doi.org/10.26398/IJAS.0029-001

STATISTICAL STUDIES OF THE BETA GUMBEL DISTRIBUTION: ESTIMATION OF EXTREME LEVELS OF PRECIPITATION

Fredrik Jonsson, Jesper Rydén[1]
Department of Mathematics, Uppsala University, Uppsala, Sweden

Abstract *Generalisations of common families of distributions are of interest in their own right as well as for applications. A Beta Gumbel distribution has earlier been introduced as a generalisation of the Gumbel distribution, suggesting that this would provide a more flexible tail behaviour compared to the Gumbel distribution. Through simulation studies, the distributions are here compared closer, e.g. with respect to estimation of quantiles. Moreover, real data in the form of extreme rainfall are analysed and assessment is made whether the proposed Beta Gumbel distribution can be superior to the standard distributions with respect to modelling tail behaviour. Estimates of return values corresponding to return periods of lengths from 100 up to 100 000 years are found, as well as the related confidence intervals. The distribution considered does indeed provide more flexibility, but to the price of computational issues.*

Keywords: *Beta distribution, Beta Gumbel distribution, Gumbel distribution, Quantiles, Rainfall.*

1. INTRODUCTION

Statistical modelling of extreme values is of importance in many fields of science and technology, e.g. in environmental applications to yearly maximal temperatures, river discharges etc. A limiting distribution for the maximum of many of the most common families of distributions is the Gumbel distribution, which is a special case of the so-called Generalised Extreme Value (GEV) distribution. In many typical applications, the tail behaviour is of particular interest, for instance when estimating quantiles. Hence, flexibility is desirable, and generalisations have been suggested, as the GEV distribution is obtained in the limit. For a practical situation with occasionally a limited amount of observations, generalisations of the Gumbel distribution may therefore be of interest (Pinheiro and Ferrari, 2016).

Generalisation of distributions has been discussed frequently. A generalised class of the Beta distribution was first given by Eugene, Lee and Famoye (2002),

[1] Corresponding author, e-mail: jesper.ryden@math.uu.se

where the Beta Normal distribution was introduced as a generalisation of the Normal distribution. Compared to the classical Normal distribution, this generalisation rendered greater flexibility of the shape of the distribution. Turning to the Gumbel distribution, Nadarajah and Kotz (2004) introduced the Beta Gumbel distribution and claimed that this allows for greater flexibility when explaining the variability of the tail compared to the Gumbel distribution. Several mathematical properties of the distribution were presented, such as moments, asymptotic results and estimation issues. However, no applied example was presented, and the present article has as its main aim to further discuss the intended flexibility of the Beta Gumbel distribution, exploring it through simulation studies. Moreover, a case study within environmental statistics is performed: estimation of return values for measurements of precipitation.

The paper is organised as follows. In Section 2, a brief introduction to statistical extreme-value analysis is given, including presentation of the Gumbel and Beta Gumbel distribution. In Section 3, we review estimation issues, in particular estimation of return levels, directly related to quantiles in the extreme-value distribution. Results from simulation studies are presented in Section 4, involving estimation of quantiles, while a real data set with annual maximum daily rainfall for two locations in Sweden is analysed in Section 5. Finally, conclusions are given in Section 6.

2. EXTREME-VALUE DISTRIBUTIONS

In this section, we first review the basic assumptions and notions from classical extreme-value theory. In light of this, the Beta Gumbel distribution is then presented.

2.1. LIMITING DISTRIBUTIONS IN EXTREME-VALUE ANALYSIS

In the classical extreme-value analysis, focus is on the quantity

$$M_n = \max(X_1, \ldots, X_n)$$

where $X_1, \ldots, X_n$ is a sequence of independent and identically distributed random variables from some distribution F. Under suitable conditions, the distribution of M_n can be approximated for large values of n. This asymptotic result, sometimes named the extremal types theorem, states that the distribution of M_n belongs to a single family of distributions, regardless of the unknown distribution F. See

e.g. Coles (2001) or Beirlant et al (2004) for a thorough presentation, including historical developments.

The extremal types theorem states that if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$\mathsf{P}((M_n - b_n)/a_n \leq x) \to G(x) \quad \text{as } n \to \infty,$$

where G is a non-degenerate distribution function, then the distribution G belongs to one of the following three families of distributions: Gumbel, Fréchet and Weibull, respectively. These are occasionally called extreme-value distribution of type I, II and III, respectively. These families of distributions can be combined into one single family called the Generalised Extreme Value (GEV) distribution. The distribution function of the GEV distribution has the following form:

$$G(x) = \exp\left\{-\left[1+\xi\left(-\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{1}$$

where x is defined for $1+\xi(x-\mu)/\sigma > 0$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$, where μ is a location parameter, σ a scale parameter, and ξ a shape parameter. For $\xi = 0$, Eq. (1) is undefined and then the limit is the distribution function

$$G(x) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}, \quad -\infty < x < \infty \tag{2}$$

with two parameters. The distribution function in Eq. (2) corresponds to the Gumbel family of distributions, or the extreme-value distribution of type I. Several common distributions belong to the Gumbel domain of maximum, for instance Weibull, exponential, Gamma, normal, lognormal.

2.2. GENERALISATIONS OF THE GUMBEL DISTRIBUTION

In this subsection, we present the Beta Gumbel distribution and discuss briefly its extremal properties. A recent comparative review of generalisations of the Gumbel distribution is made by Pinheiro and Ferrari (2016).

THE BETA GUMBEL DISTRIBUTION

Nadarajah and Kotz (2004) introduced a generalisation of the Gumbel distribution, the Beta Gumbel (BG) distribution, in hope that this would attract greater applicability in engineering. By adding two parameters, a and b, which mainly

control the skewness and kurtosis, it allows the BG more flexibility in modelling the tail behaviour compared to the Gumbel distribution.

The density function of the BG distribution is given by

$$f(x) = \frac{1}{\sigma B(a,b)} u \mathrm{e}^{-au} \left[1 - \mathrm{e}^{-u}\right]^{b-1} \quad -\infty < x < \infty, \tag{3}$$

where $u = \exp\{-(x-\mu)/\sigma\}$ and $-\infty < \mu < \infty$, $\sigma > 0$, $a > 0$, $b > 0$. Here $B(a,b)$ is the beta function:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t.$$

In this paper, we denote the BG distribution as $\mathrm{BG}(\mu, \sigma, a, b)$. Further discussion on the BG distribution, including its background and estimation issues, is found in Appendix.

