Modelling high river flows and Southern Oscillation Index jointly.

D.J. de Waal

A. Verster

University of the Free State, Bloemfontein, South Africa

Abstract: Annual volume of stream flow of the Orange river and October Southern Oscillation Index (SOI) are modeled jointly through the Gumbel copula due to upper tail dependence. The tail dependence coefficient η is estimated from its posterior density under the Pareto type distribution assumed on $T = \min(Z_1,Z_2)$ given η where Z_i , i = 1, 2 are the Fréchet transforms of the observed variables.

1 INTRODUCTION

In the paper de Waal (2009) in honor of the late Jan van Noortwijk, the prediction of the annual volume of water flowing into the Gariep Dam was discussed from a predictive point of view. It has been mentioned that the Southern Oscillation Index (SOI) is identified as a covariate, especially the October SOI and that further investigation on the joint behavior of the inflow (X) and the October SOI (Y) will be done. This paper is a report on that investigation.

The joint distribution of log(X) and Y is assumed to be normal and the fit seems to be acceptable. Proceeding with the bivariate normal as the underlying model, one can derive the joint bivariate t as the predictive density and hence use the conditional posterior predictive density of a future X_0 for predicting a future high inflow given a high October SOI Y_0. These results will be shown in section 2. This approach is however questionable (Beirlant et al, 2004, page 348), since the bivariate normal has the interesting property of lack of tail dependence no matter how large the correlation coefficient ρ<1 (see also Sibuya 1960 and Reiss 1989) .

Tail dependence plays a prominent role in estimating tail probabilities. Ledford and Tawn (1996) introduce the coefficient of tail dependence η≤1 in the tail probability

$$P(Z_1 > z, Z_2 > z) = \ell(z)z^{-1/\eta}, \quad z > 0. \tag{1}$$

Here η is a positive constant and ℓ is a slowly varying function such that $\ell(xz)/\ell(z) \to 1$ as $z \to \infty$ for all $0 < x < \infty$. The rate of decay is primarily controlled by η. Z is considered here as the Fréchet transform -1/log(F) where F denotes the marginal distribution function of either log(X) or Y.

It has been shown that for the bivariate normal η < 1, implies asymptotic tail independence. According to this, a test on η = 1 becomes necessary. Beirlant (2009) addressed this issue and

also discussed the selection of the threshold τ on T = min(Z_1, Z_2). Note that P(Z_1 >z, Z_2 >z) = P(T > z).

In section 3 we will address the issues of estimating η and τ from a Bayesian perspective and in section 4 consider the estimation of a tail probability.

2. The bivariate normal fit

The data consists of the annual volume of water flowing into the Gariep Dam, the largest dam in South Africa in the Orange river, during 1971-2009. It is of interest to ESKOM, the main supplier of electricity, to know the availability of water for hydro power generation at Gariep. It has been found (de Waal, 2009) that the October SOI of the previous year improves the prediction of the inflow of the following year under the assumption E(log(X)) = $\beta_0 + \beta_1 Y$. The October SOI data for the period 1970 – 2008 is available from the internet. We assume that

$$\tilde{X} = \begin{bmatrix} SOI \\ \log\,(Inflow) \end{bmatrix} \sim N(\mu, \Sigma).$$

A scatter plot of the data with bivariate normal contours is shown in figure 1. The estimates of the means and covariance matrix were obtained using maximum likelihood.

