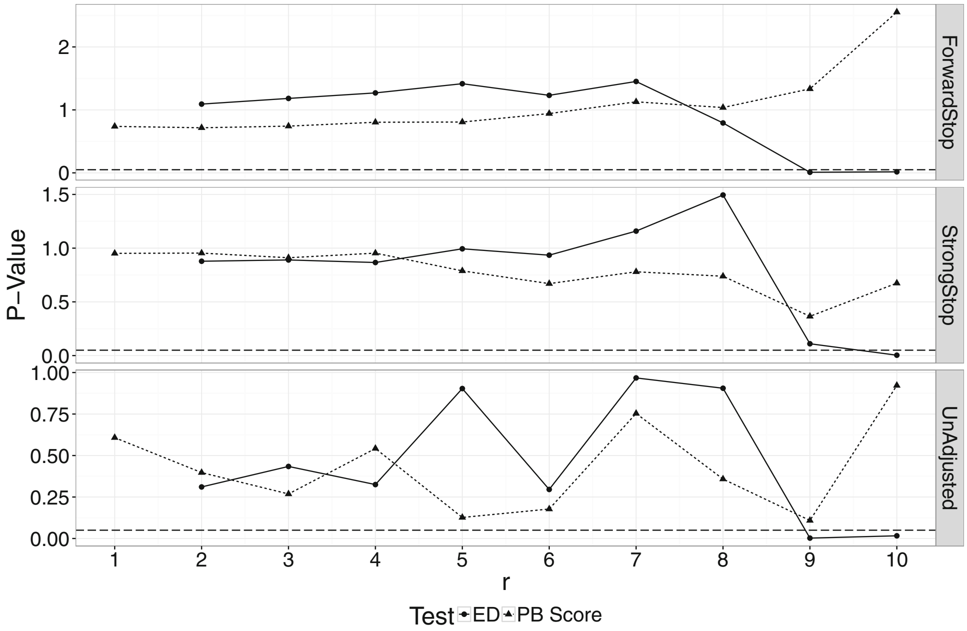

Fig. 6 P Values using ForwardStop, StrongStop, and no adjustment for the ED and PB Score tests applied to the Atlantic City precipitation data. The *horizontal dashed line* represents the 0.05 possible cutoff value

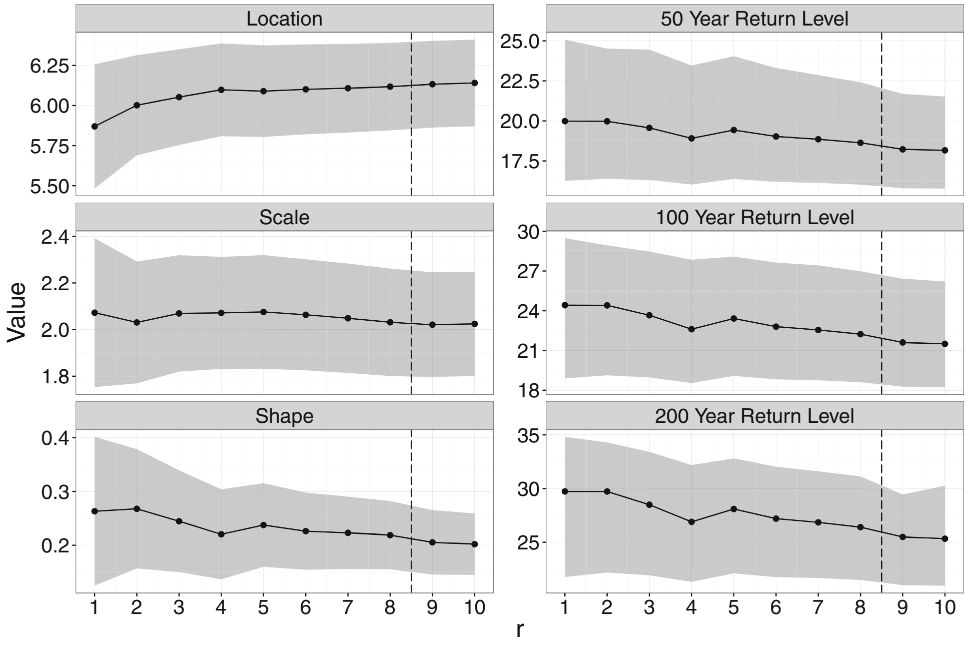

Fig. 7 Location, scale, and shape parameter estimates, with 95 % delta confidence intervals for $r = 1$ through $r = 10$ for the Atlantic City precipitation data. Also included are the estimates and 95 % profile likelihood confidence intervals for the 50, 100, and 200 year return levels

This suggests a heavy upper tail for the estimated distribution (i.e. $\hat{\xi} > 0$). The progression of parameters and certain return level estimates can be seen in Fig. 7. The 50, 100, and 200 year return level 95 % confidence intervals for $r = 8$ are calculated using the profile likelihood method and are given by (16.019, 22.411), (18.606, 26.979), and (21.489, 31.136), respectively. The advantages of using $r = 8$ versus the block maxima for analysis are quite clear from Fig. 7. The standard error of the shape parameter decreases from 0.071 to 0.032, a decrease of over 50 %. Similarly, the 50 year return level 95 % confidence intervals decreases in width by over 25 %.