In Figure 1, the density function of the BG distribution is plotted for different values of the parameters. For all curves, $\mu = 0$, $\sigma = 1$ and the influence of a and b is investigated. The solid curve corresponds to a Gumbel distribution ($a = b = 1$). An interpretation could be that the parameter b is sensitive in terms of the skewness of the density curve; the lower the parameter, the higher the skewness. This was also noted by Nadarajah and Kotz (2004) who show that a low value of b rapidly amplifies the skewness and kurtosis of the density function.

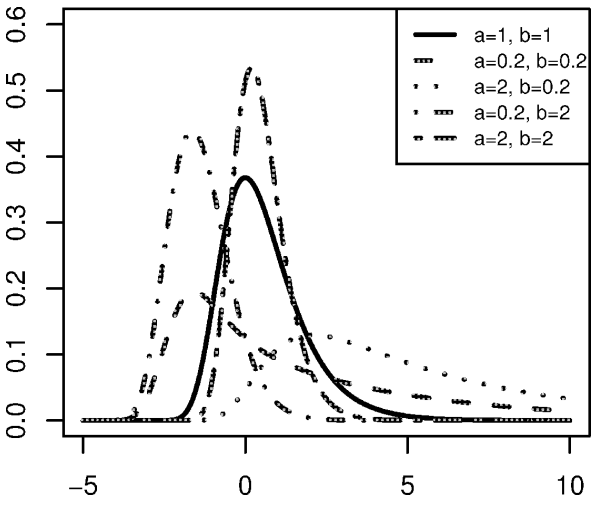

Figure 1: The density function of the BG distribution for different values of *a* and *b* with $\mu = 0$ and $\sigma = 1$ fixed.

THE EXPONENTIATED GUMBEL DISTRIBUTION

Another way of generalising the Gumbel distribution is to consider so-called exponentiated distributions. This was suggested by Nadarajah (2006) and is defined as

$$F(x) = 1 - \left[1 - \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}\right]^{\alpha}, \quad -\infty < x < \infty,$$

where $\alpha > 0$, $\sigma > 0$ and $-\infty < \mu < \infty$. Hence, compared to the standard Gumbel, an extra parameter $\alpha > 0$ is introduced. Moreover, when $a = 1$ in the BG distribution, this reduces to the exponentiated Gumbel (Pinheiro and Ferrari, 2016). An application to observations of significant wave height was studied by Persson and Rydén (2010). A recent review on exponentiated distributions is given by Cordeiro, Ortega and da Cunha (2013). We will, however, not examine the exponentiated Gumbel distribution any closer in the present paper, but focus on examining properties of the BG distribution.

3. ESTIMATION OF T-YEAR RETURN VALUES

In certain applications of extreme-value analysis, for instance in hydrology and reliability engineering, interest is typically in estimation of the T-year return value. This is defined to be the value x_T that will on average be exceeded once over a period of T years (Fernandez and Salas 1999, Rootzén and Katz 2013). The value x_T can be found by solving the equation

$$F(x_T) = 1 - 1/T \tag{4}$$

where F is the cdf. Solving for x_T in Eq. (4) by inverting the cumulative distribution function can sometimes be difficult or impossible if no closed formula exists. For the case of continuous distributions, the inverse of a cdf is usually a well-defined function on $(0,1)$ and an analytical function may sometimes be found.

3.1. CONFIDENCE INTERVALS

Consider the parameter vector $\theta = (\mu, \sigma, a, b)$ and denote its maximum-likelihood estimate (MLE) by $\widehat{\theta}$. One can show that under suitable regularity conditions as n is large, $\widehat{\theta}$ is asymptotically normally distributed (see e.g. Young and Smith (2005), Chapter 8.4). In some cases, we are interested in estimation of functions of $\widehat{\theta}$, for instance, when confidence intervals for return values are wanted. If the

regularity conditions are satisfied, a result with use of Taylor's formula enables us to find estimation errors of functions of the MLE (see e.g. Rychlik and Rydén, 2006). This method is commonly referred to as the delta method, which we will present for the particular case where F belongs to the classical Gumbel distribution.

Employing the inverse of the Gumbel cumulative distribution function yields a point estimate as

$$\widehat{x}_T = \widehat{\mu} - \widehat{\sigma}\ln\left(-\ln\left(1-1/T\right)\right), \quad T > 1$$

where the MLEs of μ and σ are found by solving through iterative methods

$$\begin{aligned} \sigma &= \bar{x} - \frac{\sum_{i=1}^{n} x_i \mathrm{e}^{-x_i/\sigma}}{\sum_{i=1}^{n} \mathrm{e}^{-x_i/\sigma}}, \\ \mu &= -\sigma \ln\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{e}^{-x_i/\sigma}\right). \end{aligned}$$

A related standard error is found through the variance

$$v = \mathbf{v}^{\mathrm{T}}\mathbf{C}\mathbf{v}$$

where

$$\mathbf{v} = \nabla x_T = \left[\frac{\partial}{\partial\mu}x_T(\mu,\sigma),\ \frac{\partial}{\partial\sigma}x_T(\mu,\sigma)\right]^{\mathrm{T}}$$

and $\mathbf{C}$ is the covariance-variance matrix evaluated at $(\widehat{\mu},\widehat{\sigma})$, obtained as the so-called observed information matrix, involving second-order derivatives of the log-likelihood function. Employing quantiles from the standard normal distribution, a confidence interval can finally be constructed.

However, since there is no analytical formula for the quantile function of the BG distribution, the delta method cannot be applied. Instead, we will use resampling to estimate standard errors to find approximate confidence intervals. The delta method (for Gumbel distribution) and resampling techniques (for BG) are thus used in the sequel.

4. SIMULATION STUDIES

In this section we will perform simulations to further explore the BG distribution, in particular, possible interpretations of the parameters a and b. Comparison will be made with the Gumbel distribution (where $a = b = 1$). Furthermore, we inves-

tigate the impact on estimation of return values, and their associated uncertainties. The computations were performed using R, version 3.1.2, with the packages `evd` (Stephenson, 2002) and `lmomco` (Asquith, 2016). To find the MLE we used the routine `optim` and the BFGS optimisation algorithm. Supplying the exact gradient function of the log-likelihood function to `optim` did not always render the maximum of the log-likelihood function but local extreme points. We decided to let the BFGS method approximate the gradient by numeric approximation.

4.1. PARAMETER ESTIMATION

Based on a sample of random numbers generated from the classical Gumbel distribution with chosen parameters μ and σ, the MLE for a BG distribution can be found.

