

# Bayesian non-parametric inference for multivariate peaks-over-threshold models

Peter Trubey and Bruno Sansó

Department of Statistics, University of California Santa Cruz, 1156 High Street, Santa Cruz, 95064, CA, USA.

\*Corresponding author(s). E-mail(s): [ptrubey@ucsc.edu](mailto:ptrubey@ucsc.edu);  
Contributing authors: [bruno@soe.ucsc.edu](mailto:bruno@soe.ucsc.edu);

## Abstract

We consider a constructive definition of the multivariate Pareto that factorizes the random vector into a radial component and an independent angular component; The former following a univariate Pareto distribution, and the latter defined on the surface of the positive orthant of the infinity norm unit hypercube. In this paper, we propose a method for inferring the distribution of the angular component. We identify its support as the limit of the positive orthant of the unit  $p$ -norm spheres, and introduce a projected gamma family of distributions defined through the normalization a vector of independent gamma random variables to the  $p$ -norm sphere. This family serves as a building block for a flexible family of distributions obtained as a Dirichlet process mixture of projected gammas. For model assessment, we discuss scoring methods appropriate to distributions on the unit hypercube. In particular, working with the energy score criterion, we develop a kernel metric that produces a proper scoring rule, and present a simulation study to compare different modeling choices using the proposed scoring rules. Finally we apply our approach to describe the dependence structure of the extreme values of the magnitude of the integrated vapor transport (IVT), data describing the flow of atmospheric moisture along the coast of California for the years of 1979 through 2020. We find clear but heterogeneous geographical dependence.

## 1 Introduction

The statistical analysis of extreme values focuses on inference for rare events that correspond to the tails of probability distributions. As such, it is a key ingredient in the risk assessment of phenomena that can have strong societal impacts like floods, heat waves, high concentration of pollutants, crashes in the financial markets, among others. The fundamental challenge of extreme value theory (EVT) is to use information, collected over limited periods of time, to extrapolate to long time horizons. This sets EVT apart from most of statistical inference, where the focus is on the bulk of the distribution. Extrapolation to the

tails of the distributions is possible thanks to theoretical results that give asymptotic descriptions of the probability distributions of extreme values.

Inferential methods for the extreme values of univariate observations are well established and software is widely available (see, for example, [Coles, 2001](#)). For variables in one dimension the application of EVT methods considers the asymptotic distribution of either the maxima calculated for regular blocks of data, or the values that exceed a certain threshold. The former leads to a Generalized Extreme Value (GEV) distribution, that depends on three parameters. The latter leads to a Generalized Pareto (GP) distribution, that depends on a shape and a scale parameter. Likelihood-based approaches to inference can be

readily implemented in both cases. In the multivariate case the GEV theory is well developed (see, for example [De Haan and Ferreira, 2006](#)), but the inferential problem is complicated by the fact that there is no parametric representation of the GEV. This problem is inherited by the peaks over threshold (PoT) approach and compounded by the fact that there is no unique definition of an exceedance of a multivariate threshold, as there is an obvious dependence on the norm that is used to measure the size of a vector.

During the last decade or so, much work has been done in the exploration of the definition and properties of an appropriate generalization of the univariate GP distribution to a multivariate setting. To mention some of the papers on the topic, the work of ([Rootzén and Tajvidi, 2006](#)) defines the generalized Pareto distribution, with further analysis on these classes of distributions presented in [Falk and Guillou \(2008\)](#) and [Michel \(2008\)](#). A recent review of the state of the art in multivariate peaks over threshold modelling using generalized Pareto is provided in [Rootzén et al \(2018\)](#) while [Rootzén et al \(2018\)](#) provides insight on the theoretical properties of possible parametrizations. These are used in [Kiriliouk et al \(2019\)](#) for likelihood-based models for PoT estimation. A frequently used method for describing the dependence in multivariate distributions is to use a copula. [Renard and Lang \(2007\)](#), and [Falk et al \(2019\)](#) provide successful examples of this approach in an EVT framework.

[Ferreira and de Haan \(2014\)](#) presents a constructive definition of the Pareto process, that generalizes the GP to an infinite dimensional setting. It consists of decomposing the process into independent radial and angular components. Such an approach can be used in the finite dimensional case, where the angular component contains the information pertaining to the dependence structure of the random vector. Based on this definition, we present a novel approach for modelling the angular component with families of distributions that provide flexibility and can be applied in a moderately large dimensional setting. Our focus on the angular measure is similar to that in [Boldi and Davison \(2007\)](#), [Sabourin and Naveau \(2014\)](#) and [Hanson et al \(2017a\)](#), that consider Bayesian non-parametric approaches. Yet, our approach differs in that it is established in the peaks-over-threshold regime, and avoids the problem of dealing with the so called moment

constraint by considering a constructive definition of the multivariate GP, based on the infinity norm. The approach proposed in this paper adds to the growing literature on Bayesian models for multivariate extreme value analysis (see, for example, [Boldi and Davison \(2007\)](#), [Guillette et al \(2011\)](#), [Sabourin and Naveau \(2014\)](#), [Hanson et al \(2017b\)](#)), providing a model that has strong computational advantages due its structural simplicity, achieves flexibility using a mixture model, imposes no moment constraints, and scales well to moderately large dimensions.

The remainder of this paper is outlined as follows. Section 2 comprises a brief review of multivariate PoT, detailing the separation of the radial measure from the angular measure. Section 3 details our approach for estimating the angular measure, based on projecting an arbitrary distribution supported in  $\mathbb{R}_+^d$  onto unit hyper-spheres defined by  $p$ -norms. Section 4 develops criteria for model selection in the support of the angular measure. Section 5 explores the efficacy of the proposed approach on a set of simulated data, and, acknowledging the relevance of extreme value theory to climatological events ([Jentsch et al, 2007](#); [Vousdoukas et al, 2018](#); [Li et al, 2019](#)), estimates the extremal dependence structure for a measure of water vapor flow in the atmosphere, used for identifying atmospheric rivers. Finally, Section 6 presents our conclusions and discussion.