8 Discussion

We proposed two model specification tests for a fixed number of largest order statistics as the basis for selecting r for the r largest order statistics approach in extreme value analysis. The score test has two versions of bootstrap procedure: the multiplier bootstrap method providing a fast, large sample alternative to the parametric bootstrap method, with a speedup of over 100 times. The ED test depends on asymptotic normal approximation of the testing statistic, which becomes acceptable for sample size over 50. It assumes that the $r - 1$ top order statistics included already fits the GEV$_{r-1}$ distribution. Therefore, the initial hypothesis at $r = 1$ needs to be tested with the score tests. Both tests hold their size better when the shape parameter is further away from the lower limit of -0.5 or sample size is larger. When only small samples are available (50 observations or less), the parametric bootstrap score test is recommended.

Alternative versions of the ED test have been explored. One may define the testing statistics as the difference in entropy between GEV$_1$ and GEV$_r$, instead of between GEV$r - 1$ and GEV$_r$. Nonetheless, it appeared to require a larger sample to hold its size from our simulation studies (not reported). In the calculation of $T_n^{(r)}$, the block maxima MLE $\hat{\theta}_n^{(1)}$ can be used as an estimate for θ in place of $\hat{\theta}_n^{(r)}$. Again, in our simulation studies, this version of the ED test was too conservative, thus reducing the power when the sample size was not large enough. This may be explained in that the resulting $\hat{S}_{Y_r}$ underestimates S_{Y_r}.

Naively, the tests may be performed sequentially for each $r \in \{1, \ldots, R\}$, for a prefixed, usually small R, at a certain significance level until $H_0^{(r)}$ is rejected. The issue of multiple, sequential testing is addressed in detail by adapting two very recent stopping rules to control the FDR and the FWER that are developed specifically for situations when hypotheses must be rejected in an ordered fashion (G'Sell et al. 2015). It is shown that these automated procedures reasonably control the error rate for the tests discussed in this paper. The naive unadjusted sequential testing procedure also appears to have decent performance at choosing the correct r and for this scenario is more conservative (selects a smaller r) than the stopping rules.

The tests can be extended to allow covariates in the parameters. For example, extremal precipitation in a year may be affected by large scale climate indexes such as the Southern Oscillation Index (SOI), which may be incorporated as a covariate in the location parameter (e.g., Shang et al. 2011). Both tests can be carried out with additional model parameters. When the underlying data falls into a rich class of dependence structures (such as time series), this dependence may be incorporated directly instead of using a procedure to achieve approximate independence (e.g. the storm length τ in Sect. 7). For example, take the GEV-GARCH model (Zhao et al. 2011) when $r = 1$. It may be extended to the case where $r > 1$ and the tests presented here may be applied to select r under this model assumption.

Acknowledgements We would like to thank Dr. Zhiyi Chi for his comments and discussion on the multiple testing issues. This research was partially supported by an NSF Grant (DMS 1521730), a University of Connecticut Research Excellence Program Grant, and a contract with Environment and Climate Change Canada.

Appendix 1: Generating from GEV$_r$ distribution

The GEV$_r$ distribution is closely connected to the GEV distribution. Let $X_1 > \cdots > X_r$ follow a GEV$_r$ distribution (1). It is obvious that the GEV$_1$ distribution is the GEV distribution with the same parameters, which is the marginal distribution of X_1. More interestingly, note that, the conditional distribution of X_2 given $X_1 = x_1$ is simply the GEV distribution righted truncated by x_1. In general, given $(X_1, \ldots, X_k) = (x_1, \ldots, x_k)$ for $1 \leq k < r$, the conditional distribution of X_{k+1} is the GEV distribution righted truncated at x_k. This property can be exploited to generate the r components in a realized GEV$_r$ observation.

The pseudo algorithm to generate a single observation is the following:

```
– Generate the first value $x_1$ from the (unconditional) GEV distribution.
– For $i = 2, \ldots, r$:
    • Generate $x_i$ from the GEV distribution right truncated by $x_{i-1}$.
```

The resulting vector $(x_1, \ldots, x_r)$ is a single observation from the GEV$_r$ distribution.