We simulated $N = 5000$ samples of sample size $n = 100$ random numbers from the Gumbel distribution with fixed location parameter $\mu = 5$. In environmental applications, data sets of yearly observations are seldom longer than a few centuries (often considerably shorter). This motivated our choice of n. The value of μ was chosen as to be a positive real value, no too close to zero. To investigate the behaviour of estimated a and b, when fitting a BG distribution, three cases with varying coefficient of variation were studied, corresponding to low, intermediate and high variability. For a Gumbel distributed random variable X with location parameter μ and scale parameter σ,

$$\mathsf{E}[X] = \mu + \gamma\sigma, \quad \mathsf{V}[X] = \frac{\pi^2}{6}\sigma^2,$$

where Euler's constant $\gamma \approx 0.5772$, and the coefficient of variation, c_v say, follows as

$$c_v = \mathsf{R}[X] = \frac{\sigma\pi}{\sqrt{6}(\mu + \gamma\sigma)}. \tag{5}$$

We can easily solve for σ in Eq. (5), given values of the location parameter μ and the coefficient of variation c_v:

$$\sigma = \frac{c_v\mu\sqrt{6}}{\pi - c_v\gamma\sqrt{6}}.$$

We chose the values of c_v to be 0.2, 0.5 and 0.9, respectively. Results are found in Table 1, where means and standard deviation of the MLE are given. We also give robust alternatives to location and spread in terms of median and median absolute deviation (MAD), since some simulations resulted in parameter estimates that could be considered outliers.

From Table 1, we may note that the parameter estimates $\hat{b}$ seem closer to the value one compared to the estimates $\hat{a}$. Using the robust measures, the medians of b estimates are even closer to one in each of the three situations. We might conclude, in light of these simulated observations, that the parameter a represents the flexibility due to the BG.

Comparing the three situations of variability, it seems natural that an increasing value of c_v would result in a larger uncertainty of parameter estimates. Indeed, from Table 1, this is true for the parameters μ and σ. For the parameter a, though, the standard deviation of the estimates is decreasing with increasing c_v. Also the MAD measure decreases. For the parameter b, the standard deviation of estimates is increasing with increasing c_v, while a slight decrease is found for the MAD measure.

Table 1: Parameter estimates resulting from simulation from the BG distribution (sample size n=100, N=5000 samples simulated for each of the three choices of *cv*).

	$c_v = 0.2$ ($\mu = 5$, $\sigma = 0.86$)				$c_v = 0.5$ ($\mu = 5$, $\sigma = 2.52$)				$c_v = 0.9$ ($\mu = 5$, $\sigma = 5.90$)			
	$\hat{\mu}$	$\hat{\sigma}$	$\hat{a}$	$\hat{b}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{a}$	$\hat{b}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{a}$	$\hat{b}$
Mean	4.16	0.89	3.38	1.25	4.04	2.66	1.97	1.27	4.03	6.21	1.47	1.26
Standard dev.	0.61	0.35	1.88	1.11	1.69	1.04	1.53	1.22	2.79	2.37	1.09	1.32
Median	4.02	0.84	3.37	0.98	4.03	2.47	1.40	0.99	4.98	5.79	1.19	0.99
MAD	0.52	0.27	1.96	0.50	1.34	0.74	0.87	0.49	1.66	1.67	0.46	0.47

4.2. QUANTILE ESTIMATION

In this subsection we discuss the behaviour of the quantiles of the BG and the Gumbel distribution. Regarding the BG distribution, a closed-form expression is not available for finding standard errors (see below).

As mentioned earlier, resampling techniques were used to find approximate confidence intervals (for the BG). For the Gumbel distribution, estimates of the quantiles and the corresponding standard errors were found by the delta method as discussed earlier.

The BG distribution can be written as a composition of functions. To see this, express the BG as a composed function where $F(X) = F_{\text{Beta}}(G(X))$, where G is the cdf of the parental distribution function. To find a random variable X using the uniform distribution U, it suffices to solve $X = F^{-1}(U)$:

$$F_{\text{Beta}}(G(X)) = U \;\Leftrightarrow\; G(X) = F^{-1}_{\text{Beta}}(U) = B \;\Leftrightarrow\; X = G^{-1}(B)$$

where the inverse of the Gumbel distribution is $G^{-1}(x) = \mu - \sigma \ln[-\ln(x)]$. With the same argument, and using that the distribution function of the Beta Gumbel

is right-continuous and strictly increasing on $p \in (0,1)$ for $F^{-1}(p)$, the quantile function of BG is

$$Q_{\mathrm{BG}} = \mu - \sigma \ln\{-\ln[Q_{\mathrm{Beta}(a,b)}(p \mid a,b)]\} \tag{6}$$

for $p \in (0,1)$, where $Q_{\mathrm{Beta}(a,b)}(p \mid a,b)$ is the quantile function of the Beta distribution, with $p = 1-q, q = 1/T$. Note that the quantile function of the Beta distribution must be calculated numerically.

The simulation was carried out as follows. We chose to simulate from a Gumbel distribution with parameters $\mu = 20$, $\sigma = 5$, the choices of parameter values guided by the application to study later. A *single* sample of 100 observations was generated from the Gumbel distribution, and resampling was thereafter performed, generating 5000 bootstrap samples.

To study the behaviour of the return values of both BG and Gumbel for longer return periods, we chose return periods from 100 up to 10 000 years and computed the corresponding confidence intervals. The results are shown in Figure 2, where it can be seen that (not surprisingly) the BG distribution has wider confidence intervals (a consequence of more parameters in the distribution). On the other hand, the point estimates of return values seem to be quite the same. Thus, based on these simulations, the Gumbel distribution would be preferred.

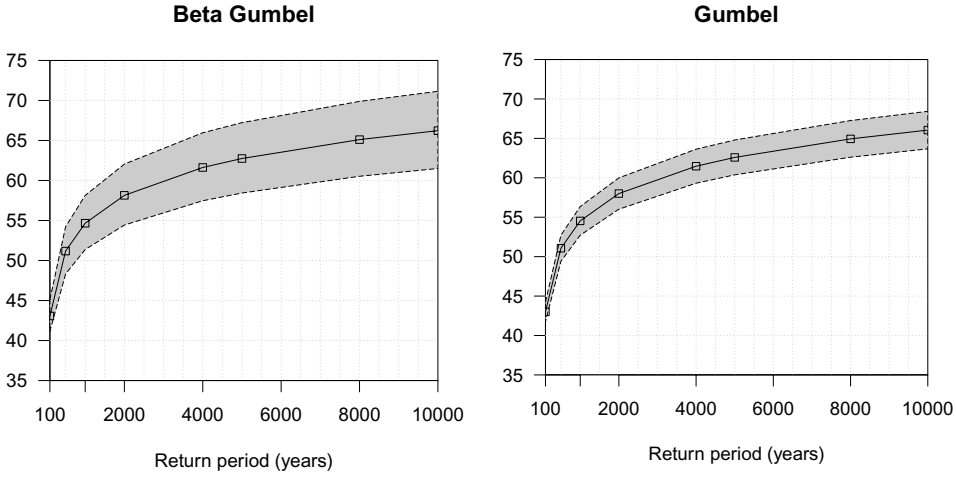

Figure 2: Return values and confidence intervals for selected return periods from 100 up to 10 000 years. Left panel: BG distribution. Right panel: Gumbel distribution.