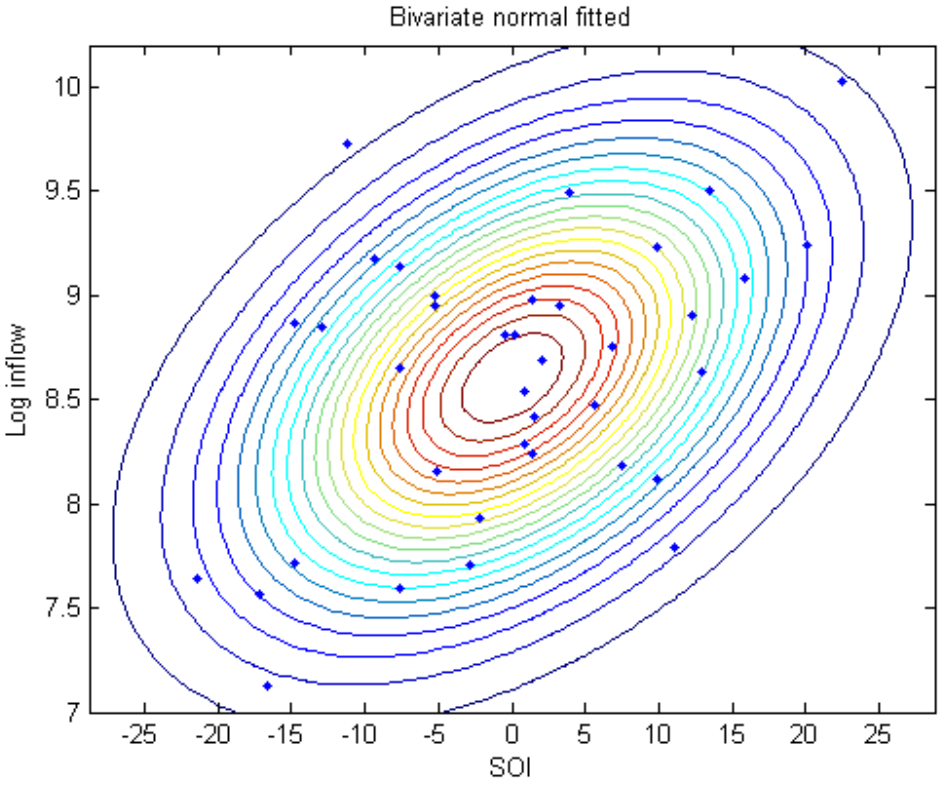

Figure 1: Scatter plot of SOI and log(Inflow) data with normal contours.

From figure 1 it seems that the bivariate normal fit may be acceptable. Proceeding with the well known Bayesian concept for deriving the predictive distribution a future vector $\tilde{X}_0$ given the data on n=39 years (see Zellner, 1971), results in $\bar{x}=\begin{pmatrix}-0.4744\\ 8.6119\end{pmatrix}$, $\hat{\Sigma}=\begin{pmatrix}110.3930 & 3.5908\\ 3.5908 & 0.4282\end{pmatrix}$ and $\tilde{X}_0 \sim T_\nu(\bar{x},S)$. $S=\frac{n+1}{n\nu}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\begin{pmatrix}116.2857 & 3.7824\\ 3.7824 & 0.4510\end{pmatrix}$.

The degrees of freedom ν = n-2 = 37. The estimate $\hat{\mu}$ of μ is given by $\bar{x}$ as the mean of the data $x_1, \ldots, x_n$ on $\tilde{X}$. If the assumption of bivariate normality could be accepted, then the path to predict future annual inflows given a specific October SOI was therefore straight forward. However the following analysis show that this assumption is questionable.

Estimating tail probabilities

(1) Marginal probabilities

Considering tail probabilities such as P(log(I*nflow*)>10) from the marginal N(8.6119, 0.4282) by plugging in the estimates, we get P(log(*Inflow)* > 10) = 0.0169. If we consider however the posterior predictive probability P(log(*Inflow) > 10*|data*),* we get from the marginal t_ν(8.6119, ν(0.4510)) a probability of 0.0229 (larger than from plugging in the estimates) as can be expected due to the heavier tail.

(2) Joint probabilities

From the bivariate normal, through 25000 simulations, we obtain a joint probability P(SOI > 17, log(*Inflow)* > 10) = 0.0052 by plugging in the estimates. From the bivariate posterior predictive T_ν above , we get from 25000 simulations a probability of P(SOI >17, log(*Inflow)* >10|data) = 0.0582 as can be expected with the T being heavier tailed.

(3) Conditional probabilities

Similar tail probabilities can be obtained if we condition on the SOI, but we will not go into this here since the above probabilities already indicate the importance of the predictive approach in under estimating them by just plugging in the estimates of the parameters in the probability equation.

The question arises: How reliable is the joint probability of 0.0582 for the SOI to be larger than 17 and the log(*Inflow)* to be larger than 10? The assumption of normality is critical since we learned that the dependence starts breaking down asymptotically in the tails for the normal. We will consider the dependence structure in the tails for this data in the next section.

3. Tail dependence and selecting the threshold

The estimation of the tail dependence coefficient η defined in (1) will be considered here from a Bayesian point of view . Let T = min(Z_1, Z_2) where Z_1 and Z_2 are Fréchet variables (Beirlant et al, 2004, pages 350-351) and assume the tail distribution a Pareto distribution with only the one parameter η (see example 1). This procedure simplifies the test of hypothesis η = 1 against η <1. We will illustrate through the next two examples.

