A MULTIVARIATE GAMMA DISTRIBUTION APPLIED TO COMPOSITIONAL DATA ANALYSIS

Daan de Waal
University of the Free State
e-mail: *dewaaldj@ufs.ac.za*

Roelof Coetzer
Sasol Group Technology and University of the Free State
e-mail: *roelof.coetzer@sasol.com*

Sean van der Merwe [1]
University of the Free State
e-mail: *vandermerwes@ufs.ac.za*

Key words: Classification, Compositional data, Multivariate extreme value distribution, Multivariate gamma distribution.

Abstract: Parametric compositional data analysis in a high dimensional simplex can be performed by employing the Dirichlet distribution, or alternatively, through the logistic normal distribution if the Dirichlet is not appropriate. In this paper, a multivariate gamma (MGAM) distribution is proposed as an alternative distribution for compositional data. In addition, the MGAM distribution is extended to a multivariate extreme value (MEV) distribution and goodness of fit statistics are calculated for comparison against the logistic normal distribution. An application is considered where the amount of gas produced from a coal gasication facility depends crucially on the size distribution of the coal, which is measured as compositional data and characterised by six variables. The observed sample space is divided into three regions of high (H), standard (S) and low (L) gas production by choosing appropriate thresholds, and new observations are classified among the regions.

1. Introduction

The amount of gas produced from a coal gasication facility depends crucially on the properties and the size distribution of the coal being used in the process. Therefore, in order to optimize gas production, the relationship between the type of coal and the gas produced must be understood and quantified. In the current application, data on six coal sizes were observed and it is expressed as compositional in a five dimensional simplex. The data are not Dirichlet since positive correlations have been found amongst the variables. Therefore, the logistic normal (LTN) distribution is an alternative parametric model to apply (Aitchison, 1986). In this paper, we propose a multivariate gamma

[1] Corresponding author.

AMS: 62H30

distribution (MGAM) as an alternative to the LTN. There exists a number of multivariate gamma distributions in the literature (see for example Prékopa and Szántai, 1978; Patil, Boswell, Ratnaparkhi and Roux, 1984). However, this distribution presented here is new, and has the advantage that it is fairly simple to fit to data. It also has the advantage that if the data is heavy tailed, the gamma is a more appropriate distribution than the logistic normal. It has fewer number of parameters than the LTN and estimates can easily be obtained. A goodness of fit test is considered using the largest trace of a difference square matrix between the observed matrix and the simulated data from the specified distribution. In the case of the LTN, we used the posterior predictive distribution (Geisser, 1983) instead of the plug-in estimates as in the case of the MGAM. The posterior predictive distribution of the MGAM is not considered since the distribution is not available explicitly.

The sample of size $n = 125$ is divided into three groups, namely $n_H = 23$ yielding high (H) gas production, $n_S = 84$ yielding standard (S) production and $n_L = 18$ yielding low (L) production. The objective is to classify a new observation as H, S or L. The sample space of the MGAM is partitioned into the three regions by partitioning the sample space of the trace of the distance measure between an observation and its simulated equivalent from the model with thresholds derived from the observed outcomes such that 18% of the observations belong to H, 68% to S and 14% to L.

Zeros do occur, but were considered to be outliers. Therefore, observations containing zeros were deleted.

2. Multivariate gamma distribution

There exist many definitions of multivariate gamma distributions. The version proposed here is a generalization of the one defined by de Waal, van Gelder and Beirlant (2004). The definition is as follows:

Definition: The random variable $V(p \times 1)$ is distributed multivariate gamma (MGAM) with shape parameter $\alpha(p \times 1)$ and correlation structure $(\rho_{ij}), i, j = 1, \ldots, p$ if the elements of $Z = \exp(W)$ are distributed independently $\text{Gamma}(\alpha_i, 1), i = 1, \ldots, p$ where $W = H^{1/2} D^{-1}\{\log V - \psi(\alpha)\} + \psi(\alpha)$. D is defined as a matrix square root of $\Lambda = \left(\rho_{ij}\sqrt{\psi'(\alpha_i)\psi'(\alpha_j)}\right), i, j = 1, \ldots, p$. Explicitly, $\Lambda = D'D$ and $H = diag(\psi'(\alpha))$. $\psi(.)$ refers to the standard digamma function applied independently to each element of its argument. $\psi'(.)$ refers to its first derivative, the trigamma function, and $diag(.)$ refers to the diagonal matrix.