Throughout the paper, we adopt the operators  $\wedge$  to denote minima, and the  $\vee$  to denote maxima. Thus  $\wedge_i s_i = \min_i s_i$ , and  $\vee_i s_i = \max_i s_i$ . These operators can also be applied component-wise between vectors, such as  $\mathbf{a} \wedge \mathbf{b} = (a_1 \wedge b_1, a_2 \wedge b_2, \dots)$ . We use uppercase to indicate random variables, lowercase to indicate points, and bold-face to indicate vectors or matrices thereof.

## 2 A multivariate PoT model

To develop a multivariate PoT model for extreme values consider a  $d$ -dimensional random vector  $\mathbf{W} = (W_1, \dots, W_d)$  with cumulative distribution  $F$ . Following [Rootzén et al \(2018\)](#), assume that there exists sequences of vectors  $\mathbf{a}_n$  and  $\mathbf{b}_n$ , and a  $d$ -variate distribution  $G$  such that  $\lim_{n \rightarrow \infty} F^n(\mathbf{a}_n \mathbf{w} + \mathbf{b}_n) = G(\mathbf{w})$ .  $G$  is a  $d$ -variate generalized extreme value distribution. It follows that

$$\lim_{n \rightarrow \infty} \Pr [\mathbf{a}_n^{-1}(\mathbf{W} - \mathbf{b}_n) \leq \mathbf{w} \mid \mathbf{W} \not\leq \mathbf{b}_n] = \frac{\log G(\mathbf{w} \wedge \mathbf{0}) - \log G(\mathbf{0})}{\log G(\mathbf{0})}$$

$H$  is a multivariate Pareto distribution, and corresponds to a joint distribution conditional on exceeding a multivariate threshold. In addition to a copula function,  $H$  depends on two  $d$ -dimensional vectors of parameters that we denote as  $\boldsymbol{\xi}$  for the shapes and  $\boldsymbol{\sigma}$  for the scales. Rootzén et al (2018) provide a number of stochastic representations for  $H$ . In particular, denote as  $Z$  a random variable with distribution  $H$  where  $\boldsymbol{\xi} = \mathbf{1}$  and  $\boldsymbol{\sigma} = \mathbf{0}$ . Then, Remark 1 justifies the representation in Ferreira and de Haan (2014), giving  $Z = \mathbf{V}R$  where  $R$  and  $\mathbf{V}$  are independent.  $R = \|\mathbf{Z}\|_\infty$  is distributed as a standard Pareto random variable, and  $\mathbf{V} = \mathbf{Z}/\|\mathbf{Z}\|_\infty$  is a random vector in  $\mathbb{S}_\infty^{d-1}$ , the positive orthant of the unit sphere under  $\mathcal{L}_\infty$  norm, with distribution  $\Phi$ . This representation is central to the methods proposed in this paper.  $R$  and  $\mathbf{V}$  are referred to, respectively, as the *radial* and *angular* components of  $H$ . The angular measure controls the dependence structure of  $\mathbf{Z}$  in the tails. In view of this, to obtain a PoT model we seek a flexible model for the distribution of  $\mathbf{V} \in \mathbb{S}_\infty^{d-1}$ .

The approach considered in Rootzén et al (2018) focuses on the limiting conditional distribution  $H$ . An alternative approach consists of assuming that regular variation (see, for example, Resnick, 2008) holds for the limiting distribution of  $\mathbf{W}$ , implying

$$\lim_{r \rightarrow \infty} n \Pr [n^{-1} \mathbf{W} \in A] = \mu(A),$$

where  $\mu$  is termed the exponent measure.  $\mu$  has the homogeneity property  $\mu(tA) = t^{-1} \mu(A)$ . Letting  $\rho = \|\mathbf{W}\|_p, p > 0$  and  $\boldsymbol{\theta} = \mathbf{W}/\rho$ , define  $\Psi(B) = \mu(\{\mathbf{w} : \rho > 1, \boldsymbol{\theta} \in B\})$ , which is referred to as the angular measure. After some manipulations, we obtain that

$$\lim_{r \rightarrow \infty} \Pr [\boldsymbol{\theta} \in A | \rho > r] = \frac{\Psi(A)}{\Psi(\mathbb{S}_p^{d-1})}. \quad (1)$$

Thus, a model for the exponent measure induces a model for the limiting distribution conditional on the observations being above a threshold defined with respect to their  $p$ -norm. The constraint that all marginals of  $\mu$  correspond to a standard Pareto distribution leads to the so called moment constraints on  $\Psi$ . Inference for the limiting distribution of the exceedances needs to account for the normalizing constant in Equation (1), as well the moment constraints.

### 3 Estimation of the angular measure

Our approach to estimating the PoT distribution considers two steps: First we estimate the shape and scale parameters for the multivariate Pareto distribution, using the univariate marginals; Then we focus on the dependence structure in extreme regions by proposing a flexible model for the distribution of  $\mathbf{V}$ . Consider  $\mathbf{w}_i, i = 1, \dots, n$  a collection of realizations of  $\mathbf{W}$ . We start by setting a large threshold  $b_{t,\ell}$  for the  $\ell$ -th marginal,  $\ell = 1, \dots, d$ . Then, the distribution of the observations for the  $\ell$ -th component, conditional on exceeding the threshold, can be approximated as  $1 - (1 + \xi_\ell(w_{i\ell} - b_{t,\ell})/\sigma_\ell)_+^{1/\xi_\ell}$ , where  $(\cdot)_+$  indicates the positive part function. We proceed by setting a threshold equal to the empirical  $(1 - 1/t)$ -quantile. Thus  $b_{t,\ell} = \hat{F}_\ell^{-1}(1 - 1/t)$ . We then use the approximate exceedance distribution to estimate  $\xi_\ell$  and  $\sigma_\ell$ , for each  $\ell$ , using likelihood based methods. Then, in order to estimate the angular distribution, we use those estimates to standardize each of the marginals. The standardization yields

$$z_{i\ell} = \left( 1 + \xi_\ell \frac{w_{i\ell} - b_{t,\ell}}{a_\ell} \right)_+^{1/\xi_\ell}. \quad (2)$$

