A METHOD FOR CHOOSING AN OPTIMUM THRESHOLD IF THE UNDERLYING DISTRIBUTION IS GENERALIZED BURR-GAMMA.

A. VERSTER AND D.J. DE WAAL

Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, RSA

e-mail: verstera@ufs.ac.za

Key Words: GBG, GP-type, threshold, tail probabilities, posterior risk, extreme value.

Summary: In this paper the Generalized Burr-Gamma (GBG) distribution is considered to model data that includes extreme values. Very often in extreme value theory the tail of the distribution is the only interest and therefore the selection of an optimum threshold plays an important role. This paper shows that the tail of the Generalized Burr-Gamma (GBG) above an optimum threshold can be approximated by a type of Generalized Pareto distribution (GP) which is dependent on the threshold value. Our GP-type of distribution discuss here differs from the usual GP distribution because it depends on the chosen threshold value and it only has one parameter known as the extreme value index (EVI) and not two as is usually the case. This paper makes a valuable contribution in choosing the best threshold if the underlying distribution is GBG. We assume throughout the paper that the GBG parameters μ, σ, k and ξ are known. A method for choosing the optimum threshold is by considering the posterior risk function as will be explained. The selection of a threshold is illustrated through a simulation study and in the case of a real data set.

1. Introduction

In extreme value analysis, the Peaks over Threshold method became a popular method in predicting high quantiles or estimating tail probabilities. Although parametric models exist to model all the data such as the Burr, Frechét, *t*, *F* and others, the generalized Burr-Gamma class is another class of distributions to fit the whole data set, Beirlant *et al.* (1999). The

Generalized Burr-Gamma distribution is fairly flexible since it consists of four parameters and is therefore a popular extreme value distribution to fit to extreme data. If we assume that the data is Generalized Burr-Gamma distributed we can show that the distribution of the tail, above an optimum threshold, can be approximated with a Generalized Pareto-type of distribution consisting of only one parameter, the EVI. This is a useful approximation since the tail of the distribution is often the only interest. This approximated GP-type of distribution is dependent on the threshold *t* and therefore a valuable contribution can be made in selecting an optimum threshold. The best threshold is chosen in this paper by considering the posterior risk function. Throughout the paper we will assume the four parameters of the GBG are know, if the parameters are not know it should first be estimated, Verster and De Waal (2009).

The layout of this paper is as follows: Section 2 gives a brief introduction on the Generalized Burr-Gamma distribution. Section 3 gives a theorem on the approximation of the tail of the GBG. Section 4 gives a theorem on how to estimate tail probabilities from the approximated GP-type. In Section 5 we discuss a method for selecting the best threshold and Section 6 gives a practical application on how to select an optimum threshold by considering a real data set.

2. The Generalized Burr-Gamma Distribution (GBG)

The GBG is a fairly flexible distribution which contains four parameters, k, μ, σ, ξ, where ξ is known as the extreme value index. If $\xi = 0$ then μ is the mean of $Y = -\log(X)$, where X is GBG distributed. Similarly if $\xi = 0$ then σ is the standard deviation of Y.

The GBG distribution models all the data, also the data in the tail and is given as follows: (Beirlant *et al*. 1999). A random variable X is GBG(k, μ, σ, ξ) distributed when the distribution function is given by

$$F(x) = P(X \le x) = \frac{1}{\Gamma(k)} \int_0^{\upsilon_\xi(x)} e^{-u} u^{k-1} du \qquad (1)$$

where

$$\upsilon_{\xi}(x)=\frac{1}{\xi}\log\left(1+\xi\nu(x)\right)>0 \tag{2}$$

and

$$\nu(x)=e^{\left\{\psi(k)+\frac{\log x+\mu}{\sigma}\sqrt{\psi'(k)}\right\}}$$

for

$$1+\xi v(x)>1.$$

$\psi(k)=\frac{\partial}{\partial k}\log\Gamma(k)$ and $\psi'(k)=\frac{\partial}{\partial k}\psi(k)$ represent the digamma and trigamma functions respectively.

The parameter space is defined as $\Omega=\left\{-\infty<\mu<\infty,\sigma>0,k>0,-\infty<\xi<\infty\right\}$.

A very important characteristic shown by Beirlant *et al*. (1999, p. 115) is that $V_{\xi}\sim\mathrm{Gam}(k,1)$.

3 Approximating the tail of the GBG

Because the tail is often the only interest in extreme value theory the following theorem shows an approximation of the GBG tail. The theorem proves that the tail of the GBG above a reasonable high threshold can be approximated though a Generalized Pareto-type of distribution.

Theorem 1.

If $V_{\xi}\sim\mathrm{Gam}(k,1)$ then for a large threshold value (t), $V=e^{\left\{\psi(k)+\frac{\log X+\mu}{\sigma}\sqrt{\psi'(k)}\right\}}>t_v$ is distributed as a Generalized Pareto-type of distribution with survival function given by

$$P\left(V > v \mid V > t_v\right) = \left(1 + \frac{\xi(v - t_v)}{1 + \xi t_v}\right)^{\frac{-1}{\xi}},\ v > 0; \sigma_t, \xi > 0, i = 1, ..., N_t\,. \tag{3}$$

Where t_v is the threshold in terms of V and N_t is the number of observations above the threshold.