Remarks:

1. $E(W) = \psi(\alpha)$ and $\text{cov}(W, W') = H$.

2. We refer to the distribution of $\log V$ as a multivariate extreme value (MEV) distribution with $E(\log V) = \psi(\alpha)$ and $\text{Cov}(\log V) = \Lambda$. Cov(.) refers to the covariance matrix. The special case $p = 1$ and $\alpha = 1$ reduces the MGAM of V to the standard exponential distribution with the distribution of $\log V$ known as the extreme value distribution (Kotz and Nadarajah, 2000).

3. It is known that the $\log \Gamma(\alpha, \beta)$ distribution belongs to the Pareto class (Beirlant, Goegebeur, Segers, Teugels, de Waal and Ferro, 2004) with extreme value index $1/\alpha$. Therefore, it may be specified that $W = \exp(V)$ belongs to a class of multivariate Pareto distributions (the Fréchet domain).

4. Let us consider the marginal distributions of $V_i,\ i = 1,\dots,p$:

   a. If $p = 1$, then V_1 $\sim$Gamma$(\alpha_1, 1)$ as the matrix inverse becomes a simple reciprocal and $\rho_{11} = 1$.

   b. In the case where $\rho_{ij} = 0, i \neq j$, the V_i are distributed independent Gamma$(\alpha_i, 1)$, as all matrices become diagonal and the transformation then cancels out trivially.

   c. In the general case where $\rho_{ij} \neq 0$ for some $i \neq j$, the marginal distributions of V_i and V_j do not appear to follow any standard distribution. See Appendix A for an explanation.

5. Let $\log V = \Sigma^{1/2}Y + \mu$, then Y is distributed as a generalised multivariate gamma distribution with parameters $\mu, \Sigma, \alpha, \Lambda$ denoted by GMGAM$(\mu, \Sigma, \alpha, \Lambda)$, with $E(Y) = \mu$, Cov$(Y, Y') = \Sigma$.

6. If $\mu = 0$ and $\Sigma = I_p$, we refer to $Y \sim$ MGAM(α, ρ), the multivariate gamma distribution.

7. If $\rho_{ij} = 0, i \neq j$, then $Y \sim$ GMGAM(α, μ, Σ) defined in de Waal et al. (2004).

8. To fit a MGAM distribution to compositional data on X in the p dimensional simplex, let $Y = -\log X$ and assume $Y \sim$ MGAM(α, ρ).

9. We refer to the distribution of $\log(Y)$ as a multivariate extreme value distribution (MEV).

10. To simulate an x value on X, we need to go backwards in the definition above. The steps are:

    a. Generate $w_i \in$ Gamma$(\alpha_i, 1)$ independently for each $i = 1,\dots,p$.

    b. Let $\mu_i = \log(w_i)$ and standardise by subtracting the mean $\psi(\alpha_i)$ and dividing by $\sqrt{\psi'(\alpha_i)}$, namely $\mu_s = diag^{-1}(\sqrt{\psi'(\alpha)})\{\mu - \psi(\alpha)\}$.

    c. Transform μ_s to have covariance Λ and mean $\psi(\alpha)$, namely $v = \Lambda^{-1/2}\mu_s + \psi(\alpha)$.

    d. Let $x = \exp(-v)$.

    e. Rescale x such that $\sum_{i=1}^{p} x_i < 1$. One way to do this is to divide each element of x by $p - \sum_{i=1}^{p} x_i$. $x_1,\dots,x_p$ becomes the required generated compositional observation in the p dimensional simplex.