For $\xi \to 0$, caveat is needed in numerical evaluation. Using function `expm1` for $\exp(1 + x)$ for $x \to 0$ improved accuracy in comparison to a few implementations in existing R packages. For readability, here is a simplified version of our implemetation in R package `eva` (Bader and Yan 2015).

```
## Quantile function of a GEVr(loc, scale, shape)
qgev <- function(p, loc = 0, scale = 1, shape = 0,
                 lower.tail = TRUE, log.p = FALSE) {
  if (log.p) p <- exp(p)
  if(shape == 0) {
    loc - scale * log(-log(p))
  } else
    loc + scale * expm1(log(-log(p)) * -shape) / shape)
  }

## Random number generator of GEVr;
## Returns a matrix of n rows and r columns,
## each row a draw from GEVr
rgevr <- function(n, r, loc = 0, scale = 1, shape = 0) {
  umat <- matrix(runif(n * r), n, r)
  if (r > 1) {
    matrix(qgev(t(apply(umat, 1, cumprod)),
                loc, scale, shape),
           ncol = r)
  } else {
    qgev(umat, loc, scale, shape)
  }
}
```

Appendix 2: Asymptotic distribution of $T_n^{(r)}(\theta)$

Proof (Theorem 1) Consider a random vector $(X_1, X_2, ..., X_r)$ which follows a $\text{GEV}_r(\theta)$ distribution. The following result given by Tawn (1988, p. 248) will be used:

$$\begin{aligned} h(j|\theta, a, b, c) &\equiv E[Z_j^a(1+\xi Z_j)^{-(\frac{1}{\xi}+b)} \log^c(1+\xi Z_j)] \\ &= \frac{(-\xi)^{c-a}}{\Gamma(j)} \sum_{\alpha=0}^{a} (-1)^\alpha \binom{a}{\alpha} \Gamma^{(c)} \\ &\quad (j + b\xi - \alpha\xi + 1) \qquad (5) \end{aligned}$$

where $Z_j = (X_j - \mu)/\sigma$ and $\Gamma^{(c)}$ is the cth derivative of the gamma function, for $a \in \mathbb{Z}$, $b \in \mathbb{R}$, and $c \in \mathbb{Z}$, such that $(j + b\xi - \alpha\xi + 1) \notin \{0, -1, -2, \ldots\}$, $\alpha = 0, 1, \ldots, a$.

Assume that $\xi \neq 0$ and $1+\xi Z_j > 0$ for $j = 1, \ldots, r$. The difference in log-likelihoods for a single observation from the $\text{GEV}_r(\theta)$ and $\text{GEV}_{r-1}(\theta)$ distribution is given by (3) in Sect. 4. Thus, the first moment of Y_{ir} is

$$\begin{aligned} E[Y_{1r}] = &-\log\sigma - h(r|\theta, 0, 0, 0) + h(r-1|\theta, 0, 0, 0) \\ &- \left(\frac{1}{\xi} + 1\right) h(r|\theta, 0, -\xi^{-1}, 1) \\ = &-\log\sigma - 1 + (1+\xi)\psi(r) \end{aligned}$$

where $\psi(x) = \frac{\Gamma^{(1)}(x)}{\Gamma(x)}$.

To prove that the second moment of Y_{ir} is finite, note that

$$\begin{aligned} |Y_{1r}| \leq 4 \max \Bigg\{ &\left|\log\sigma\right|, \left|(1+\xi Z_{1r})^{-\frac{1}{\xi}}\right|, \\ &\left|(1+\xi Z_{1r-1})^{-\frac{1}{\xi}}\right|, \left|\left(\frac{1}{\xi}+1\right)\log(1+\xi Z_{1r-1})\right| \Bigg\}, \end{aligned}$$

which implies

$$\begin{aligned} Y_{1r}^2 \leq 16 \Bigg(\max \Bigg\{ &\left|\log\sigma\right|, \left|(1+\xi Z_{1r})^{-\frac{1}{\xi}}\right|, \\ &\left|(1+\xi Z_{1r-1})^{-\frac{1}{\xi}}\right|, \left|\left(\frac{1}{\xi}+1\right)\log(1+\xi Z_{1r-1})\right| \Bigg\} \Bigg)^2. \end{aligned}$$

The bound of $E(Y_{1r}^2)$ can be established by applying (5) to the last three terms in the max operator,

$$\begin{aligned} &E[(1+\xi Z_{1r})^{-\frac{2}{\xi}}] = h(r|\theta, 0, \xi^{-1}, 0) < \infty, \\ &E[(1+\xi Z_{1r-1})^{-\frac{2}{\xi}}] = h(r-1|\theta, 0, \xi^{-1}, 0) < \infty, \\ &E[\log^2(1+\xi Z_{1r-1})] = h(r-1|\theta, 0, -\xi^{-1}, 2) < \infty. \end{aligned}$$

The desired result then follows from the central limit theorem and Slutsky's theorem.