Note that  $z_{i\ell} > 1$  implies that  $w_{i\ell} > b_{t,\ell}$ , meaning that the observation  $\mathbf{w}_i$  is extreme in the  $\ell$ -th dimension. Thus,  $r_i = \|\mathbf{z}_i\|_\infty > 1$  implies that at least one dimension has an extreme observation, and corresponds to a very extreme observation when  $t$  is large. We focus on the observations that are such that  $r_i > 1$ . These provide a subsample of the standardized original sample. We define  $\mathbf{v}_i = \mathbf{z}_i/r_i \in \mathbb{S}_\infty^{d-1}$ . These vectors are used for the estimation of  $\Phi$ .

#### 3.1 Projected gamma family

The  $\mathcal{L}_p$ -norm, for  $p > 0$  is defined as

$$\|\mathbf{x}\|_p = \left( \sum_{\ell=1}^d |s_\ell|^p \right)^{\frac{1}{p}},$$

The absolute and Euclidean norms are obtained for  $p = 1$  and  $p = 2$  respectively, and the  $\mathcal{L}_\infty$  norm can be obtained as a limit,

$$\|\mathbf{s}\|_\infty = \lim_{p \rightarrow \infty} \|\mathbf{s}\|_p = \bigvee_{\ell=1}^d s_\ell.$$

To obtain a distribution on  $\mathbb{S}_p^{d-1} = \{\mathbf{s} : \mathbf{s} \in \mathbb{R}_+^d, \|\mathbf{s}\|_p = 1\}$  we start with a vector in  $\mathbf{x} \in \mathbb{R}_+^d$ , and normalize it to obtain  $\mathbf{y} = \mathbf{x}/\|\mathbf{x}\|_p \in \mathbb{S}_p^{d-1}$ . A natural distribution to consider in  $\mathbb{R}_+^d$  is given by a product of independent univariate Gamma distributions. Let  $\mathbf{X} \sim \prod_{\ell=1}^d \text{Ga}(X_\ell | \alpha_\ell, \beta_\ell)$ ,  $\alpha_\ell$  and  $\beta_\ell$  are the shape and scale parameters, respectively. For any finite  $p > 0$ , letting  $y_d = (1 - \sum_{\ell=1}^{d-1} y_\ell^p)^{1/p}$ , the transformation

$$T(x_1, \dots, x_d) = \left( \|\mathbf{x}\|_p, \frac{x_1}{\|\mathbf{x}\|_p}, \dots, \frac{x_{d-1}}{\|\mathbf{x}\|_p} \right) = (r, y_1, \dots, y_{d-1}) \quad (3)$$

is invertible with

$$T^{-1}(r, y_1, \dots, y_{d-1}) = \left( ry_1, \dots, ry_{d-1}, r \left( 1 - \sum_{\ell=1}^{d-1} y_\ell^p \right)^{1/p} \right) \quad (4)$$

The Jacobian of the transformation takes the form

$$r^{d-1} \left[ \left( 1 - \sum_{\ell=1}^{d-1} y_\ell^p \right)^{\frac{1}{p}} + \sum_{\ell=1}^{d-1} y_\ell^p \left( 1 - \sum_{l=1}^{d-1} y_l^p \right)^{\frac{1}{p}-1} \right] \quad (5)$$

The normalization provided by  $T$  maps a vector in  $\mathbb{R}_+^d$  onto  $\mathbb{S}_p^{d-1}$ . With a slight abuse of terminology we refer to it as a projection. Using Equations (3)–(5) we have the joint density

$$f(r, \mathbf{y}) = \prod_{\ell=1}^d \left[ \frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} (ry_\ell)^{\alpha_\ell-1} \exp\{-\beta_\ell ry_\ell\} \right] \times r^{d-1} \left[ \left( 1 - \sum_{\ell=1}^{d-1} y_\ell^p \right)^{\frac{1}{p}} + \sum_{\ell=1}^{d-1} y_\ell^p \left( 1 - \sum_{l=1}^{d-1} y_l^p \right)^{\frac{1}{p}-1} \right] \quad (6)$$

Integrating out  $r$  yields the resulting *Projected Gamma* density

$$\text{PG}(\mathbf{y} | \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{\ell=1}^d \left[ \frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} y_\ell^{\alpha_\ell-1} \right] \times \left[ y_d + \sum_{\ell=1}^{d-1} y_\ell^p y_d^{1-p} \right] \times \frac{\prod_{\ell=1}^d \beta_\ell y_\ell}{\left( \sum_{\ell=1}^d \beta_\ell y_\ell \right)^{\sum_{\ell=1}^d \alpha_\ell}} \quad (7)$$

defined for  $\mathbf{y} \in \mathbb{S}_p^{d-1}$ , and for any finite  $p > 0$ . To avoid identifiability problems when estimating the shape and scale parameters, we set  $\beta_1 = 1$ . Núñez-Antonio and Genevro (2019) obtain the density in Equation (7) for  $p = 2$  as a multivariate distribution for directional data, using spherical coordinates. For  $\mathbf{y} \in \mathbb{S}_1^{d-1}$ , and  $\beta_\ell = \beta$  for all  $\ell$ , the density in Equation (7) corresponds to that of a Dirichlet distribution.