3. Two dimensional MGAM

As an illustration we simulated data from a MGAM distribution in two dimensions. Figure 1 shows scatter plots of two sets of data simulated from a MEV distribution with different α's and ρ's. The data is clearly heavy tailed and the marginals are skew distributed. The marginals are of course gamma distributed. The figures highlight the ability of this distribution to capture some non-traditional relationships.

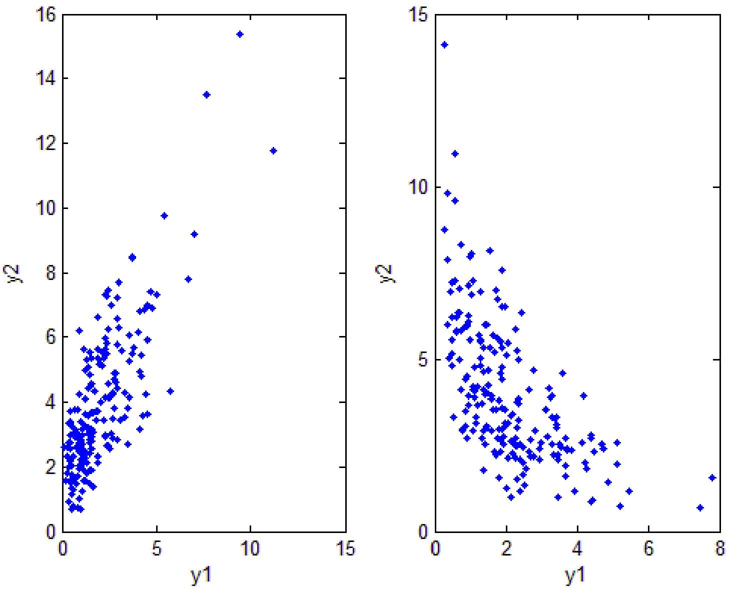

Figure 1: Two sets of size $n = 200$ simulated from two dimensional MGAM distributions with α = [2 4], and $\rho_{12} = 0.7$ and -0.7, respectively.

4. Application of the MGAM

Consider a MGAM(α,ρ) fit to the compositional data on sizes of coal in a five dimensional simplex. Let $y = -\log(x)$ where x is the $(n = 130, p = 5)$ data matrix and we want to fit a MGAM(α,ρ) to y. It is however advisable to rather fit a MEV(α,Λ) distribution to $z = \log(y)$. The moment estimates are simple to obtain. Since $E\log(Y) = \psi(\alpha)$, it can be estimated by the mean of $z = \log(y)$, say $\tilde{z}$. From the inverse function $\hat{\alpha} = \psi^{-1}(\tilde{y})$, estimates of the shape parameters can be obtained by numerically calculating the inverse values. Also, since the covariance of $Z = \log(Y)$ is Λ, it is easy to estimate through standard algorithms. An estimate of ρ can be obtained instead. The estimates are:

$$\hat{\alpha} = \begin{bmatrix}1.4700 & 2.0700 & 1.8800 & 4.2100 & 2.7700\end{bmatrix},$$

$$\hat{\rho} = \begin{bmatrix} 1.0000 & -0.7716 & -0.4760 & -0.0475 & -0.3735 \\ -0.7716 & 1.0000 & -0.0954 & 0.0808 & 0.2949 \\ -0.4760 & -0.0954 & 1.0000 & -0.1252 & -0.1743 \\ -0.0475 & 0.0808 & -0.1252 & 1.0000 & -0.1160 \\ -0.3735 & 0.2949 & -0.1743 & -0.1160 & 1.0000 \end{bmatrix}.$$

The estimate of Λ becomes $\hat{\Lambda} = \left(\hat{\rho}_{ij}\sqrt{\tilde{z}_i\tilde{z}_j}\right), i, j = 1,\ldots,p$.