Proof.

From (2), $V = \frac{e^{\xi V_\xi} - 1}{\xi}$ now the survival function is given by

$$\begin{aligned} P\left(V > v \mid V > t_v\right) &= P\left(\frac{e^{\xi V_\xi} - 1}{\xi} > v \,\middle|\, \frac{e^{\xi V_\xi} - 1}{\xi} > t_v\right), \\ &= \frac{P\left(V_\xi > \frac{\log\left(\xi(v) + 1\right)}{\xi}\right)}{P\left(V_\xi > \frac{\log\left(\xi t_v + 1\right)}{\xi}\right)}. \end{aligned} \tag{4}$$

Let $\frac{\log\left(\xi v + 1\right)}{\xi} = a$ and $\frac{\log\left(\xi t_v + 1\right)}{\xi} = b,$ then

$$P\left(V > v \mid V > t_v\right) = \frac{\Gamma\left(k, a\right)}{\Gamma\left(k, b\right)} \tag{5}$$

can be expressed as a ratio of two incomplete Gamma functions. The incomplete Gamma functions are given by the integrals

$$\Gamma\left(k, a\right) = \int_a^\infty h^{k-1} e^{-h} dh \text{ and } \Gamma\left(k, b\right) = \int_b^\infty h^{k-1} e^{-h} dh\,. \tag{6}$$

The following equations are approximations of the incomplete gamma functions for large values of *a* and *b*, Amore (2005)

$$\Gamma\left(k, a\right) \approx e^{-a}\left(1 + a\right)^{k-1} \text{ and } \Gamma\left(k, b\right) \approx e^{-b}\left(1 + b\right)^{k-1} \tag{7}$$

Thus equation (4) can be expressed as follows

$$\frac{\Gamma(k,a)}{\Gamma(k,b)} = \frac{(1+\xi v)^{\frac{-1}{\xi}}}{(1+\xi t_v)^{\frac{-1}{\xi}}} \frac{\left\{1+\frac{1}{\xi}\left[\ln(1+\xi v)\right]\right\}^{k-1}}{\left\{1+\frac{1}{\xi}\left[\ln(1+\xi t_v)\right]\right\}^{k-1}}$$
$$= \left(1+\frac{\xi(v-t_v)}{1+\xi t_v}\right)^{-\frac{1}{\xi}} \left(\frac{1+\frac{1}{\xi}\left[\ln(1+\xi v)\right]}{1+\frac{1}{\xi}\left[\ln(1+\xi t_v)\right]}\right)^{k-1}. \quad (8)$$

When using L'Hospital's rule, Salas *et al*. (1999), $\left(\frac{1+\frac{1}{\xi}\left[\ln(1+\xi v)\right]}{1+\frac{1}{\xi}\left[\ln(1+\xi t_v)\right]}\right)^{k-1} \to 1$, as $t_v \to \infty$

Therefore we conclude that for large *t*, $\frac{\Gamma(k,a)}{\Gamma(k,b)} = \left(1+\frac{\xi(v-t_v)}{1+\xi t_v}\right)^{-\frac{1}{\xi}}$,which is the distribution function of the Generalized Pareto distribution with parameter ξ on the exceedances above *t*. Therefore, for large values of the threshold the tail of the GBG distribution can be approximated with a Generalized Pareto-type of distribution with extreme value index ξ. To avoid confusion we will now refer to the extreme value index of the Generalized Pareto-type of distribution as η.

■

The following figures show how the approximated Generalized Pareto-type of distribution (3) fits the data above the threshold when compared to the ratio of the incomplete gamma distributions (8). In Figure 1 a threshold is chosen at *t* = 10, η is chosen as 0.95 and *k* takes on different values between 0.5 and 1.3. In Figure 2 a threshold is again chosen at *t* = 10, but now η is chosen as $\eta = 0.2$ and *k* is again chosen as different values between 0,5 and 1,3. In Figure 3 a threshold is chosen at a larger value *t* = 30, η is chosen again as $\eta = 0.2$ and *k* is chosen again as different values between 0,5 and 1,3.

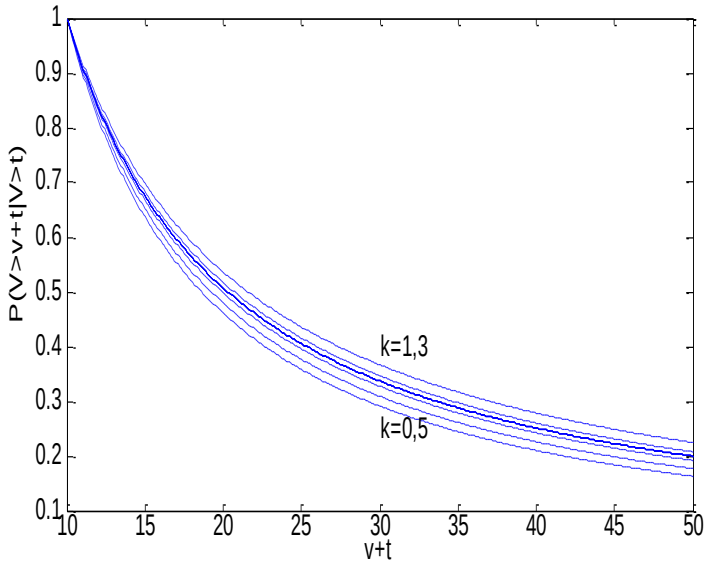

Figure 1 Comparison between the approximated GP-type and the incomplete gamma ratio given in equation 8