The projected gamma family is simple to specify and has very tractable computational properties. Thus, we use it as a building block for the angular measure  $\Phi$  models. To build a flexible family of distributions in  $\mathbb{S}_p^{d-1}$  we consider

mixtures of projected gamma densities defined as

$$f(\mathbf{y}) = \int_{\Theta} \mathcal{P}\mathcal{G}(\mathbf{y} | \boldsymbol{\theta}) dG(\boldsymbol{\theta}), \quad (8)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ . Following a Bayesian non-parametric approach (Ferguson, 1974; Antoniak, 1974; Müller et al, 2015), we assume that  $G$  is drawn from a random measure. In particular, assuming a Dirichlet process prior for  $G$ , we have a hierarchical formulation of the mixture model that, for a vector of observations  $\mathbf{y}_i$ , is given by

$$\mathbf{y}_i \sim \text{PG}(\mathbf{y}_i | \boldsymbol{\theta}_i) \quad \boldsymbol{\theta}_i \sim G \quad G \sim \mathcal{DP}(\eta, G_0) \quad (9)$$

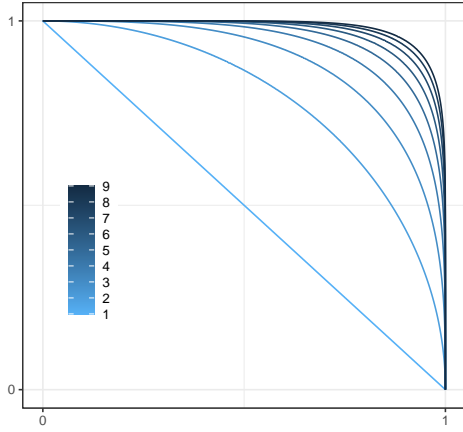
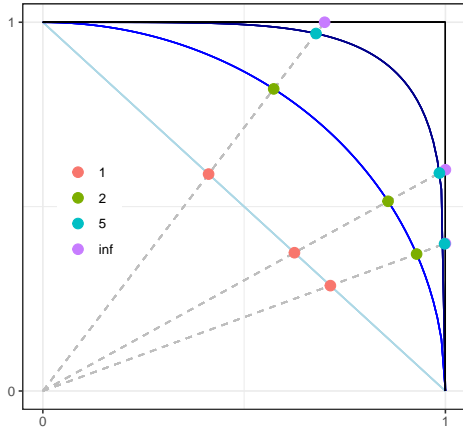
where  $\mathcal{DP}$  denotes a Dirichlet process,  $\eta$  is the precision parameter, and  $G_0$  is the centering distribution.

Unfortunately, in the limit when  $p \rightarrow \infty$ , the normalizing transformation is not differentiable. Thus, a closed form expression like Equation (7) for the projected gamma density on  $\mathbb{S}_\infty^{d-1}$  is not available. Instead, we observe that for a sufficiently large  $p$ ,  $\mathbb{S}_p^{d-1}$  will approach  $\mathbb{S}_\infty^{d-1}$ . With that in mind, our strategy consists of describing the angular distribution  $\Phi$  using a sample based approach with the following steps: (i) Apply the transformation in Equation (2) to the original data; (ii) Obtain the subsample of the standardized observations that satisfy  $R > 1$ ; (iii) Take a finite  $p$  and project the observations onto  $\mathbb{S}_p^{d-1}$ ; (iv) Fit the model in Equation (8) to the resulting data and obtain samples from the fitted model; (v) Project the resulting samples onto  $\mathbb{S}_\infty^{d-1}$ . For step (iv) we use a Bayesian approach that is implemented using a purposely developed Markov chain Monte Carlo sampler described in the next section.

A measure that is used to characterize the strength of the dependence, in the tail, for two random variables  $Z_1$  and  $Z_2$ , with marginal distributions  $F_1$  and  $F_2$  is given by (Coles, 2001)

$$\chi_{12} = \lim_{u \uparrow 1} \Pr [F_1(Z_1) > u | F_2(Z_2) > u].$$

$\chi_{12}$  provides information about the distribution of extremes for the variable  $Z_1$  given that  $Z_2$  is very large. When  $\chi_{12} > 0$ ,  $Z_1$  and  $Z_2$  are said to be asymptotically dependent, otherwise they are asymptotically independent. The following result provides the asymptotic dependence coefficient between two components of  $\mathbf{Z}$  after PoT limit.

(a)  $\mathbb{S}_p^1$  for  $p = 1, \dots, 10$ (b) Projection of data on  $\mathbb{S}_\infty^1$  onto  $\mathbb{S}_p^1$ **Fig. 1:** The positive orthant of the  $p$ -norm sphere for  $d = 2$ .

**Proposition 1.** Suppose that  $\mathbf{Z} = R\mathbf{V}$  with  $R \sim Pa(1)$ ,  $\Pr[V_\ell > 0] = 1$  and  $E[V_\ell]$  exists, for  $\ell = 1, \dots, d$ , then

$$\chi_{j\ell} = E \left[ \frac{V_j}{E[V_j]} \wedge \frac{V_\ell}{E[V_\ell]} \right] \quad (10)$$

*Proof:* Denote as  $F_\ell$  the marginal distribution of  $Z_\ell$ . To obtain  $\chi_{j\ell}$  we need  $\Pr(Z_j > z_j, Z_\ell > z_\ell)$ , where  $z_\ell = F_\ell^{-1}(u) = E[V_\ell] / (1 - u)$ ,  $\ell = 1, \dots, d$ . Using the fact that  $V_\ell > 0, \forall \ell$  almost surely, we have that the former is equal to

$$\Pr \left[ R > \frac{z_j}{V_j} \vee \frac{z_\ell}{V_\ell} \right] = E \left[ 1 \wedge \left( \frac{z_j}{V_j} \vee \frac{z_\ell}{V_\ell} \right)^{-1} \right] = E \left[ \frac{V_j}{z_j} \wedge \frac{V_\ell}{z_\ell} \right]$$

where the second identity is justified by the fact that  $V_i$  is bounded and  $z_i \rightarrow \infty$ . The proof is completed by noting that  $\Pr[F_i(Z_i) > u] = 1 - u$ .  $\square$

Equation (10) implies that  $\chi_{j\ell} > 0$ , and so, no asymptotic independence is possible under our proposed model. For the analysis of extreme values it is of interest to calculate the multivariate conditional survival function. The following result provides the relevant expression, as a function of the angular measure.