According to the definition of the MGAM, we need to transform the data y to have marginals Gamma$(\alpha,1)$. We therefore transform $Y_i \sim$ Gamma(a,b) to $Y_i \sim$ Gamma$(\alpha_i,1)$ independently for all $i = 1,\ldots,p$. We then have that $Z = \log Y \sim \text{MEV}(\alpha,\rho)$. Thereafter, we can proceed with the analysis on z.

We are now able to generate y from $\text{MGAM}(\alpha,\rho)$ or z from $\text{MEV}(\alpha,\rho)$. The Mahalanobis distance measure

$$D(z) = \frac{1}{n} trace[(z_0 - z)'(z_0 - z)].$$

between an observed matrix z and a generated z_0 from a given model can be used to select $\hat{\alpha}$. The closer z_0 gets to z, that is, the smaller $D(z)$, the better the $\hat{\alpha}$.

Since $D(z)$ varies from simulation to simulation, 500 z_0 values were generated and the distribution of $D(z)$ considered. Figure 2 shows the empirical cdf of such $D(z)$ values for the above selected $\hat{\alpha}$. It also serves as a goodness of fit check in the sense that it provides the $\hat{\alpha}$ where the generated z_0 is close to the observed z.

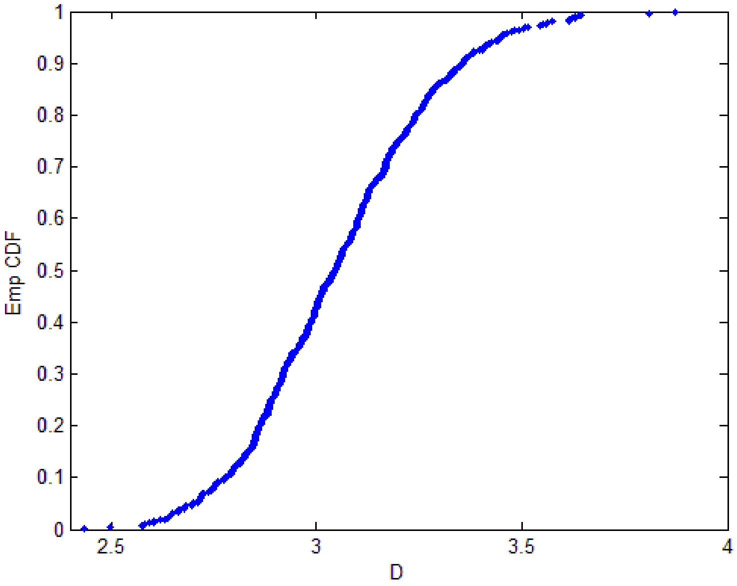

Figure 2: Empirical cdf's of 30 repetitions on distances D between z and a simulated dataset z_0 with 500 simulations from $\text{MEV}(\alpha,\rho)$. Dotted cdf on D is obtained for the observed z.

4.1. Classification of new observations under the MGAM model

We consider now the classification of a new observation z under the MEV model between three groups H, S and L on coal sizes. We need a measure to classify a new observation as H, S or L. As a measure we consider the linear function $w = \beta_0 + \beta_1 z_1 + \cdots + \beta_p z_p$ with z_i the MEV variables and w the percentiles of the cdf of the observed gas production. The sample space of w will then be divided into three regions representing H, S and L.

Consider the classification of a new observation z on coal sizes obtained through the transformations discussed in section 2 under the assumption that $Z \sim \mathrm{MEV}(\alpha, \rho)$. We want to classify the observation as H, S or L.

Estimates of β's defined in the linear function $w = \beta[1\, z_1 \ldots z_p]^T$ are obtained through least squares minimization. From the dataset of $n = 112$ observations, we get

$$\hat{\beta} = \begin{bmatrix} 1.5467 & 0.2394 & -0.2877 & 0.4000 & 0.1341 & 0.2767 \end{bmatrix}.$$

The R^2 is equal to 0.48 for the linear model. This is increased to 0.56 if interaction terms are included in the linear function. The specific value of R^2 for this example is not meaningful on its own, we use it merely as a starting point for model comparison. Fitting more complex models in an attempt to raise this value is possible, but outside the scope of this article.