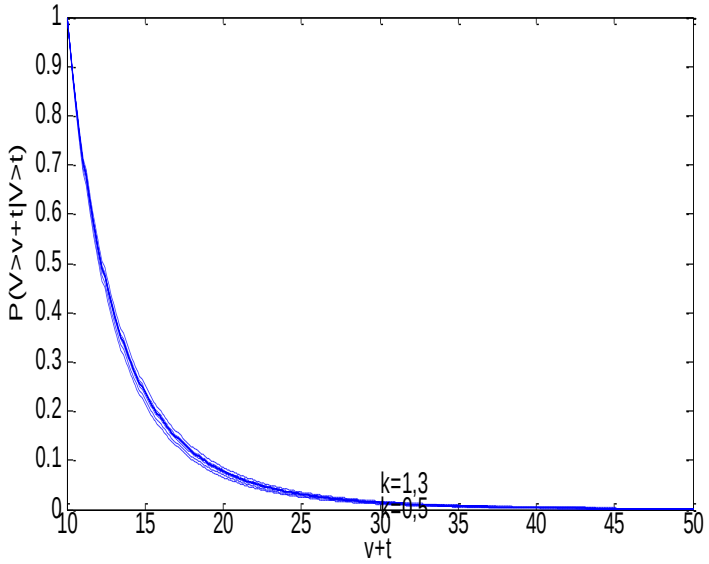

Figure 2 Comparison between the approximated GP-type and the incomplete gamma ratio given in equation 8 with a smaller value of η

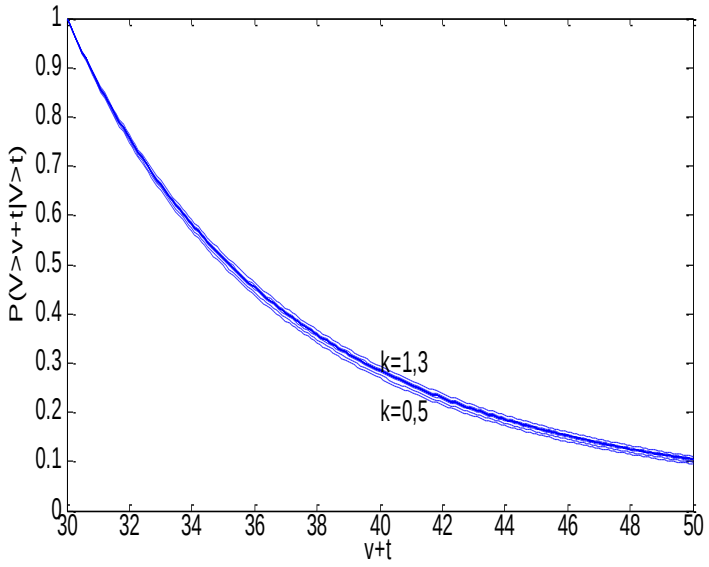

Figure 3 Comparison between the approximated GP-type and the incomplete gamma ratio given in equation 8 with a larger value of *t*

From the above figures it can be seen that, for small values of η**,** the approximated Generalized Pareto-type of distribution follows the ratio of the incomplete gamma distributions more closely. If η becomes large, close to 1, a higher threshold should be chosen to make sure that the second term of equation (8) strives to 1.

4 Estimating tail probabilities

In this section we estimate an approximation for tail probabilities when the data is GBG distributed and the tail of the GBG is approximated with a GP-type of distribution.

Theorem 2.

If $X \sim GBG(\mu, \sigma, k, \xi)$ and $V_\xi \sim GAM(k, 1)$ then the tail probability $P(X > x)$ is estimated by

$$\hat{P}(X > x) = \left(1 + \frac{\eta(v - t_v)}{1 + \eta t_v}\right)^{\frac{-1}{\eta}} \frac{N_t}{n}, x > t, v > t_v, \sigma_t, \eta > 0, i = 1, .., N_t. \quad (9$$

Proof.

$$P(X > x|X > t) = \frac{P(X > x \text{ and } X > t)}{P(X > t)}$$
$$= \frac{P(X > x)}{P(X > t)}$$
$$= \frac{P(V > v)}{P(V > t_v)} \quad (10)$$

therefore

$$P(V > v) = P(V > v|V > t_v)P(V > t_v)$$
$$= P(V > v|V > t_v)P(X > t) \quad (11)$$