5. ANALYSIS OF DATA SET: DAILY RAINFALL

To investigate the applicability for modelling real data, we will study annual maximum daily rainfall in Sweden at two different locations: Stockholm and Härnösand. The series of annual maximum daily rainfall covers the period from 1961 to 2011 and was retrieved from a website[2] which provides weather data with courtesy of SMHI, Swedish Meteorological and Hydrological Institute. Stockholm and Härnösand are located in areas of Sweden where some of the most extreme rainfall events (defined as at least 90 mm precipitation during 24 hours) have occurred, especially the latter one[3].

5.1 NOTES ON MEASUREMENTS

Precipitation can be measured in two main ways: either at a fixed geospatial point location (say, a weather station) or over a geographical region, by collecting data from numerous weather stations scattered around a large area and then picking the most extreme record.

Measuring the amount of rainfall is done by rain gauges which gather and measure the accumulated amount of liquid over a specific period of time. Due to limitations, the amount of precipitation cannot be measured accurately. During hurricanes or windy weather it is difficult to gather the rainfall which leads to under-estimation of the precipitation. Moreover, any evaporation will reduce the amount of measured precipitation. In numbers, the total under-estimation is on average of 5–10 %, see Wern (2012).

In winter any snow gathered by the instrument will be melted and the melted water is measured. For definitions on how precipitation is measured, see Wern (2012).

5.2 INTRODUCTORY ANALYSIS

In Figure 3, the time series of annual daily maxima are plotted for Stockholm and Härnösand, respectively. By visual inspection, Härnösand seems to have on average a higher annual maximum daily rainfall. (From data, we find the means 31.7 mm and 42.0 mm, respectively.) Furthermore, no apparent trend is visible. To investigate possible dependence between observations in each sequence, sam-

---

[2] http://www.hurvarvadret.se

[3] Extrem punktnederbörd, (2015, 14th of August). Retrieved December 20, 2016, from http://www.smhi.se/kunskapsbanken/meteorologi/extrem-punktnederbord-1.23041

ple autocorrelation plots up to lag 15 are shown in Figure 4. Most values fall within the confidence limits and there is thus no major concern of dependence. The ACF of Härnösand shows a cut-off at lag 4, but the dependence seems overall weak.

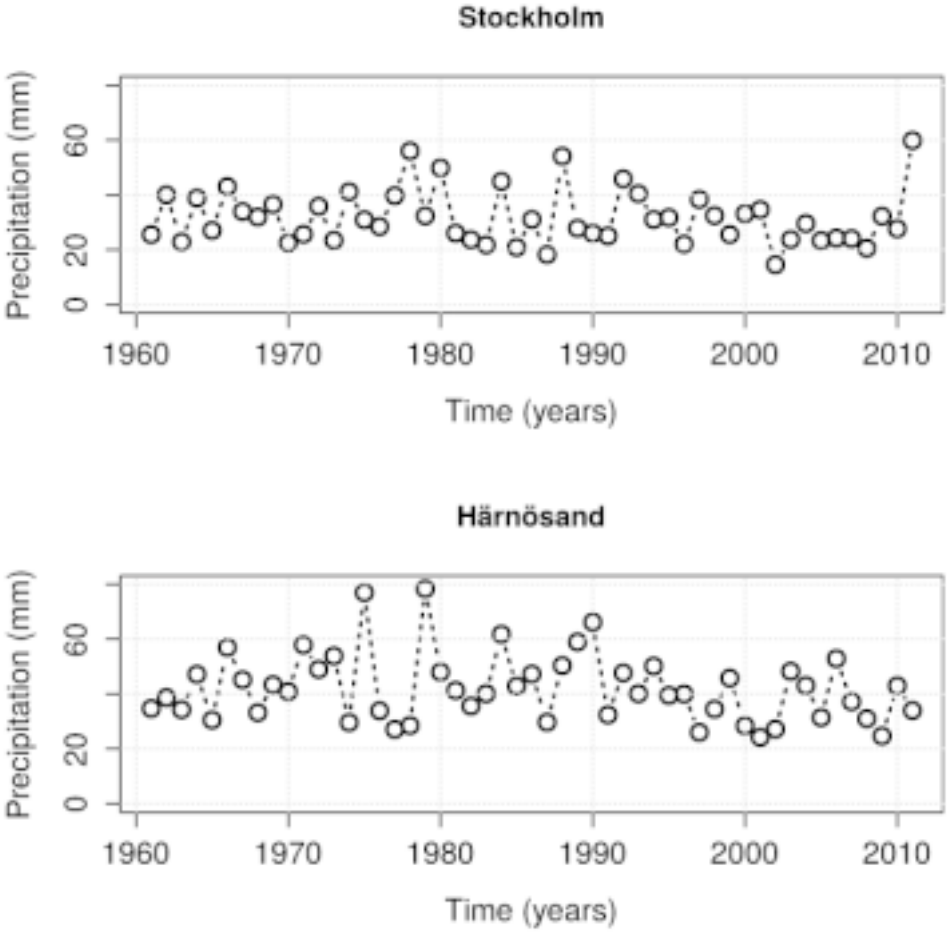

Figure 3: Annual maximum daily rainfall records in Stockholm (top) and Härnösand (bottom).