The empirical cdf of the calculated w's is shown in Figure 3. Selecting thresholds on the sample space of w as best guesses, we can specify classification rules: say if $w > 2.3$ classify the observation as L, $w < 1.6$ classify the observation as H and otherwise classify it as S. The thresholds can be chosen optimally by striving for the outcome on the classification accuracy to be as close as possible to the real outcomes H = 20, S = 76 and L = 16. The classification results are H = 24, S = 67 and L = 21 of which 13 H were classified correctly, 55 S were correct and 11 L observations classified correctly. Adjusting the thresholds, we may get improved classifications. The success rate is 71% . Although the multiple linear regression model does not fit that well on the MEV variables, it does contain information on the relationship between the coal sizes and the gas production.

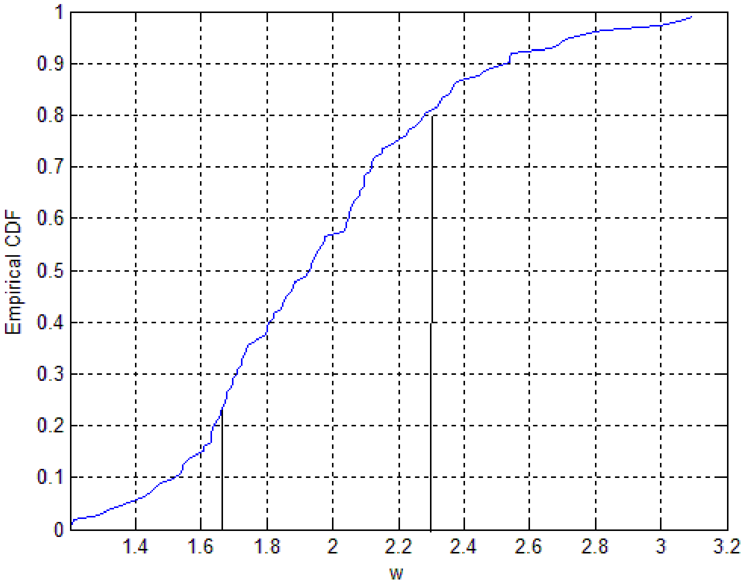

Figure 3: Empirical cdf of W with thresholds dividing the sample space.

5. Logistic normal

Under the logistic normal model, we assume that $X \sim$ logistic normal(μ,Σ) or $Y \sim N(\mu,\Sigma)$ where $Y_i = \log\left(\frac{X_i}{1-\sum_{i=1}^{p} X_i}\right)$. It is known (Geisser, 1983) that the predictive distribution of a future Y_0 is multivariate t with $n-p$ degrees of freedom, mean $\bar{y}$ and precision matrix $T = \frac{n(n-p)}{n^2-1}S^{-1}$. $S = \frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})'(y_j-\bar{y})$ is the estimated covariance matrix of Y.

Considering the dataset y (124 $\times$ 5) on coal sizes, the estimates of μ and Σ are

$$\bar{y} = \begin{bmatrix} 2.7617 & 2.1730 & 2.3919 & 0.0584 & 1.5043 \end{bmatrix}$$

and

$$S = \begin{bmatrix} 0.6416 & 0.1496 & 0.3571 & 0.3134 & 0.4040 \\ 0.1496 & 0.6737 & 0.3751 & 0.2859 & 0.5625 \\ 0.3571 & 0.3751 & 0.7714 & 0.3239 & 0.4942 \\ 0.3134 & 0.2859 & 0.3239 & 0.8963 & 0.3551 \\ 0.4040 & 0.5625 & 0.4942 & 0.3551 & 1.1189 \end{bmatrix}.$$

A goodness of fit test is employed with measure $D(y) = \frac{1}{n}trace[(y_0-y)'(y_0-y)]$.

If y denotes the data and y_0 simulated from a multivariate normal with parameters above and repeated 500 times, the cdf of $D(y)$ is shown in Figure 4.