The $P(X > t)$ can be estimated by $\frac{N_t}{N}$ and from Theorem 1 $P(V > v|V > t_v) \approx \left(1 + \frac{\eta(v-t_v)}{1+\eta t_v}\right)^{\frac{-1}{\eta}}$, therefore an estimate of $\hat{P}(X > x)$ is

$$\hat{P}(V > v) \approx \left(1 + \frac{\eta(v - t_v)}{1 + \eta t_v}\right)^{\frac{-1}{\eta}} \frac{N_t}{N} \quad (12)$$

■

η is simulated from the posterior of the approximated GP-type of distribution. The posterior distribution is given by

$$\pi(\eta|\boldsymbol{v}) \propto \prod_{i=1}^{N_t} \frac{1}{1+\eta t_v}\left[1 + \frac{\eta(v_i-t_v)}{1+\eta t_v}\right]^{\frac{-1}{\eta}-1} \pi(\eta) \quad (13)$$

where

$$\pi(\eta) \propto \frac{e^{-(\eta)}}{1+\eta t_v}. \quad (14)$$

is the maximal data information (MDI) prior. The MDI prior is derived in Appendix B. Equation (12) shows the estimated approximated tail probability of $P(X > x)$, the true tail probability however can be obtained explicitly by the following equation

$$P(X > x) = 1 - \Gamma\left(V_{\xi}(x), k, 1\right) \tag{15}$$

where $X \sim GBG(\mu, \sigma, k, \xi)$. We expect that the two tail probabilities (equations 12 and 15) should be close to one another at the optimum threshold level. Thus, at the optimum threshold level the difference between the two tail probabilities should be close to zero.

5 Selecting an optimum threshold

Under ideal circumstances with an optimum threshold level, we would expect that the values of ξ and η should be close or the difference between them should be close to zero. At different threshold values the posterior risk function, the expected squared difference between the simulated η values and fixed GBG parameter ξ, can be calculated. The threshold value that minimizes the posterior risk function (Rice, 1995) will be considered as the best threshold to choose. The posterior risk function for a specific threshold is given in the following equation

$$R_t = E_{\eta}\left((\eta_t - \xi)^2 | x_{N-N_t+1}, \dots, x_N\right) \tag{16}$$

where η_t is the simulated values from the posterior at a certain threshold and ξ is the know EVI of the GBG. R_t cannot be solved explicitly but can be estimated as

$$\hat{R}_t = \frac{1}{m}\sum_j^m \left\{\left(\eta_{t,j} - \xi\right)^2\right\} \tag{17}$$

where m is the number of simulated η values from the posterior at a certain threshold.

The selection of a threshold is illustrated next through a simulation study.

Simulate 1000 observations from a GBG with the following parameters: $\mu = 2\,, \sigma = 2, k = 1$ and $\xi = 0.1$. Figure 4 shows the simulated values. Different threshold values are now chosen from the smallest observation to some large observation in small steps. For this simulation the smallest observation is 2.5379 and a large observation of 5 is chosen. The threshold is chosen now as different values from 2.5379 to 5 in steps of 0.01. At each threshold value a vector of η's are simulated from the posterior distribution and the posterior risk function is calculated. Figure 5 shows a plot of the different R_t values at the corresponding threshold values. From Figure 5 it can be seen that the posterior risk is a minimum at a threshold of 3.2679. Therefore we consider 3.2679 to be the best threshold to choose. 3.2679 is the 80.5^{th} percentile. In Figure 6 the chosen threshold is indicated through a solid line in the graph of the simulated data. Assume it is of interest to know $P(X > 4)$. If 3.2679 is an appropriate threshold value, then the mean squared difference between the true tail probability $P(X > 4)$ given in (12) and the approximated tail probability (15) should be very small. The mean squared difference between the tail probabilities at $t = 4$ is calculated as 0.00384, which is small as we expected.

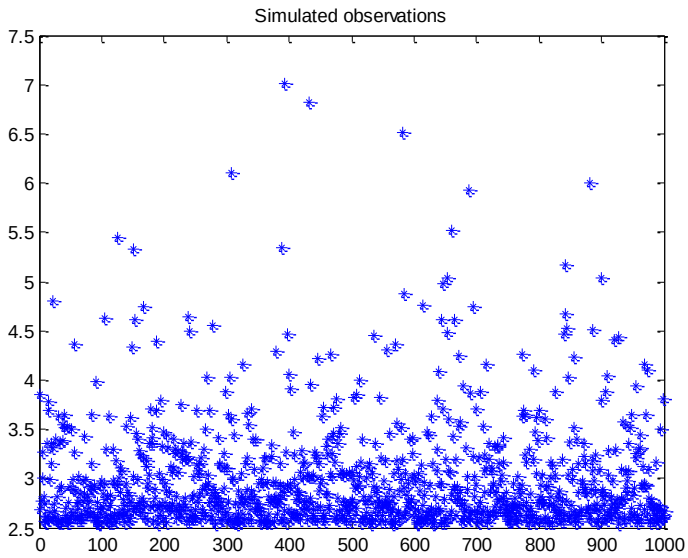

Figure 4 1000 simulated observations from a GBG($\mu = 2\,, \sigma = 2, k = 1, \xi = 0.1$)