The bivariate normal

The bivariate normal is a classical case of asymptotically independence even in the case of high correlation ρ < 1. The next example illustrates this from a Bayesian perspective:

Example 1. The posterior distribution of the tail dependence coefficient η described in (1) by assuming a Pareto type (PT) distribution (Verster and de Waal, 2009) to the tail of T from a simulated dataset of n = 500 observations from a standard bivariate normal with means zeros, variances one and covariance ρ will be shown. The observations were transformed to Fréchet values through the transformation z=-1/log(F) with F being the cdf of a standard N(0,1). Beirlant et al (2004), page 351, mentioned that for estimating η, the peaks-over-threshold setting can be applied to T such as fitting a Generalized Pareto (GP) distribution. We will consider a different form of the GP which has only the one parameter η. Verster (2009) showed that the tail of the Generalized Burr Gamma (GBG) can be approximated by a Pareto. The GBG distribution is defined as (Beirlant et al, 2002) as follows: Let T ~ GBG(α,β,κ,ξ), then the distribution function is given by

$$F(t) = \Gamma(\kappa, \log(1+\xi(te^{\alpha})^{1/\beta})/\xi), \quad t > 0 \qquad (2)$$

The parameter κ denotes the scale parameter and ξ the extreme value index. The other two parameters α and β are functions of location, scale and spread of the distribution. α and β appearing in (2) can be written in terms of μ =-E(log(T)|ξ=0) and σ = std(log(T)|ξ=0) as $\beta = \sigma/\sqrt{\psi'(\kappa)}$ and $\alpha = \mu + \beta\psi(\kappa)$ (Beirlant et al, 2002). ψ(κ) denotes the digamma function, ψ'(κ) the trigamma function and Γ(κ, .) denotes the incomplete gamma integral with scale parameter κ and location parameter one. This distribution generalizes the Burr and the Gamma distributions and falls in the Pareto class. Verster (2009) showed that for T exceeding a large threshold τ,

$$S(t|\tau) = P(T > t|T > \tau) \cong \{1 + \frac{\eta}{1+\eta v(\tau)}(v(t) - v(\tau))\}^{-1/\eta}, \; t > \tau \qquad (3)$$

where v(t) = $(te^{\alpha})^{1/\beta}$ and v(τ) = $(\tau e^{\alpha})^{1/\beta}$. (3) falls clearly in the class of the GP with the difference that it has only one parameter. We will refer to (3) as Pareto type (PT). Since v(t) on a log scale is a linear transform of t, we consider in this paper the direct application of PT to the random variable T instead of first fitting the GBG and then apply (3) to the tail of the GBG. We did the GBG fit and then applied (3), but the question arised: Why do all the work to fit the GBG – can we bypass this and fit the PT directly? There was not much difference in fitting the GBG first. We propose

$$S(t|\tau) = P(T > t|T > \tau) \cong \{1 + \frac{\eta}{1+\eta\tau}(t-\tau)\}^{-1/\eta}, \quad t > \tau. \tag{4}$$

As a prior distribution for η, Zellner's (1977) maximal data information (MDI) prior is used namely

$$\pi(\eta) \propto \frac{1}{1+\eta\tau} e^{-\eta} . \tag{5}$$

The proof follows by using the fact that E(log(S(t|τ))) = -1 and that π(η) ∝ exp(E(log(f(t|τ))). f (t|τ) denotes the density function of T.

The posterior of η given the dataset t_1, …, t_k of k exceedances above τ, becomes

$$\pi(\eta|\text{data}) \propto (1+\eta\tau)^{-(k+1)} e^{-\eta}(1+\eta)\prod_{i=1}^{k} S(t_i|\tau)^{(1+\eta)} . \tag{6}$$

We can proceed to calculate the posterior (6) of η. The threshold τ has been selected based on the largest k = 68 observations. This corresponds to exceedances above a threshold τ = 2.94 where the maximum observed value is 105.0149. The mode of (6) has been chosen as the estimate of η and for this data the estimate is $\hat{\eta} = 0.67$. τ has been selected as that value where the mode of (6) is closest to 1. This criterion is chosen because we need to test if η= 1 against η < 1 and we considered the posterior (6) with η closest to one to give it the benefit of being 1. A plot of the posterior is shown in figure 2 and we have to reject the hypothesis that η = 1. This simulation has been repeated a number of times and in almost all cases we accept that η < 1. We confirm the result that the bivariate normal has asymptotical tail independence even for a larger ρ.