**Proposition 2.** Assume the same conditions of Proposition 1. Let  $\alpha \subset \{1, \dots, d\}$  be a collections of indexes. Then

$$\Pr \left[ \bigcap_{\ell \in \alpha} Z_\ell > z_\ell \mid \bigcap_{\ell \notin \alpha} Z_\ell > z_\ell \right] = \frac{E \left[ \bigwedge_{k=1}^d 1 \wedge \frac{V_k}{z_k} \right]}{E \left[ \bigwedge_{k \notin \alpha} 1 \wedge \frac{V_k}{z_k} \right]} \quad (11)$$

The proof uses a similar approach to the proof of Proposition 1.

Equations (10) and (11) provide relevant tools for inference on the tail behavior of the joint distribution of the observations. The expressions can be readily calculated within a sample-based inferential approach like the one considered in the following section.

### 3.1.1 Inference for the projected Gamma mixture model

In a sample-based inference approach, for a given iteration the Dirichlet process mixture model groups observations into stochastically assigned clusters, where members of a cluster share distributional parameters (Müller et al, 2015; Ascolani et al, 2022). Building out the methods of inference for Equation (9), let  $n_j^{-i}$  be the number of observations in cluster  $j$  not including observation  $i$ . Let  $J^{-i}$  be the number of extant clusters, not including any singleton containing observation  $i$ . Under this model, the probability of cluster membership for a given observation is proportional to

$$\Pr[\delta_i = j \mid \dots] \propto \begin{cases} n_j^{-i} \mathcal{P}\mathcal{G}(\mathbf{y}_i \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) \\ \eta \int \mathcal{P}\mathcal{G}(\mathbf{y}_i \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) dG_0(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j), \end{cases}$$

where the top case is for a *new* cluster  $j = 1, \dots, J^{-i}$ , and the bottom case is for a *new* cluster. If  $G_0$  is not a conjugate prior for the kernel

density, the integral in the above formula may not be available in closed form. We sidestep this using Algorithm 8 from Neal (2000): by Monte Carlo integration, we draw  $m$  candidate clusters,  $\alpha_j, \beta_j$  for  $j = J^{-i} + 1, \dots, J^{-i} + m$  from  $G_0$ . Then, we sample the cluster indicator  $\gamma_i$  from extant or candidate clusters, where the probability of cluster membership is proportional to

$$\Pr[\delta_i = j | \dots] \propto \begin{cases} n_j^{-i} \mathcal{P}\mathcal{G}(\mathbf{y}_i | \alpha_j, \beta_j) \\ \frac{n_j}{m} \mathcal{P}\mathcal{G}(\mathbf{y}_i | \alpha_j, \beta_j). \end{cases} \quad (12)$$

Again, the top case is iterating over extant clusters, and now the bottom case is iterating over new *candidate* clusters. If a candidate cluster is selected, then  $\gamma_i = J = J^{-i} + 1$ , and the associated cluster parameters are saved.

A key feature of the the projected Gamma distribution is its computational properties. We augment  $\mathcal{P}\mathcal{G}(\mathbf{y}_i | \alpha_i, \beta_i)$  by introducing a latent radial component  $r_i$ , for each observation. Using Equation (6) we observe that the full conditional of  $r_i$  is easy to sample from, as it is given as

$$r_i | \alpha_i, \beta_i \sim \mathcal{G} \left( r_i \left| \sum_{\ell=1}^d \alpha_{i\ell}, \sum_{\ell=1}^d \beta_{\ell} y_{i\ell} \right. \right). \quad (13)$$

Moreover, the full conditional for  $\alpha_j, \beta_j$  is then proportional to

$$f(\alpha_j, \beta_j | \mathbf{Y}, \mathbf{r}, \delta, \dots) \propto \prod_{i:\gamma_i=j} \prod_{\ell=1}^d \mathcal{G}(r_i y_{i\ell} | \alpha_{j\ell}, \beta_{j\ell}) \times dG_0(\alpha_j, \beta_j). \quad (14)$$

Note that the ordering of the products can be reversed in Equation (14), indicating that with appropriate choice of centering distribution, the full conditionals for  $\alpha_j, \beta_j$  become separable by dimension. We first consider a centering distribution given by a product of independent Gammas:

$$G_0(\alpha_j, \beta_j | \xi, \tau, \zeta, \sigma) = \prod_{\ell=1}^d \mathcal{G}(\alpha_{j\ell} | \xi_{\ell}, \tau_{\ell}) \times \prod_{\ell=2}^d \mathcal{G}(\beta_{j\ell} | \zeta_{\ell}, \sigma_{\ell}). \quad (15)$$

This model is completed with independent Gamma priors on  $\xi_{\ell}, \tau_{\ell}, \zeta_{\ell}, \sigma_{\ell}$ . We also assume a Gamma prior on  $\eta$ , that is updated via the procedure outlined in Escobar and West (1995). We refer to this model as the *projected gamma-gamma* (PG-G) model. An advantage of the PG-G model is that, thanks to conjugacy, the rate

parameters can easily be integrated out for inference on  $\alpha_j$ . Then, the full conditional for  $\alpha_{j\ell}$  takes the form

$$\pi(\alpha_{j\ell} | \mathbf{r}, \mathbf{Y}, \gamma, \xi_{\ell}, \tau_{\ell}, \zeta_{\ell}, \sigma_{\ell}) \propto \frac{\left( \prod_{i:\gamma_i=j} r_i y_{i\ell} \right)^{\alpha_{j\ell}-1} \alpha_{j\ell}^{\xi_{\ell}-1} e^{-\tau_{\ell} \alpha_{j\ell}}}{\Gamma^{n_j}(\alpha_{j\ell})} \times \quad (16)$$

for  $\ell = 2, \dots, d$ . For  $\ell = 1$ , as  $\beta_1 := 1$ , the full conditional takes the simpler form

$$\pi(\alpha_{j1} | \mathbf{r}, \mathbf{Y}, \gamma, \xi_1, \tau_1) \propto \frac{\left( \prod_{i:\gamma_i=j} r_i y_{i1} \right)^{\alpha_{j1}-1} \alpha_{j1}^{\xi_1-1} e^{-\tau_1 \alpha_{j1}}}{\Gamma^{n_j}(\alpha_{j1})}. \quad (17)$$