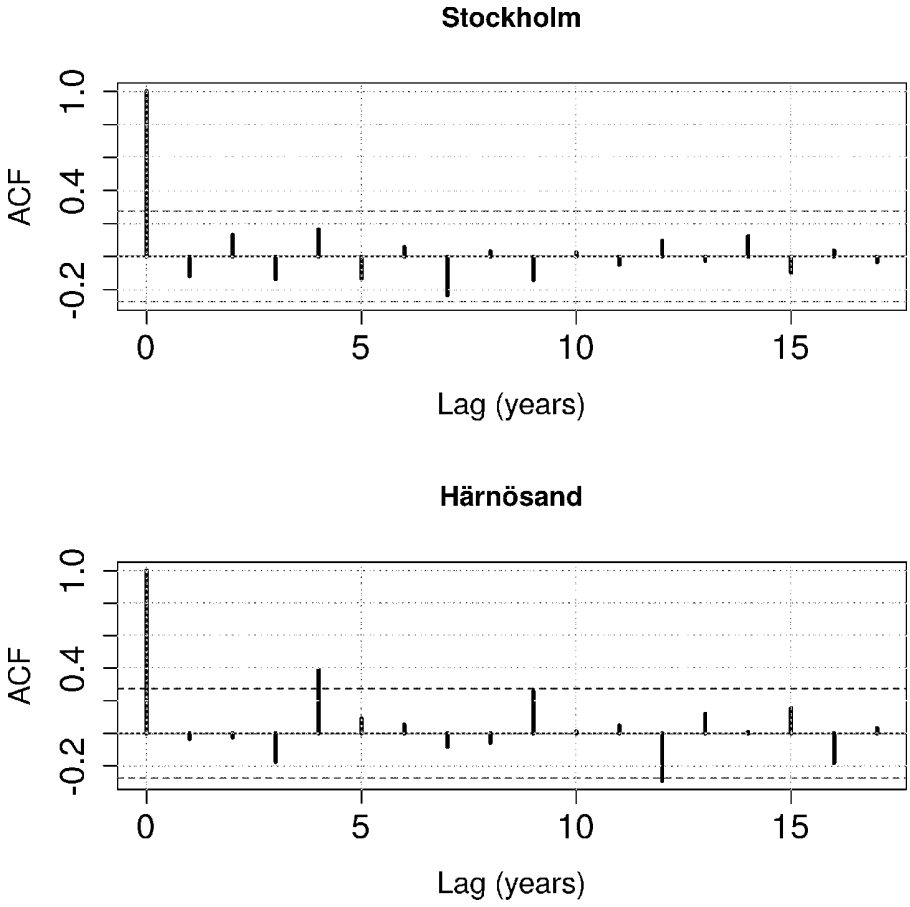

Figure 4: Sample autocorrelation functions for both datasets: Stockholm (top) and Härnösand (bottom).

5.3 ANALYSIS OF FITTED DISTRIBUTIONS

An important problem in statistical methodology of today is check of model assumptions and, if several models are possible, model choice. We first investigate the fit of the BG distribution to the two datasets by graphical means. In Figure 5 the empirical distribution and the fitted BG model are plotted. For both locations, our model agrees reasonably well with the empirical cdf. We also provide the QQplot in Figure 6, and notice no apparent departures from the straight line except at a few points for dataset 2 (Härnösand). From these plots, the BG distribution seems to be a plausible model.

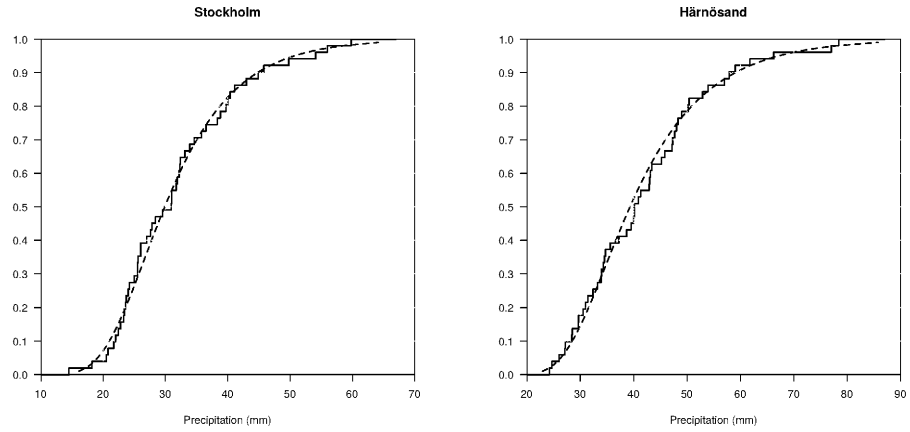

Figure 5: Empirical distribution versus fitted distribution functions for BG, Stockholm (left) and Härnösand (right).

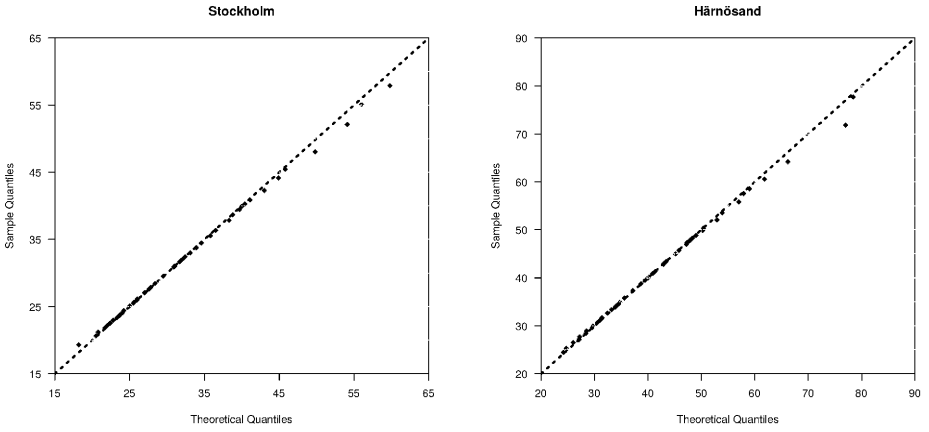

Figure 6: QQ-plots for the data sets, Stockholm (left) and Härnösand (right).

We now turn to the problem of model choice, comparing BG to other candidate models such as the Gumbel distribution and the GEV, and we will then perform likelihood-ratio tests, and also investigate using the Akaike information criterion (AIC).

Table 2: The maximum log-likelihood values for each distribution.

	$\ln L(\hat{\theta})$
Stockholm	
BG	–184.0948
GEV	–184.1427
Gumbel	–184.1654
Härnösand	
BG	–195.9717
GEV	–196.1058
Gumbel	–196.1798

Likelihood-ratio tests

The maximum log-likelihood values for each distribution are given in Table 2. For these data sets we had to use the Nelder–Mead algorithm for optimising the log-likelihood function of the BG distribution. Note that the highest log-likelihood value is obtained by fitting the BG. However, the differences in the log-likelihood values between the distributions are very small. We can use the log-likelihood-ratio test to check whether one higher-order parameter model describes the variability significantly better. A log-likelihood statistic is $D = 2[\log(M_1) - \log(M_0)]$ where M_0 is a reduction of the model M_1. The statistic D is chi-square distributed with $p-k$ degrees of freedom, where p and k are the dimensions of the parameter space of M_1 and M_0, respectively. The null hypothesis is rejected if $D > \chi^2_{p-k}$, favouring the M_1 model which describes the variability of the data significantly better.