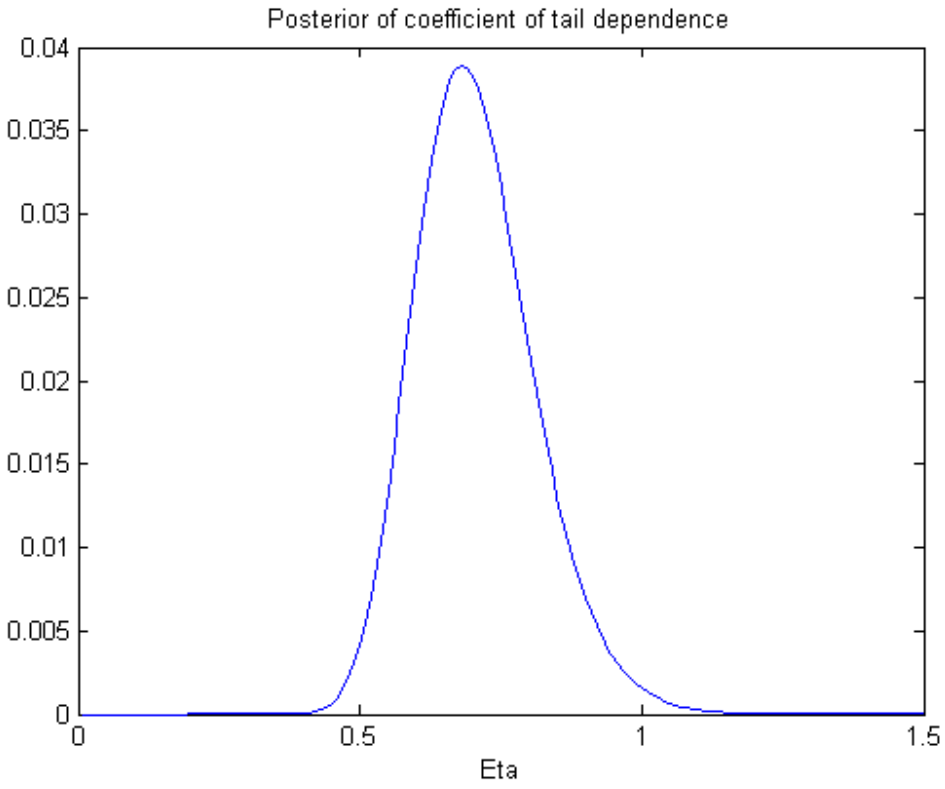

Figure 2: Posterior of η for bivariate normal case

The bivariate t

The bivariate t_ν is considered a case where we have asymptotic tail dependence if ν is small (say 1.5) (see Chen, Wu and Yi, 2009). The following example illustrates this:

Example 2. The same procedure has been followed as in the previous example from a simulated dataset of n = 500 observation from a bivariate t distribution with ν= 1.5 degrees of freedom, zero means and covariance structure the same with ρ = 0.5. The threshold τ = 2.4 was obtained where the posterior (6) had a mode closest to η = 1. The maximum observation was 184.97, clearly more extreme data than in the normal case. Figure 3 shows the posterior of η based on the largest k = 105 observations and again without any doubt it can be assumed it is the case η = 1, namely the tail dependent case.