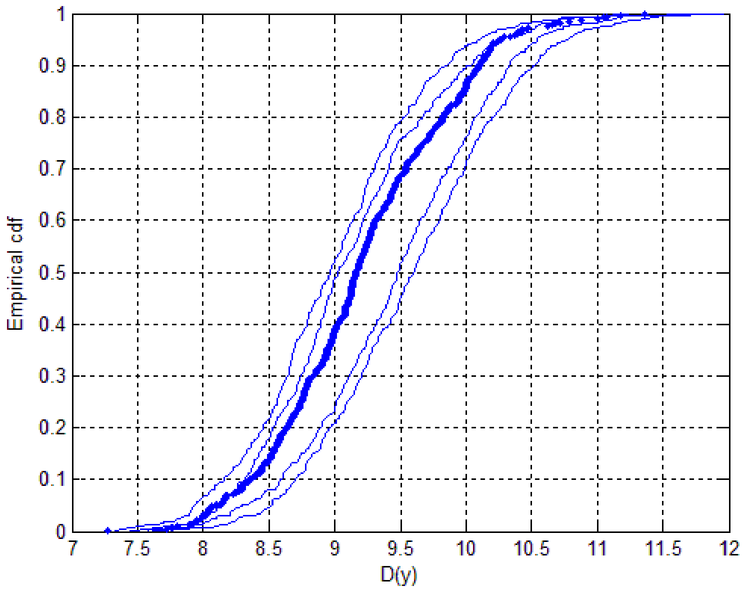

Figure 4: Empirical cdf of $D(y)$ shown by dots based on 500 simulations. Lines show a few repetitions if normal data is used instead of real data.

The logistic normal fit is also acceptable, as shown in Figure 4 from a few repetitions of simulated data instead of the data.

Alternatively, similar to the distance measure above, we can specify

$$U(Y_0) = \frac{1}{p}(Y_0 - \bar{y})'T(Y_0 - \bar{y}) \sim F(p, n-p).$$

The statistic U can also be considered as a distance measure between Y_0 and the mean $\bar{y}$. From the distribution of U, 95% critical values can be obtained to be able to make a decision on judging the validity of the model.

We cannot conclude whether the MGAM fits better than the LTN. The objective of the paper is to show that the MGAM is an alternative distribution to consider and especially for heavier tailed data it may be more appropriate. For the classification outcomes, the logistic normal model also gives good results in the following application.

5.1. Classifying a new observation under the LTN model

We will consider the classification of a new observations under the logistic normal model. We use the linear function $w = \beta_0 + \beta_1 y_1 + \cdots + \beta_p y_p$ for classification. Figure 5 shows the empirical cdf of w for the observed data with thresholds chosen to classify an observation between H, S and L. As a classification rule, if $w < 1.7$, the observation is classified as H, if $w > 2.26$ as L and as S otherwise. These thresholds can be optimised by minimising the difference between the outcomes of the classification on all the data and the observed, namely 20 H, 76 S and 16 L observations.

The two cdf's of w, the one for the MGAM (Figure 3) and the other for the logistic normal (Figure 5), are very similar except that the cdf corresponding to the MGAM model has a heavier tail for H observations.

The classification matrix in Table 1 shows the outcomes.

Table 1: Classification outcome matrix A.

Observed \ Assigned	H	S	L
H	13	7	0
S	13	49	4
L	0	4	11

The percentage correct classification is 65%. This is lower than the outcome under the MGAM model, but by changing the thresholds, this can be higher.

6. Conclusion

In conclusion we can state that MGAM and logistic normal give very similar classification results on the dataset as a whole. Both models seem to fit well. The largest eigenvalues on the square distances between observed and simulated values as a classification measure is giving satisfactorily results, but other measures can be used. The choice of the thresholds is also debatable and other choices than squared determinant differences may be appropriate. The assignment of new observations to H, S or L can with at least 55% confidence on properties and sizes be done under both models.