We consider two situations. Comparing the Gumbel distribution to the GEV distribution is equal to testing

$$H_0 : \xi = 0 \qquad \text{against} \quad H_1 : \xi \neq 0.$$

We can also test the Gumbel against the BG distribution since it is a reduction of the BG distribution of the parameter space $a \times b$

$$H_0 : a = 1,\ b = 1 \qquad \text{against} \quad H_1 : a \neq 1,\ b \neq 1$$

and hence a χ^2 distribution with $4 - 2 = 2$ degrees of freedom. In Table 3, the values of the observed test statistics are given, along with p-values (obtained via $\chi^2(1)$ and $\chi^2(2)$ distributions, respectively).

Table 3: Values of observed test statistics and related p-values.

Model comparison	Location	D	p-value
Gumbel vs. GEV	Stockholm	0.045	0.83
	Härnösand	0.15	0.70
Gumbel vs.	BG Stockholm	0.14	0.93
	Härnösand	0.42	0.81

From Table 3, we note that for all situations of model comparison and at all locations, the hypothesis of the simpler model (i.e. Gumbel) cannot be rejected. From a modelling perspective, the Gumbel distribution seems adequate.

Akaike Information Criterion (AIC)

In general, it is not desirable to use too complicated models; frequently, the simplest model is most likely to be correct, and one can test whether the more complicated model explains the variability significantly better. To test whether one model with a higher number of parameters models the data significantly better than another candidate model with a lower number of parameters, we can use the Akaike information criterion, a test statistic that penalises over-fitting. The test statistic is given by $\text{AIC} = -2\ln L(\hat{\theta}) + 2p$, where p is the number of model parameters.

Table 4: AIC values and parameter estimates of the different distributions.

		Parameter estimate				
	AIC	$\hat{\mu}$	$\hat{\sigma}$	$\hat{a}$	$\hat{b}$	$\hat{\xi}$
Stockholm						
BG	376.19	16.53	6.44	3:89	0:77	
GEV	374.29	27.19	7.51			0:023
Gumbel	372.33	27.28	7.57			
Härnösand						
BG	399.94	19.17	6.50	6:43	0:56	
GEV	398.21	36.16	9.36			0:048
Gumbel	396.36	36.41	9.55			

In Table 4, AIC values and parameter estimates are presented. We first discuss AIC values, and note that at both locations, the Gumbel alternative has the lowest AIC and should be preferred. Comparing BG and GEV, for both locations GEV has the smaller AIC and should be an option rather than BG. Turning now to parameter estimates of the shape parameter ξ, we observe that the estimate for

GEV has a quite low value at both locations. To test whether ξ is significantly nonzero, we use that the MLE is asymptotically normally distributed. The standard errors of the estimates are 0.13 (Stockholm) and 0.11 (Härnösand), resulting in two-sided p-values 0.83 and 0.71, respectively. Hence, for both data sets, the null hypothesis of $\xi = 0$ cannot be rejected. Therefore, one may argue that the Gumbel distribution models the data equally well. A reduction would be preferable here, but for the sake of comparison we will keep the GEV in the sequel.

5.4. ESTIMATION OF RETURN VALUES

In this subsection, we present estimated return values for each distribution and the corresponding confidence intervals. The confidence intervals of the GEV and Gumbel distribution were derived as usual with the delta method. To find approximate confidence intervals of the BG distribution, we used resampling methodology. Both samples were resampled 2000 times.

Table 5: Return level estimates and corresponding confidence intervals.Return levels with 95% C.I.

	Return levels with 95% C.I.		
	T = 100 (x100)	T = 500 (x500)	T = 1000 (x1000)
Stockholm			
BG	64.3 (51.2, 78.0)	77.8 (58.5, 97.6)	83.6 (61.7, 106.0)
GEV	63.6 (46.2, 81.0)	77.3 (45.4, 109.2)	83.3 (43.6, 123.1)
Gumbel	62.0 (53.6, 70.6)	74.3 (63.2, 85.4)	79.5 (67.4, 91.7)
Härnösand			
BG	85.3 (66.5, 105.9)	103.4 (74.6, 133.6)	111.2 (77.9, 145.7)
GEV	84.4 (57.6, 111.1)	104.0 (52.1, 155.8)	112.9 (47.0, 178.8)
Gumbel	80.3 (69.5, 91.1)	95.7 (81.6, 109.8)	102.3 (86.8, 117.9)

From Table 5, we note regarding point estimates that both GEV and BG distributions render higher estimates of the return values compared to the Gumbel distribution; especially for the higher 1000-year return period. The estimates do not differ largely from a practical point of view. Moreover, we see that GEV and BG are more conservative in estimating the return value. This is valid for both datasets.

Regarding confidence intervals, when comparing GEV and BG, GEV has wider intervals. Note, though, that different methods were employed (delta method vs. resampling). In any case, the confidence intervals are wide from an applied point of view.

5.5 LONGER RETURN PERIODS

Next, we investigate the behaviour of longer return periods for the different distributions.

Estimation of longer return periods is only meaningful if the assumption of stationarity is valid, but still, as a risk measure, the notion of return periods of up 10 000 years is useful. For instance, in dike design in the Netherlands (Botzen et al 2009), the 10 000-year return flood level is used. It could be mentioned that Rootzén and Katz (2013) propose a notion of design life level in order to quantify risk in a changing climate.

In Figures 7 and 8, estimates of the return values for varying return periods are presented, and we note e.g. that the Gumbel distribution gives the lowest estimates (cf. the findings in Table 5). Moreover, the GEV renders higher estimates compared to the BG distribution, which are higher the longer the return period.

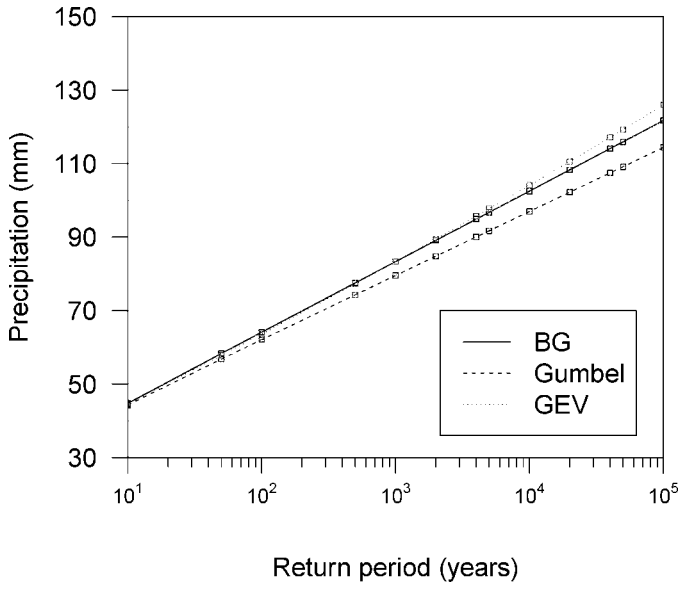

Figure 7: Return values for Stockholm.