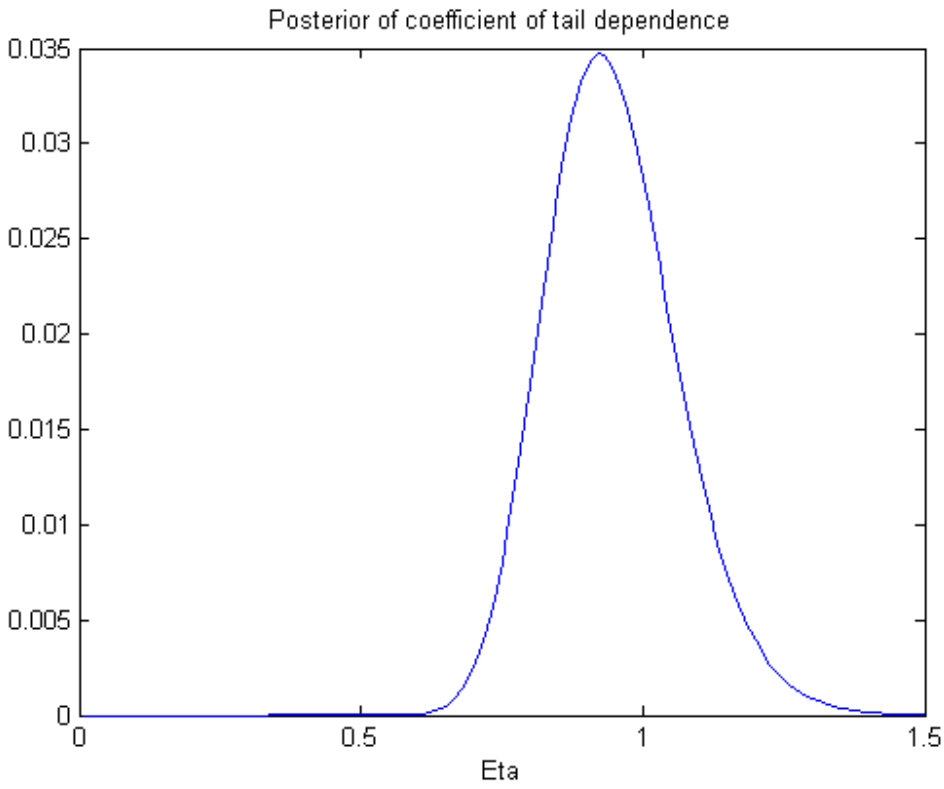

Figure 3: Posterior of η for the bivariate t_v case

4. The posterior of the tail dependence index for the streamflow and SOI data

We will apply the above procedure to estimate η on the October SOI and log(Streamflow) data described in section 2.

The Octber SOI (1970-2008) and logs of annual volume of inflow (1971-2009) data were transformed to Fréchet values using the empirical cdf's, namely $z_{ij} = -1/\log(u_{ij})$, I = 1,2; j = 1,..,n where $u_{ij} = \frac{1}{n+1}\sum_{k=1}^{n} \mathbf{1}(x_{ik} \leq x_{ij})$.

Pareto fit: The threshold was selected as τ = 1.56 (maximum of T is 9.49) and based on the largest k = 7 observations exceeding τ, the posterior of the tail dependence coefficient is shown in figure 4. It is clear that we can accept η = 1 and therefore tail dependence.

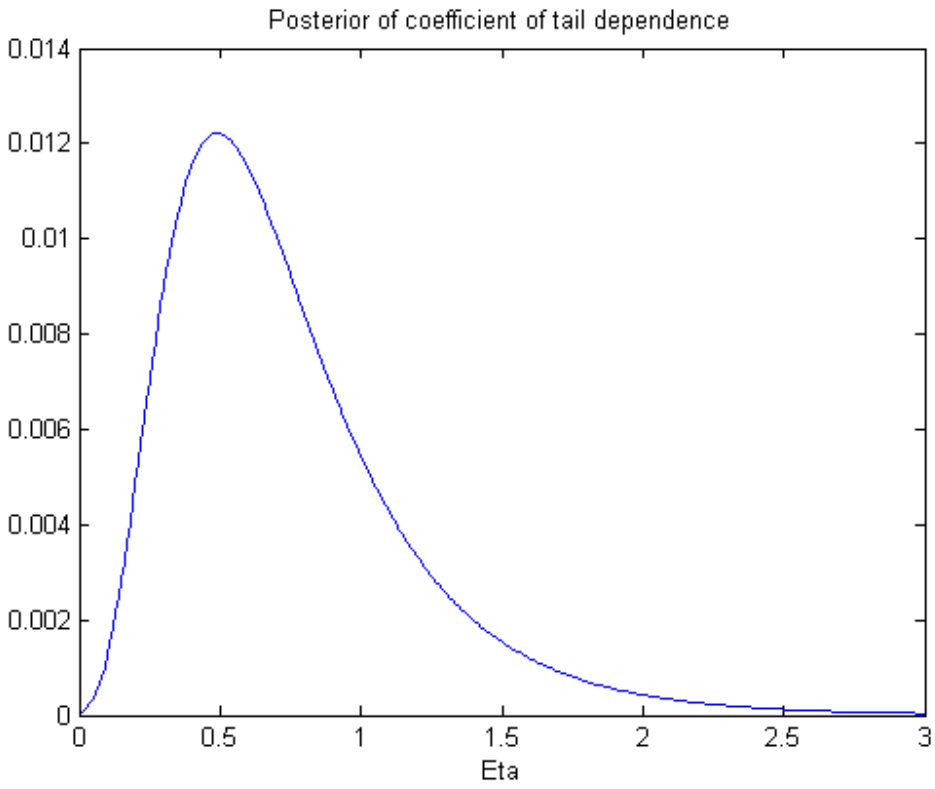

Figure 4: Posterior of η for the October SOI and log(Inflow) data.