Samples of  $\alpha_{j\ell}$  can thus be obtained using a Metropolis step. In our analysis, we first transform  $\alpha_{j\ell}$  to the log scale, and use a normal proposal density. The full conditional for  $\beta$  is

$$\beta_{j\ell} | \mathbf{r}, \mathbf{Y}, \alpha, \zeta_{\ell}, \sigma_{\ell} \sim \mathcal{G} \left( \beta_{j\ell} \left| n_j \alpha_{j\ell} + \zeta_{\ell}, \sum_{i:\gamma_i=j} r_i y_{i\ell} + \sigma_{\ell} \right. \right), \quad (18)$$

for  $\ell = 2, \dots, d$ . Updating  $\beta_{j\ell}$  is done via a Gibbs step. The hyper-parameters  $\xi_{\ell}, \tau_{\ell}, \zeta_{\ell}, \sigma_{\ell}$  follow similar gamma-gamma update relationships. We also explore a restricted form of this model, where  $\beta_{\ell} = 1$  for all  $\ell$ . Under this model, we use the full conditional in Equation (17) for all  $\ell$ , and omit inference on  $\zeta, \sigma$ . We refer to this model as the *projected restricted gamma-gamma* (PRG-G)

The second form of centering distribution we explore is a multivariate log-normal distribution on the shape parameters  $\alpha_j$ , with independent gamma  $\beta_{j\ell}$  rate parameters.

$$G_0(\alpha_j, \beta_j | \mu, \Sigma, \zeta, \sigma) = \mathcal{LN}(\alpha_j | \mu, \Sigma) \times \prod_{\ell=2}^d \mathcal{G}(\beta_{j\ell} | \zeta_{\ell}, \sigma_{\ell}). \quad (19)$$

This model is completed with a normal prior on  $(\mu, \Sigma)$ , an inverse Wishart prior on  $\Sigma$ , and Gamma priors on  $\zeta_{\ell}$  and  $\sigma_{\ell}$ , and  $\eta$ . This model is denoted as the *projected gamma-log-normal* (PG-LN) model. We also explore a restricted Gamma form of this model as above, where  $\beta_{\ell} = 1$  for all  $\ell$ . This is denoted as the *projected restricted gamma-log-normal* (PRG-LN) model. Updates for  $\alpha$  can be accomplished using a joint Metropolis step, where  $\beta_{j\ell}$  for  $\ell = 2, \dots, d$  have been integrated

out of the log-density. That is,

$$\pi(\boldsymbol{\alpha}_j \mid \mathbf{Y}, \mathbf{r}, \boldsymbol{\delta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\zeta}, \boldsymbol{\sigma}) \propto \exp \left\{ -\frac{1}{2} (\log \boldsymbol{\alpha}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\alpha}_j - \boldsymbol{\mu}) \right\} \times \frac{1}{\prod_{\ell=1}^d \alpha_{j\ell}} \times \frac{\left( \prod_{i:\gamma_i=j} r_i y_{i1} \right)^{\alpha_{j1}-1}}{\prod_{\ell=1}^d \Gamma^{n_j}(\alpha_{j\ell})} \times \prod_{\ell=1}^d \frac{S_{\text{ES}}(P, \mathbf{x}_i) = \mathbb{E}_p[g(\mathbf{X}_i, \mathbf{x}_i)] - \frac{1}{2} \mathbb{E}_p[g(\mathbf{X}_i, \mathbf{X}'_i)]}{\Gamma(n_j \alpha_{j\ell} + \zeta_\ell)} \quad (20)$$

The inferential forms for  $\beta_{j\ell}$  and its priors are the same as for PG-G. The normal prior for  $\mu$  is conjugate for the log-normal  $\boldsymbol{\alpha}_j$ , and can be sampled via a Gibbs step. Finally, the inverse Wishart prior for  $\boldsymbol{\Sigma}$  is again conjugate to the log-normal  $\boldsymbol{\alpha}_j$ , implying that it can also be sampled via a Gibbs step.

To effectively explore the sample space with a joint Metropolis step, as well as to speed convergence, we implement a parallel tempering algorithm (Earl and Deem, 2005) for the log-normal models. This technique runs parallel MCMC chains at ascending temperatures. That is, for chain  $s$ , the posterior density is exponentiated to the reciprocal of temperature  $t_s$ . For the *cold* chain,  $t_1 := 1$ . Let  $E_s$  be the log-posterior density under the current parameter state for chain  $s$ . Then states between chains  $r, s$  are exchanged via a Metropolis step with probability

$$\alpha_{rs} = \min \left[ 1, \exp \left\{ (t_r^{-1} - t_s^{-1})(E_r - E_s) \right\} \right].$$

Higher temperatures serve to *flatten* the posterior distribution, meaning hotter chains have a higher probability of making a given transition, or will make larger transitions. As such, they will more quickly explore the parameter space, and share information gained through state exchange.

## 4 Scoring criteria for distributions on the infinity–norm sphere

In order to assess and compare the estimation of a distribution on  $\mathbb{S}_\infty^{d-1}$  we consider the theory of proper scoring rules developed in Gneiting and Raftery (2007). As mentioned in Section 3.1, our approach does not provide a density on  $\mathbb{S}_\infty^{d-1}$ , restricting our ability to construct model selection criteria to sample-based approaches. To this end, we employ the *energy score* criterion introduced therein.