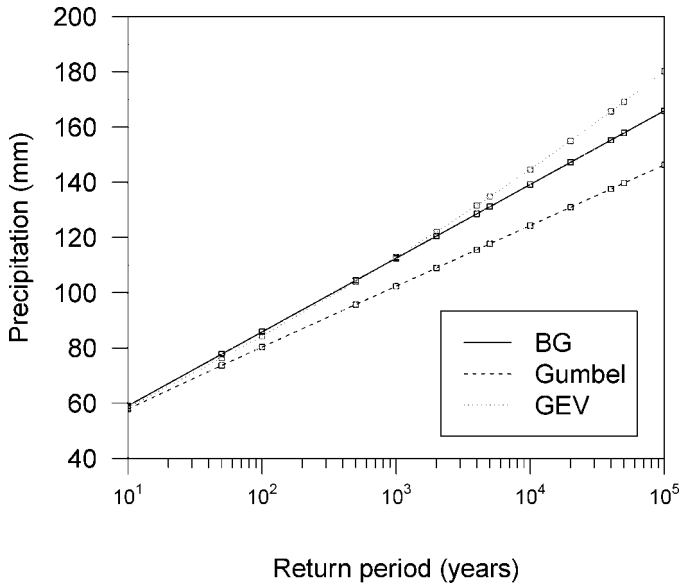

Figure 8: Return values for Härnösand.

5.6 REMARKS ON NUMERICAL COMPUTATIONS

When applying the BG to real data, we had problems finding the standard errors. This was related to the calculation of the inverse of the BG distribution function (i.e. the quantile function). The formula for the quantiles given in Eq. (6) involves the inverse of the Beta distribution function which has to be calculated numerically. After resampling the real data 2000 times, thus yielding 2000 parameter sets of resampled data, we encountered problems in calculating the inverse of the Beta distribution. More precisely, for some parameter sets, it was difficult to find quantiles q, $F_{\text{Beta}}(q) = p$, for p close to 1 since the formula was highly sensitive to precision errors giving us indefinite quantiles.

As noted, the inverse is injective only on (0,1) and for some parameter sets, the part of the formula involving the inverse of the Beta distribution yielded us a value of 1.0, whenever we tried to find quantiles for p close to 1.0. This is of course not well-defined and gives quantiles that are indefinite (infinite).

All numerical computations were done in R which uses finite-precision arithmetic which basically means that about up to 16 digits are correct (or accurate) in the computations. We therefore had to rely on an alternative software that uses arbitrary-precision arithmetic to do the computations, meaning that any number of precision of digits can be used. When using Mathematica (a computer algebra

system) we found that a precision of at least 30 up to 39 digits had to be used to perform the computations of the quantiles. (Mathematica was used to produce the estimates in Table 5.) Even with a difference as negligible as 10^{-39} (i.e.$1 - 10^{-39} = 1$), the logarithmic function in the formula in Eq. (6) is not defined at 1. The quantile function also involves computation of a composition of functions (logarithmic function within a logarithmic function). This gives large differences in evaluating the formula (6) for values close to 1.

6. DISCUSSION

We have studied a generalisation of the Gumbel distribution, the Beta Gumbel (BG) distribution, which has two additional parameters that allow for skewness and variability of the tail weights. From simulations, we found some evidence that the BG does indeed provide more flexibility than the Gumbel distribution.

Finding the MLEs for the BG distribution was also in some situations tricky since we were then faced with computational problems. For arbitrary values of a and b, estimates could sometimes not be found (especially with b close to 0). If this is related to the curvature of the four-dimensional function or a matter of numerical issue is unclear. The BG distribution involves the incomplete beta function and also makes it more difficult to work with. The numerical problems related to the Beta distribution have been pointed out e.g. by Cordeiro and Castro (2011), where another generalisation of the Gumbel distribution was presented, a so-called *Kw*-Gumbel.

We compared the BG to the Gumbel distribution as well as to the GEV distribution when modelling real data. Likelihood-ratio tests as well as comparisons by AIC were made. Since tail behaviour is of interest, the Anderson–Darling test could be applied; however, this would have implied finding critical points for the test statistic with respect to the BG distribution, and this extra work was not performed. It should be noted that BG, which is a four-parameter model compared to the three-parameter model of GEV, makes the numerical work more problematic. Optimisation in four dimensions is highly more difficult than in a threedimensional space. We conclude that for the analysed data, the simpler Gumbel distribution would be a preferred option (Tables 3 and 4).

We cannot conclude whether BG is a better candidate model to use. As has been pointed out by Pinheiro and Ferrari (2016), the BG is non-identifiable. Further work has to be done to study this distribution and its applicability in various fields and contexts, but the numerical obstacles and its non-identifiability suggest that there are other candidate distributions to investigate.

REFERENCES

Asquith, W.H. (2016). lmomco-L-moments, censored L-moments, trimmed Lmoments, L-comoments, and many distributions. R package version 2.2.5, Texas Tech University, Lubbock, Texas.

Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes. Theory and Applications.* Wiley & Sons, Chichester, UK.

Botzen, W.J.W., Aerts, J.C.J.H. and van den Bergh, J.C.J.M. (2009). Dependence of flood risk perceptions on socioeconomic and objective risk factors. *Water Resources Research,* 45: 1-15.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London.

Cordeiro, G. and Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation,* 81: 883-893.

Cordeiro, G.M., Ortega, M.M. and da Cunha, D.C.C. (2013). The exponentiated generalized class of distributions. *Journal of Data Science,* 11: 29-41.

Eugene, N., Lee, C. and Famoye, F. (2002). Beta-Normal Distribution and its applications, *Communications in Statistics – Theory and Methods*, 31:4: 497-512.

Fernandez, B. and Salas, J.D. (1999). Return period and risk of hydrologic events. I: mathematical foundation. *Journal of Hydrologic Engineering,* 4: 297-307.

Morais, A. (2009). *A Class of Generalized Beta Distributions, Pareto Power Series andWeibull Power Series.* M.s. Thesis, Universidade Federal de Pernambuco, Recife-PE.

Nadarajah, S. (2006). The exponentiated Gumbel distribution with climate application. *Environmetrics,* 17: 13-23.

Nadarajah, S. and Kotz, S. (2004). The Beta Gumbel Distribution. *Mathematical Problems in Engineering,* 2004(4): 323-332.

Persson, K. and Rydén, J. (2010). Exponentiated Gumbel distribution for estimation of return levels of significant wave height. *Journal of Environmental Statistics,* 1(3): 1-12.

Pinheiro, E.C. and Ferrari, S.L.P. (2016). A comparative review of generalizations of the Gumbel extreme value distribution with an application to wind speed data. *Journal of Statistical Computing and Simulation,* 86(11): 2241-2261.