5. Bivariate Gumbel copula

From figure 1 we can suspect upper tail dependence. Accepting tail dependence in the upper tail as declared in section 4, we opted for the bivariate Gumbel copula with distribution function (Chen, Wu and Yi, 2009)

$$C(u_1,u_2) = \exp(-[(-\log u_1)^{\alpha} + (-\log u_2)^{\alpha}]^{1/\alpha}, \quad 0 < u_1, u_2 < 1, \; 1 \leq \alpha < \infty. \qquad (6)$$

u_i is taken as the empirical cdf of the i-th variable. The Gumbel copula has Kendall's tau $\kappa_\tau = 1 - 1/\alpha$. An estimate of Kendall's tau is 0.3684 which gives an estimate for α as 1.5833. Tail probabilities can be estimated according to

$$P(Z_1 > z_1, Z_2 > z_2) = \bar{C}(e^{-1/z_1}, e^{-1/z_2}). \qquad (7)$$

Suppose we want to estimate P(SOI > 17, log(Inflow) > 10), then $z_1 = 2.5443$ and $z_2 = 1.1687$ or equivalently $u_1 = 0.6750$ and $u_2 = 0.4250$. Since $\bar{C}(u_1,u_2) = 1 - u_1 - u_2 + C(u_1,u_2)$, we get an estimate of this probability after substituting the relevant values in (6) and (7) of 0.2657. This turns out to be a much larger as that obtained in section 2.2 under normal and t assumptions.

5. Conclusion

In this paper we presented a method to test for tail dependency in bivariate data containing possible extremes. The method propose the fitting of the Pareto type distribution to the minimum of the Fréchet transforms and then select the threshold by considering the posterior distribution of the tail dependence coefficient. The parameter of the Pareto distribution reflects the tail dependence coefficient η and its posterior distribution is fairly easy to derive. From the posterior distribution of η one can judge if $\eta = 1$ or if $\eta < 1$. This knowledge is necessary to proceed to estimate tail probabilities under the appropriate copula.

References

Chen X., Wu W.B. and Yi Y. (2009): Efficient estimation of Markov models. *The Annals of Statistics, Vol 37, 6B, pp 4214-4253*.

Beirlant J.; Goegebeur Y.; Segers J. and Teugels J.L. (2004): *Statistics of Extremes. Theory and Applications.* Wiley.

Beirlant J., de Waal D.J. and Teugels J.L. (2002): The Generalised Burr-Gamma family of distributionswith applications in Extreme value Analysis. *Limit theorems in Probability and Statsitics I (I. Berkes, E. Csaki, M. Csorgo, eds.). Budapest, pp 113-132.*

Beirlant J.; Dierckx G. and Guillou A. (2009): Biased reduced estimators in joint tail modeling. *Collection of papers in commemoration of D de Waal's 34 years of service as head of department of Mathematical Statistics. (M. Finkelstein eds). Dept Math Stats, Univ Free State, Bloemfontein. S.A.*

De Waal D.J (2009): Posterior predictions on river discharges. (2009): *Risk and Decision Analysis in Maintenance Optimization and Flood Management. (M.J. Kallen & S.P. Kuniewski eds). IOS Press*

Ledford A.W. and Tawn J.A. (1996): Statistics for near independence in multivariate extreme values. *Biometrika 83, pp 169-187.*

Reiss R.D. (1989): *Approximate Distributions of Order Statistics.* Springer-Verlag.

Sibuya M. (1960): Bivariate extreme statistics. *Annals of the Institute of Statistical Mathematics, 11,pp 195-210.*

Verster A. (2009): *Modelling of Multivariate Extreme data.* Unpublished PhD thesis. Dept Math Statistics, University of the Free Sate, Bloemfontein, S.A.

Verster A. and de Waal D.J. (2009): Modelling Risk on Losses due to spillage for hydro Power Generation. *Tech. report nr 394, Dept Math Statistics. University Free Sate.*

Zellner A. (1971): *An Introduction to Bayesian Inference in Econometrics.* Wiley

Zellner A. (1977): *Bayesian Analysis in Econometrics.* Edward Elgar.