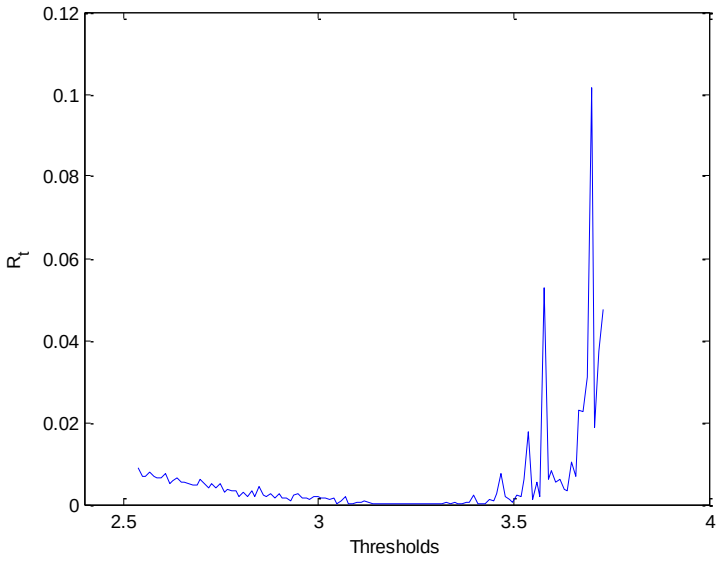

Figure 5 The posterior risk plotted against the thresholds

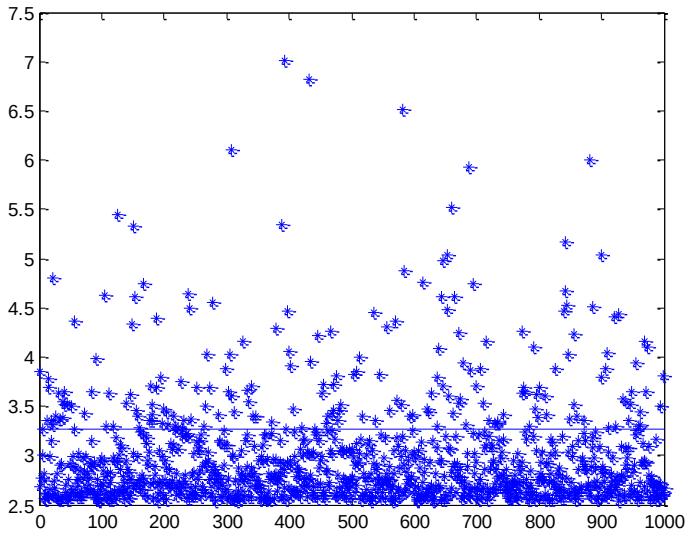

Figure 6 The chosen threshold at t = 3.2679

6 Choosing an optimum threshold in a real data set

The data considered here is the total annual water spillage at the Gariep Dam during 1971 to 2006. The Gariep Dam is the largest reservoir in South Africa and lies in the upper Orange River. At full supply it stores 5943 million cubic meters of water. ESKOM, the main supplier of electricity in South Africa, has a hydro power station at the dam wall consisting of four turbines, each turbine can let through 162 cubic meters per second. If all 4 turbines are operating, the total release of water through the turbines is 648 m^3/s. Spillage over the wall will occur if the dam is 100% full with all 4 turbines running and the inflow into the dam exceeds 648 m^3/s. The total loss observed at Gariep due to spillage during 1971 to 2006 is 1.7693×10^{10} million cubic meters and in terms of South African Rand it was calculated as R76, 950, 708 which is a major loss. It is however important to note that out of the 36 years, 23 appeared without losses. Figure 7 shows the spillage during this period.