The case where $\xi = 0$ in Theorem 1 can easily be derived by taking the limit as $\xi \to 0$ in (3) and in (5) by the Dominated Convergence Theorem. □

References

An, Y., Pandey, M.D.: The r largest order statistics model for extreme wind speed estimation. J. Wind Eng. Ind. Aerodyn. **95**(3), 165–182 (2007)

Bader, B., Yan, J.: eva: Extreme Value Analysis with Goodness-of-Fit Testing (2015)

Benjamini, Y.: Discovering the false discovery rate. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(4), 405–416 (2010a)

Benjamini, Y.: Simultaneous and selective inference: current successes and future challenges. Biom. J. **52**(6), 708–721 (2010b)

Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**(1), 289–300 (1995)

Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Stat. **29**(4), 1165–1188 (2001)

Blanchard, G., Roquain, É.: Adaptive false discovery rate control under independence and dependence. J. Mach. Learn. Res. **10**, 2837–2871 (2009)

Casella, G., Berger, R.L.: Statistical Inference, 2nd edn. Duxbury, Pacific Grove (2002)

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, 1st edn. Springer, New York (2001)

Dupuis, D.J.: Extreme value theory based on the r largest annual events: a robust approach. J. Hydrol. **200**(1), 295–306 (1997)

Eljabri, S.S.M.: New statistical models for extreme values. PhD thesis, The University of Manchester, Manchester, UK (2013)

Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling ExtremalEvents. Springer, New York (1997)

G'Sell, M.G., Wager, S., Chouldechova, A., Tibshirani, R.: Sequential selection procedures and false discovery rate control. J. R. Stat. Soc. Ser. B (Stat. Methodol) (2015). doi:10.1111/rssb.12122

Guedes Soares, C., Scotto, M.G.: Application of the r largest-order statistics for long-term predictions of significant wave height. Coast. Eng. **51**(5), 387–394 (2004)

Hosking, J.R.M.: L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. R. Stat. Soc. Ser. B (Methodol.l) **52**(1), 105–124 (1990)

Kojadinovic, I., Yan, J.: Goodness-of-fit testing based on a weighted bootstrap: a fast large-sample alternative to the parametric bootstrap. Can. J. Stat. **40**(3), 480–500 (2012)

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., Houston, T.G.: An overview of the global historical climatology network-daily database. J. Atmos. Ocean. Technol. **29**(7), 897–910 (2012)

Rao, C.R.: Score test: historical review and recent developments. In: Balakrishnan, N., Kannan, N., Nagaraja, H.N. (eds.) Advances in Ranking and Selection, Multiple Comparisons, and Reliability, pp. 3–20. Springer, New York (2005)

Ribereau, P., Guillou, A., Naveau, P.: Estimating return levels from maxima of non-stationary random sequences using the generalized PWM method. Nonlinear Proces. Geophys. **15**(6), 1033–1039 (2008)

Scarf, P.A., Laycock, P.J.: Estimation of extremes in corrosion engineering. J. Appl. Stat. **23**(6), 621–644 (1996)

Scarrott, C., MacDonald, A.: A review of extreme value threshold estimation and uncertainty quantification. REVSTAT-Stat. J. **10**(1), 33–60 (2012)

Shaffer, J.P.: Multiple hypothesis testing. Annu. Rev. Psychol. **46**(1), 561–584 (1995)

Shang, H., Yan, J., Zhang, X.: El Niño–Southern Oscillation influence on winter maximum daily precipitation in California in a spatial model. Water resour. Res. **47**, W11,507—W11,515 (2011). doi:10.1029/2011WR010415

Singh, V.P.: Entropy Theory and Its Application in Environmental and Water Engineering. Wiley, Oxford (2013)

Singo, L.R., Kundu, P.M., Odiyo, J.O., Mathivha, F.I., Nkuna, T.R.: Flood frequency analysis of annual maximum stream flows for Luvuvhu river catchment, Limpopo Province. University of Venda, Department of Hydrology and Water Resources, South Africa. Technical report (2012)

Smith, R.L.: Maximum likelihood estimation in a class of nonregular cases. Biometrika **72**(1), 67–90 (1985)

Smith, R.L.: Extreme value theory based on the *r* largest annual events. J. Hydrol. **86**(1), 27–43 (1986)

Tawn, J.A.: An extreme-value theory model for dependent observations. J. Hydrol. **101**(1), 227–250 (1988)

Weissman, I.: Estimation of parameters and large quantiles based on the *k* largest observations. J. Am. Stat. Assoc. **73**(364), 812–815 (1978)

Zhao, X., Scarrott, C.J., Oxley, L., Reale, M.: GARCH dependence in extreme value models with Bayesian inference. Math. Comput. Simul. **81**(7), 1430–1440 (2011)