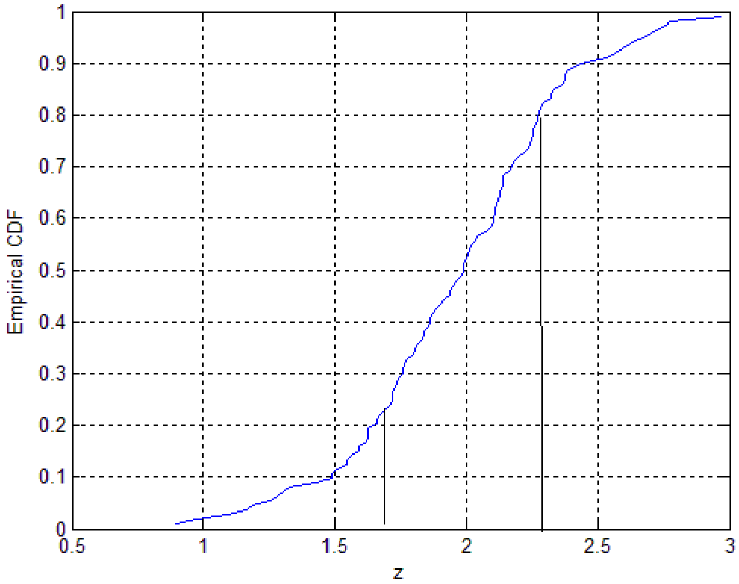

Figure 5: Empirical cdf of w with lines showing the thresholds.

Notes

All analyses were programmed in MATLAB (The Mathworks Inc., 2013).

References

AITCHISON, J. (1986). *The Statistical Analysis of Compositional Data*. Springer: Netherlands. doi:10.1007/978-94-009-4109-0.

BEIRLANT, J., GOEGEBEUR, Y., SEGERS, J., TEUGELS, J., DE WAAL, D., AND FERRO, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons: Chichester.
URL: `https://books.google.co.za/books?id=GtIYLAlTcKEC`

DE WAAL, D. J., VAN GELDER, P. H. A. J. M., AND BEIRLANT, J. (2004). Joint modelling of daily maximum wind strengths through the multivariate Burr-Gamma distribution. *Journal of Wind Engineering and Industrial Aerodynamics*, **92** (12), 1025–1037. doi:http://dx.doi.org/10.1016/j.jweia.2004.06.001.
URL: `http://www.sciencedirect.com/science/article/pii/S0167610504000832`

GEISSER, S. (1983). *On The Prediction Of Observables: A Selective Update*. University of Minnesota, School of Statistics.

KOTZ, S. AND NADARAJAH, S. (2000). *Extreme Value Distributions: Theory and Applications*. Imperial College Press: London.
URL: `http://books.google.co.za/books?id=b40P_o3yXuUC`

PATIL, G. P., BOSWELL, M. T., RATNAPARKHI, M. V., AND ROUX, J. J. J. (1984). *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work*, volume 3: Multivariate Models. International Co-operative Publishing House: Burtonsville, MD.

PRÉKOPA, A. AND SZÁNTAI, T. (1978). A new multivariate gamma distribution and its fitting to empirical streamflow data. *Water Resources Research*, **14** (1), 19–24. doi:10.1029/WR014i001p00019.
URL: `http://dx.doi.org/10.1029/WR014i001p00019`

THE MATHWORKS INC. (2013). Matlab.
URL: `http://www.mathworks.com/`

Appendix A: Marginal and conditional distributions

Marginal distributions in the general case

To understand the case where there is non-zero correlation, we consider $p = 2$ and provide the following derivation:

First, let R be the matrix of correlation parameters $[\rho_{ij}] = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, then $\Lambda = H^{1/2}RH^{1/2}$.

$$\therefore D = R^{1/2}H^{1/2}, \text{ where } R^{1/2} = \frac{1}{\sqrt{2+2\sqrt{1-\rho^2}}}\begin{bmatrix} 1+\sqrt{1-\rho^2} & \rho \\ \rho & 1+\sqrt{1-\rho^2} \end{bmatrix}.$$