The energy score criterion defined for a general probability distribution  $P$ , with finite expectation, is developed as

$$S_{\text{ES}}(P, \mathbf{x}_i) = \mathbb{E}_p[g(\mathbf{X}_i, \mathbf{x}_i)] - \frac{1}{2} \mathbb{E}_p[g(\mathbf{X}_i, \mathbf{X}'_i)], \quad (20)$$

where  $g(\cdot, \cdot)$  is a kernel function. The score defined in Equation (20) can be evaluated using samples from  $P$ , with the help of the law of large numbers. Moreover, Theorem 4 in Gneiting and Raftery (2007), states that if  $g(\cdot, \cdot)$  is a negative definite kernel, then  $S(P, \mathbf{x})$  is a *proper* scoring rule. Recall that a real valued function  $g$  is a negative definite kernel if it is symmetric in its arguments, and  $\sum_{i=1}^n \sum_{j=1}^n a_i a_j g(x_i, x_j) \leq 0$  for all positive integers  $n$ , and any collection  $a_1, \dots, a_n \in \mathbb{R}$  such that  $\sum_{i=1}^n a_i = 0$ .

In a Euclidean space, these conditions are satisfied by the Euclidean distance (Berg et al, 1984). However, for observations on different faces of  $\mathbb{S}_\infty^{d-1}$ , the Euclidean distance will under-estimate the geodesic distance, the actual distance required to travel between the two points. Let

$$\mathbb{C}_\ell^{d-1} = \{\mathbf{x} : \mathbf{x} \in \mathbb{S}_\infty^{d-1}, x_\ell = 1\}$$

comprise the  $\ell$ th *face*. For points on the same face, the Euclidean distance corresponds to the length of the shortest possible path in  $\mathbb{S}_\infty^{d-1}$ . For points on different faces, the Euclidean distance is a lower bound for that length.

For a finite  $p$ , the shortest connecting path between two points in  $\mathbb{S}_p^{d-1}$  is the minimum geodesic; its length satisfying the definition of a distance. Thus its length can be used as a negative definite kernel for the purpose of defining an energy score. Unfortunately as  $p \rightarrow \infty$  the resulting surface  $\mathbb{S}_\infty^{d-1}$  is not differentiable, implying that routines to calculate geodesics are not readily available. However, as  $\mathbb{S}_\infty^{d-1}$  is a portion of a  $d$ -cube, we can borrow a result from geometry (Pappas, 1989) stating that the length of the shortest path between two points on a geometric figure corresponds to the length of a straight line drawn between the points on an appropriate unfolding, rotation, or *net* of the figure from a  $d$ -dimensional to a  $d - 1$ -dimensional space. The optimal net will have the shortest straight line between the points, as long as that line is fully contained within such net. As  $\mathbb{S}_\infty^{d-1}$  has  $d$  faces—each face pairwise adjacent, there are  $d!$  possible nets. However, we are only interested in

nets that begin and end on the source and destination faces respectively, reducing the number of nets under consideration to  $\sum_{k=0}^{d-2} \binom{d-2}{k}$ . This is still computationally burdensome for a large number of dimensions. However, we can efficiently establish an upper bound on the geodesic length. We use this upper bound on geodesic distance as the kernel function for the energy score.

To calculate the energy score we define the kernel

$$g(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{c} \in \mathbb{C}_j^{d-1} \cap \mathbb{C}_\ell^{d-1}} \{\|\mathbf{c} - \mathbf{a}\|_2 + \|\mathbf{b} - \mathbf{c}\|_2\}.$$

where  $\mathbf{a} \in \mathbb{C}_\ell^{d-1}$ , and  $\mathbf{b} \in \mathbb{C}_j^{d-1}$ , for  $\ell, j \in \{1, \dots, d\}$ . Evaluating  $g$  as described requires the solution of a  $(d-2)$ -dimensional optimization problem. The following proposition provides a straightforward approach.

**Proposition 3.** *Let  $\mathbf{a} \in \mathbb{C}_\ell^{d-1}$ , and  $\mathbf{b} \in \mathbb{C}_j^{d-1}$ , for  $\ell, j \in \{1, \dots, d\}$ . For  $\ell \neq j$ , the transformation  $P_{j\ell}(\cdot)$  required to rotate the  $j$ th face along the  $\ell$ th axis produces the vector  $\mathbf{b}'$ , where*

$$\mathbf{b}'_i = P_{j\ell}(\mathbf{b}) = \begin{cases} b_i & \text{for } i \neq j, \ell \\ 1 & \text{for } i = \ell \\ 2 - b_\ell & \text{for } i = j \end{cases}. \quad (21)$$

Then  $g(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}'\|_2$ .

*Proof:* Notice that for  $\mathbf{c} \in \mathbb{C}_j^{d-1} \cap \mathbb{C}_\ell^{d-1}$ ,  $\|\mathbf{b} - \mathbf{c}\|_2 = \|\mathbf{b}' - \mathbf{c}\|_2$ . We then have that

$$\begin{aligned} g(\mathbf{a}, \mathbf{b}) &= \min_{\mathbf{c} \in \mathbb{C}_j^{d-1} \cap \mathbb{C}_\ell^{d-1}} \{\|\mathbf{c} - \mathbf{a}\|_2 + \|\mathbf{b} - \mathbf{c}\|_2\} \\ &= \min_{\mathbf{c} \in \mathbb{C}_j^{d-1} \cap \mathbb{C}_\ell^{d-1}} \{\|\mathbf{c} - \mathbf{a}\|_2 + \|\mathbf{b}' - \mathbf{c}\|_2\} \\ &= \|\mathbf{a} - \mathbf{b}'\|_2. \end{aligned}$$

The last equality is due to the fact that  $\mathbf{a}$  and  $\mathbf{b}'$  belong to the same hyperplane.  $\square$

Using the rotation in Proposition 3 we obtain the following result.