Rigby, R.A., Stasinopoulous, D.M., Heller, G. and Voudouris, V. (2014). The distribution toolbox of GAMLSS. www.gtamlss.org

Rootzén, H. and Katz, R.W. (2013). Design life level: Quantifying risk in a changing climate. *Water Resources Research,* 49: 5964-5972.

Rychlik, I., and Rydén J. (2006). *Probability and Risk Analysis: An Introduction for Engineers.* Springer-Verlag, Berlin.

Stephenson, A.G. (2002). evd: Extreme Value Distributions. *R News,* 2(2):31-32.

Wern L., (2012). Extrem nederbörd i Sverige under 1 till 30 dygn, 1900 – 2011. *SMHI Meteorologi,* 143: 5-22.

Young, G.A. and Smith, R.L. (2005). *Essentials of Statistical Inference.* Cambridge University Press, Cambridge.

APPENDIX: THE BETA GUMBEL DISTRIBUTION

THE BETA GUMBEL DISTRIBUTION

We here review the background of the derivation of the BG distribution, following Nadarajah and Kotz (2004). Let G be the cumulative distribution function. Then a generalised class of Beta distribution functions can be defined by

$$F(x) = I_{G(x)}(a,b) \tag{7}$$

where $I_{G(x)}(a,b)$ is the incomplete beta ratio function. In this paper we study the Beta Gumbel distribution in which $G(x)$ belongs to the Gumbel distribution. The generalization in Eq.(7) can be rewritten as

$$I_{G(x)}(a,b) = \frac{B_{G(x)}(a,b)}{B(a,b)}, \quad a>0,\ b>0$$

where $B(a,b)$ is the beta function

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

and $B_{G(x)}(a,b)$ is the incomplete beta function given by

$$B_{G(x)}(a,b) = \int_0^{G(x)} t^{a-1}(1-t)^{b-1}\,\mathrm{d}t, \quad a>0,\ b>0.$$

If we in Eq. (7) let G correspond to the Gumbel distribution, this gives a generalisation of the original (parental) distribution G which we call the Beta Gumbel distribution and denote $\mathrm{BG}(\mu,\sigma,a,b)$. For the special case where $a=1$ and $b=1$ the distribution coincides with the Gumbel distribution. We can now define the probability-density function as

$$\begin{aligned} f(x) := F'(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{B(a,b)}\int_0^{G(x)} t^{a-1}(1-t)^{b-1}\,\mathrm{d}t \\ &= \frac{g(x)}{B(a,b)}G(x)^{a-1}\left[1-G(x)\right]^{b-1} \end{aligned} \tag{8}$$

where $g(x)$ is the density function of the parental distribution. From Eq.(8) it follows that the density function of the Beta Gumbel distribution is given by

$$f(x) = \frac{1}{\sigma B(a,b)}u\mathrm{e}^{-au}\left[1-\mathrm{e}^{-u}\right]^{b-1} \quad -\infty<x<\infty, \tag{9}$$

for $-\infty<\mu<\infty$, $\sigma>0$, $a>0$, and $b>0$, where $u=\exp\{-(x-\mu)/\sigma\}$.

ESTIMATION ISSUES

To find point estimates of the parameters (μ, σ, a, b) of the BG distribution, the method of maximum likelihood is employed. The log-likelihood function was given by Nadarajah and Kotz (2004) and is as follows:

$$\begin{aligned}\ln L(\mu,\sigma,a,b \mid x) = &-n\ln\sigma + (b-1)\sum_{i=1}^{n}\ln\left[1-\exp\left\{-\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right\}\right] \\ &-\sum_{i=1}^{n}\frac{x_i-\mu}{\sigma} - a\sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right) - n\ln B(a,b).\end{aligned} \tag{10}$$

Taking the partial first-order derivatives of Eq. (10) with respect to each parameter, we obtain

$$\begin{aligned}\frac{\partial \ln L}{\partial \mu} =& \frac{n}{\sigma} - \frac{a}{\sigma}\sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right) \\ &+ \frac{b-1}{\sigma}\sum_{i=1}^{n}\frac{\exp(-(x_i-\mu)/\sigma)\exp\{-\exp(-(x_i-\mu)/\sigma)\}}{1-\exp\{-\exp(-(x_i-\mu)/\sigma)\}},\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L}{\partial \sigma} =& -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{x_i-\mu}{\sigma^2}\left\{1-a\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right\} \\ &+ \frac{b-1}{\sigma^2}\sum_{i=1}^{n}\frac{(x_i-\mu)\exp(-(x_i-\mu)/\sigma)\exp\{-\exp(-(x_i-\mu)/\sigma)\}}{1-\exp\{-\exp(-(x_i-\mu)/\sigma)\}},\end{aligned}$$

$$\frac{\partial \ln L}{\partial a} = n\psi(a+b) - n\psi(a) - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\sigma}\right),$$

$$\frac{\partial \ln L}{\partial b} = n\psi(a+b) - n\psi(b) + \sum_{i=1}^{n}\ln\left[1-\exp\left\{-\exp\left(-\frac{x_i-\mu}{\sigma}\right)\right\}\right].$$

where ψ is the digamma function, $\psi(x) = \mathrm{d}\ln\Gamma(x)/\mathrm{d}x = \Gamma'(x)/\Gamma(x)$. Note that $\frac{\partial}{\partial\mu}\ln L$ is slightly different from the one given by Nadarajah and Kotz (2004) which is likely due to a misprint in the original source. Estimates of μ, σ, a and b are found by setting the partial derivatives to zero and solving the subsequent simultaneous equations.

APPENDIX: DATASETS

Table 6: Stockholm data set.

25.5	40.0	22.8	38.8	27.0	43.0	33.9	31.9
36.5	22.4	25.6	35.8	23.4	41.1	30.9	28.4
39.7	56.0	32.3	49.8	26.0	23.6	21.7	44.9
20.8	31.0	18.2	54.1	27.8	26.0	25.0	45.8
40.4	31.0	31.7	22.0	38.3	32.4	25.5	33.1
34.6	14.5	23.7	29.5	23.3	24.2	24.0	20.5
32.2	27.6	59.8					

Table 7: Härnösand data set.

34.7	38.7	34.3	47.2	30.5	57.0	45.2	33.2
43.4	40.9	57.9	49.0	53.9	29.6	77.0	33.9
27.1	28.5	78.4	48.0	41.3	35.6	40.1	61.8
42.9	47.3	29.7	50.4	59.0	66.2	32.4	47.7
40.1	50.2	39.5	40.0	26.0	34.5	45.8	28.4
24.2	27.2	48.3	43.1	31.4	52.9	37.2	31.0
24.7	43.0	34.0					