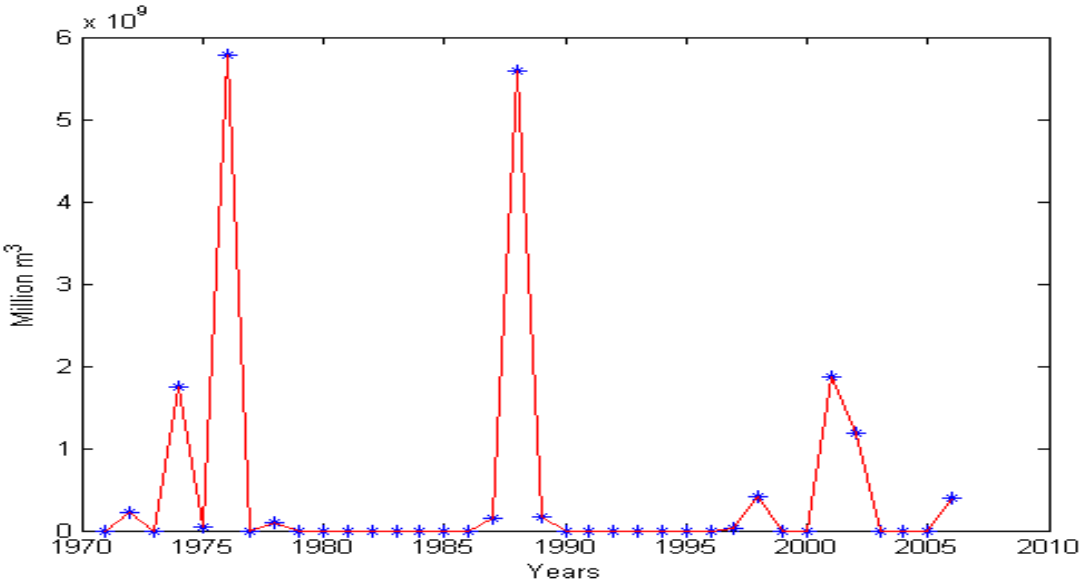

Figure 7 Spillage at Gariep Dam during 1971-2006

It is shown by Verster and De Waal (2009) that the water spillage data set can be considered to be Generalized Burr-Gamma (GBG) distributed with the following set of parameters, $\mu = -19.3732, \sigma = 1.4375, k = 5.2$ and $\xi = 0.1$. Different threshold values are now chosen from the smallest observation to the largest observation. For this simulation the smallest observation is 2.2371×10^7 and the largest observation is 5.7889×10^9, the threshold is taken in steps of 10×10^7. At each threshold value a vector of η's are simulated from the posterior distribution and the posterior risk function is calculated. Figure 8 shows a plot of the different R_t values at the corresponding threshold values. From Figure 8 it can be seen that the posterior risk is a minimum at 0.8524×10^9.

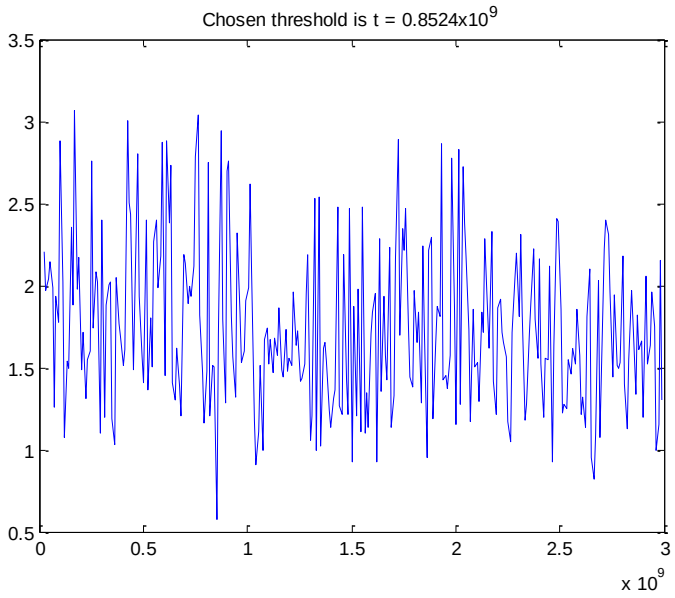

Figure 8 The posterior risk plotted against the thresholds

7 Conclusion

This study shows that the tail of a GBG distribution can be approximated with an Generalized Pareto-type of distribution which is more convenient to work with since it only has one parameter η.

The problem around choosing an optimum threshold is addressed here by considering the posterior risk function at different threshold level. The threshold at which the posterior risk becomes a minimum is then chosen as the best threshold. As shown in this paper, our method for choosing an optimum threshold is simple to work with and effective.

References

AMORE, P (2005). Asymptotic and exact series representations for the incomplete Gamma function. Retrieved August 30, 2007, from www.edpsciences.org/articles/epl/abs/2005/13/epl8802/epl8802.html/

BEIRLANT, J., DE WAAL, D.J., AND TEUGELS, J.L. (1999). 'The generalized Burr-Gamma family of distribution with applications in extreme value analysis', *Limit Theorems in Probability and Statistics*, **1**,113-132.

BEIRLANT, J., GOEDGEBEUR, Y., SEGERS, J. AND TEUGELS, J. (2004). *Statistics of Extremes Theory and Applications*, Weily, England.

Rice, J.A. (1995). *Mathematical Statistics and Data-Analysis Second Edition,* Duxbury Press, California.

SALAS, S.L., HILLE, E., AND ETGEN, G.J. (1999). *Calculus one and several variables Eight Edition*, John Wily & Sons inc., USA.

Peng, L (2009). 'A practical method for analyzing heavy tailed data', *The Canadian Journal of Statistics*, **37**(2), 235-248.

VERSTER, A., AND De Waal, D.J. (2009). ' Modelling Risk on Losses due to Water Spillage for Hydro Power Generation', from http://www.uovs.ac.za/faculties/documents/04/117/TechnicalReports/Teg394.pdf

Appendix A

Three more simulation examples are shown.

Simulation 1

Simulate 1000 observations from a GBG with the following parameters: $\mu = 2\,, \sigma = 1, k = 4$ and $\xi = 0.3$.