And so, defining $r = 1+\sqrt{1-\rho^2}$,

$$H^{1/2}D^{-1} = (2r)^{-0.5}\begin{bmatrix} r & -\rho \\ -\rho & r \end{bmatrix}.$$

$$\therefore\ W = (2r)^{-0.5}\begin{bmatrix} r(\log V_1 - \psi(\alpha_1)) - \rho(\log V_2 - \psi(\alpha_2)) \\ -\rho(\log V_1 - \psi(\alpha_1)) + r(\log V_2 - \psi(\alpha_2)) \end{bmatrix} + \begin{bmatrix} \psi(\alpha_1) \\ \psi(\alpha_2) \end{bmatrix}$$

$$\begin{aligned} \text{and } Z &= \begin{bmatrix} V_1^{(2r)^{-0.5}r} V_2^{-(2r)^{-0.5}\rho} e^{(2r)^{-0.5}[-r\psi(\alpha_1)+\psi(\alpha_2)\rho]+\psi(\alpha_1)} \\ V_1^{-(2r)^{-0.5}\rho} V_2^{(2r)^{-0.5}r} e^{(2r)^{-0.5}[-r\psi(\alpha_2)+\psi(\alpha_1)\rho]+\psi(\alpha_2)} \end{bmatrix} \\ &= \begin{bmatrix} V_1^{(2r)^{-0.5}r} V_2^{-(2r)^{-0.5}\rho} c_1 \\ V_1^{-(2r)^{-0.5}\rho} V_2^{(2r)^{-0.5}r} c_2 \end{bmatrix}. \end{aligned}$$

The joint density of Z is $f_Z(\mathbf{z}) = [\Gamma(\alpha_1)\Gamma(\alpha_2)]^{-1} z_1^{\alpha_1-1} z_2^{\alpha_2-1} \exp\{-z_1 - z_2\}$, so

$$\begin{aligned} f_V(\mathbf{v}) &= c_3 v_1^{(2r)^{-0.5}[r(\alpha_1-1)-(\alpha_2-1)\rho]} v_2^{(2r)^{-0.5}[r(\alpha_2-1)-(\alpha_1-1)\rho]} \\ &\quad \times \exp\{-v_1^{(2r)^{-0.5}r} v_2^{-(2r)^{-0.5}\rho} c_1 - v_1^{-(2r)^{-0.5}\rho} v_2^{(2r)^{-0.5}r} c_2\} \\ &\quad \times c_1 c_2 v_1^{(2r)^{-0.5}r-2-(2r)^{-0.5}\rho} v_2^{(2r)^{-0.5}r-2-(2r)^{-0.5}\rho} \times |c_4| \\ &= c_5 v_1^{(2r)^{-0.5}[r\alpha_1-\alpha_2\rho]-1} v_2^{(2r)^{-0.5}[r\alpha_2-\alpha_1\rho]-1} \\ &\quad \times \exp\{-c_1 v_1^{(2r)^{-0.5}r} v_2^{-(2r)^{-0.5}\rho} - c_2 v_1^{-(2r)^{-0.5}\rho} v_2^{(2r)^{-0.5}r}\}. \end{aligned}$$

$\therefore\ f_{v_1}(v_1) = \int_0^\infty a v_2^b e^{-c v_2^d - g v_2^h}\, dv_2$ where a, b, c, g, h are functions of v_1 and the parameters; while d, h are just functions of ρ. This integral does not have a closed form solution for $\rho \neq 0, 1$.

Note that if $\rho = 0$ then $r = 2$ and we again arrive at the product of independent gamma densities.

Conditional distributions

Since the marginals do not have a closed form, neither do the conditional distributions. However, given specific values of one or more components, it is possible to evaluate the conditional densities empirically, up to an unknown constant. It is thus possible to simulate from the conditional distributions and then calculate desired quantities from the simulations.

Manuscript received, 2016-03-23, revised, 2016-06-22, accepted, 2016-06-26.