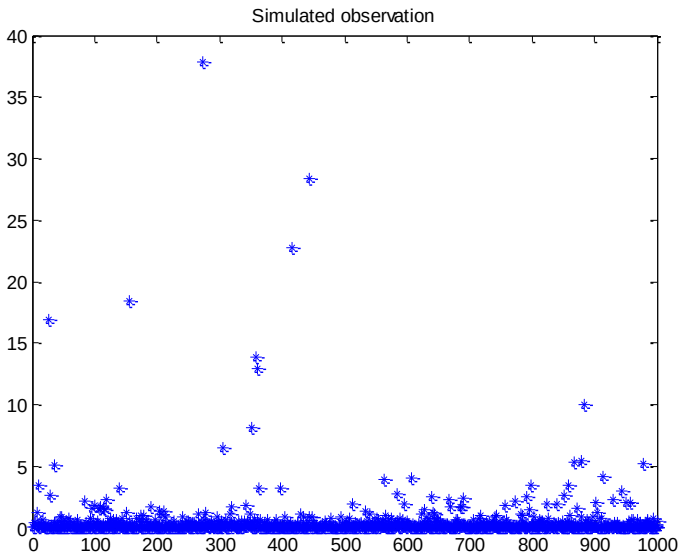

Figure 9 Simulated observations

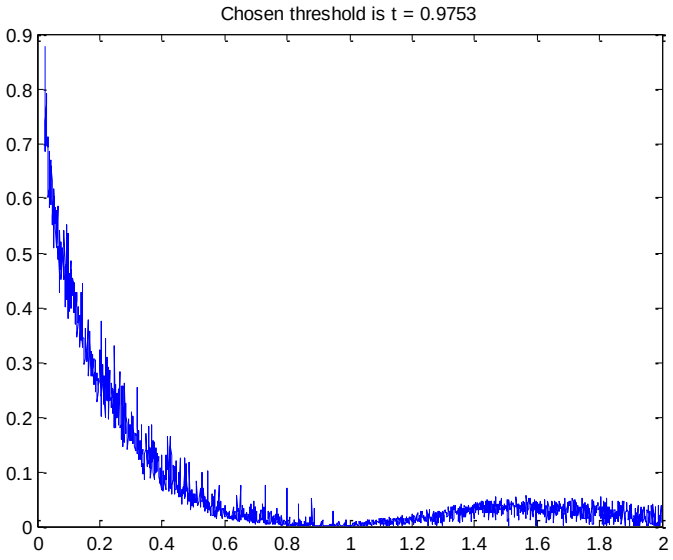

Figure 10 A minimum posterior risk is obtained at t = 0.9753

Simulation 2

Simulate 1000 observations from a GBG with the following parameters: $\mu = 0$, $\sigma = 1$, $k = 2$ and $\xi = 0.5$.

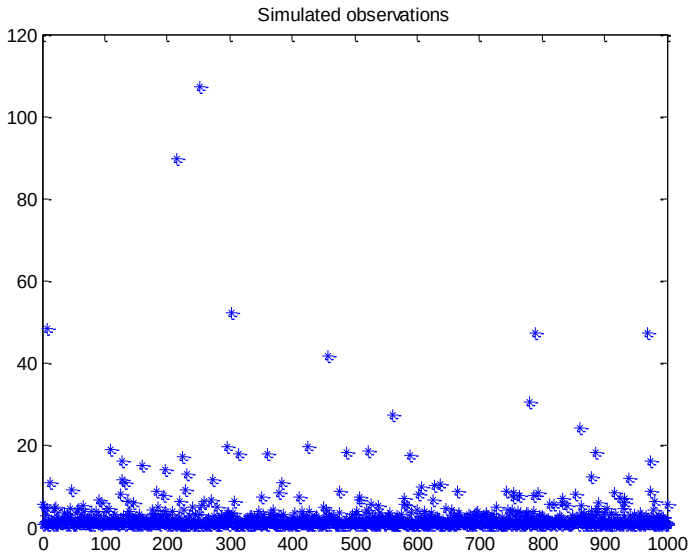

Figure 11 Simulated observations

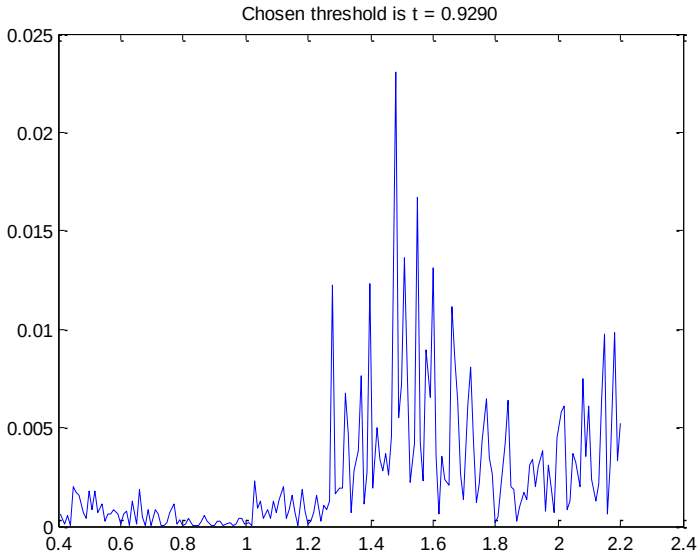

Figure 12 A minimum posterior risk is obtained at $t = 0.9290$

Simulation 3

Simulate 1000 observations from a GBG with the following parameters: $\mu = 2$, $\sigma = 1$, $k = 2$ and $\xi = 0.8$.

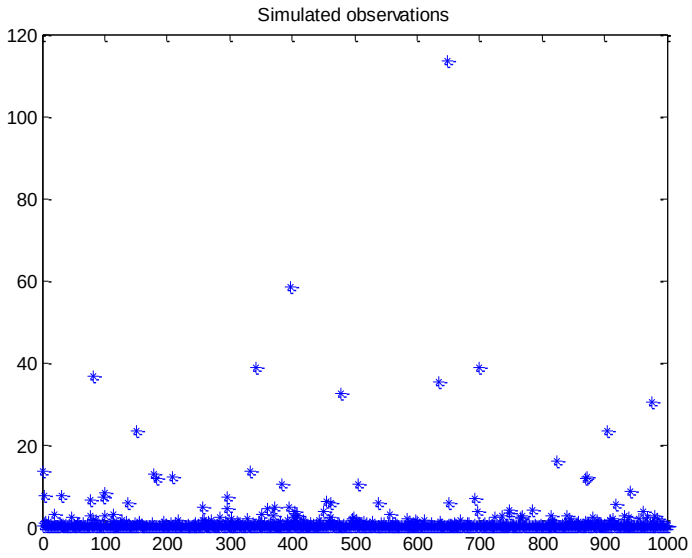

Figure 13 Simulated observations

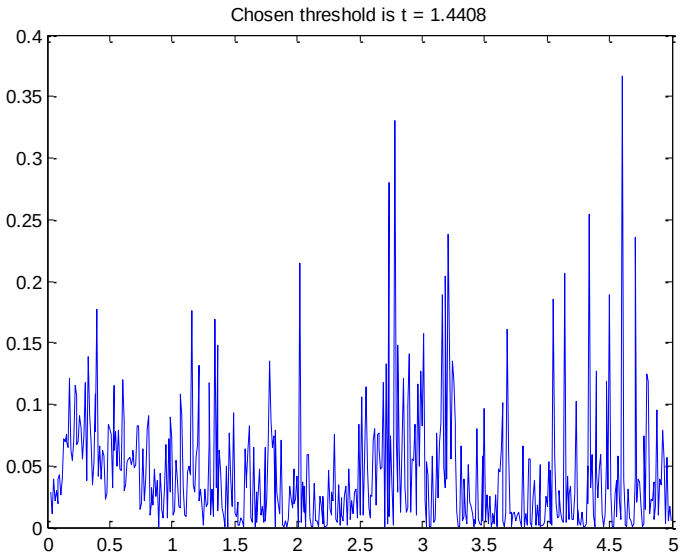

Figure 14 A minimum posterior risk is obtained at t = 1.4408

Appendix B

The MDI prior for η is defined at $\pi(\eta) \propto expE\{\log f(Y|\xi)\}$, Beirlant *et al.* (2004).

$$\overline{F}(v) = \left[1 + \frac{\eta(v-t_v)}{1+\eta t_v}\right]^{\frac{-1}{\eta}}$$

$$f(v) = \frac{1}{1+\eta t_v}\left[1 + \frac{\eta(v-t_v)}{1+\eta t_v}\right]^{-\frac{1}{\eta}-1}$$

Let $\omega = \frac{\eta}{1+\eta t_v}$ then $\eta = \frac{\omega}{1-\omega t_v}$.

$$f(v) = (1-\omega t_v)[1+\omega(v-t_v)]^{-\frac{(1-\omega t_v)}{\omega}-1}$$

$$log(f(v)) = log(1-\omega t_v) - \left(1 + \frac{1-\omega t_v}{\omega}\right) log[1+\omega(v-t_v)]$$

$$E[log(f(v))] = log(1-\omega t_v) - \left(1 + \frac{1-\omega t_v}{\omega}\right) E[log(1+\omega(v-t_v))]$$

$$= log(1-\omega t_v) + \left(1 + \frac{1-\omega t_v}{\omega}\right)\left(\frac{\omega}{1-\omega t_v}\right) E\left[log\left(\overline{F}(v)\right)\right]$$

According to Rice (1995 p. 61) $-log\left[\overline{F}(v)\right] \sim Exp(1)$, and $E\left[log\left(\overline{F}(v)\right)\right] = -1$.

Therefore

$$E[log(f(v))] = log(1-\omega t_v) - \left(1 + \frac{1-\omega t_v}{\omega}\right)\left(\frac{\omega}{1-\omega t_v}\right)$$

$$exp\{E[log(f(v))]\} = (1-\omega t_v)e^{-\left(1+\frac{1-\omega t_v}{\omega}\right)\left(\frac{\omega}{1-\omega t_v}\right)}$$

and

$$\pi(\omega) = (1-\omega t_v)e^{-\left(\frac{\omega}{1-\omega t_v}+1\right)}.$$

Because $\eta = \frac{\omega}{1-\omega t_v}$

$$\pi(\eta) = \frac{e^{-(\eta+1)}}{1+\eta t_v}. \blacksquare$$