**Proposition 4.**  *$g$  is a negative definite kernel.*

*Proof:* For a given  $n$  consider an arbitrary set of points  $\mathbf{a}_1, \dots, \mathbf{a}_n \in \mathbb{S}_\infty^{d-1}$ , and  $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ , such that  $\sum_{i=1}^n \alpha_i = 0$ . Then

$$\sum_{i,j} \alpha_i \alpha_j g(\mathbf{a}_i, \mathbf{a}_j) = \sum_{i,j} \alpha_i \alpha_j \|\mathbf{a}_i - \mathbf{a}'_j\|_2 \leq 0,$$

where  $\mathbf{a}'_j$  is defined as in Proposition 3. The last equality holds as  $\|\mathbf{x} - \mathbf{x}'\|_2$ ,  $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$  is negative definite (Gneiting and Raftery, 2007)  $\square$

Proposition 3 provides a computational efficient way to evaluate the proper scoring rule  $S_{\text{ES}}$  defined on  $\mathbb{S}_\infty^{d-1}$ , for each observation. For the purpose of model assessment and comparison, we report the average  $S_{\text{ES}}$  taken across all observed data, and notice that the smaller the score, the better.

## 5 Data illustrations

We apply the aforementioned models to simulated angular data. We then consider the analysis of atmospheric data. To tackle the difficult problem of assessing the convergence an MCMC chain for a large-dimensional model, we monitor the log-posterior density. In all the examples considered, MCMC samples produced stable traces of the log-posterior in less than 40 thousand iterations. We use that as a burn-in, and thereafter sample 10 thousand additional iterations. We then thin the chain by retaining one every ten samples, to obtain 1000 total samples. These are used to generate samples from the posterior predictive densities. We used two different strategies to implement the MCMC samplers. For the models whose DP prior is centered around a log-normal distribution we used parallel tempering. This serves to overcome the very slow mixing that we observed for these cases. The temperature ladder was set as  $t_s = 1.3^s$ , for  $s \in \{0, 1, \dots, 5\}$ . This was set empirically in order to produce acceptable swap probabilities both for the simulated data, and real data. Parallel tempering produces chains with good mixing properties, but has a computational cost that grows linearly with the number of temperatures. Thus, for the gamma-centered models, we used a single chain. We leverage the fast speed of each iteration, to obtain a large number of samples, that are appropriately thinned to deal with a mild autocorrelation. In summary, the strategy for log-normal centered models is based on a costly sampler with good mixing properties. The strategy for the gamma-centered models is based on a cheap sampler that can be run for a large number of iterations.

Our hyperprior parameters are set as follows: for the gamma-centered models (PG-G, PRG-G), the shape parameter for the centering distribution  $\xi_\ell \sim \mathcal{G}(1, 1)$ , and rate parameter  $\tau_\ell \sim \mathcal{G}(2, 2)$ . For the log-normal centered



models (PG–LN, PRG–LN), the centering distribution’s log-mean  $\boldsymbol{\mu} \sim \mathcal{N}_d(0, \mathbf{I}_d)$ , and covariance matrix  $\boldsymbol{\Sigma} \sim \mathcal{IW}(d+10, (d+10)\mathbf{I}_d)$ . These values are intended such that draws from the prior for  $\boldsymbol{\Sigma}$  will weakly tend towards the identity matrix. For models learning rate parameters  $\beta_{j\ell}$  (PG–G, PG–LN), the centering distribution’s shape parameter  $\zeta \sim \mathcal{G}(1, 1)$  and rate parameter  $\sigma \sim \mathcal{G}(2, 2)$ . The choice of the  $\mathcal{G}(2, 2)$  for rate parameters places little mass near 0, in order to draw estimates for the value away from 0 for numerical stability.

---

**Algorithm 1** Simulated Angular Dataset Generation Routine.  $\mu_j, \Sigma_j$  are the parameters of the mixture component distribution;  $\pi$  is the probability vector assigning weight mixture components;  $\delta_i$  is the mixture component identifier for each simulated observation.

---

```

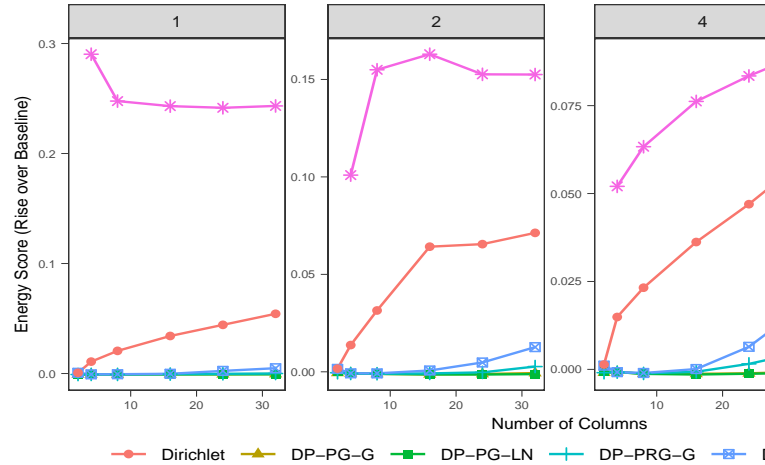
for  $n_{\text{iter}}$  in  $[1, \dots, 10]$  do
  for  $n_{\text{mix}}$  in  $[1, 2, 4, 8]$  do
    for  $j$  in  $1, \dots, n_{\text{mix}}$  do
      Generate  $\mu_j \sim \mathcal{N}_{32}(\mathbf{0}, \mathbf{I})$ 
      Generate  $\Sigma_j \sim \mathcal{IW}_{32}(70, 70\mathbf{I})$ 
    end for
    Generate  $\pi \sim \text{Dirichlet}(\mathbf{10}_{n_{\text{mix}}})$ 
    for  $i$  in  $1, \dots, 1000$  do
      Generate  $\delta_i \sim \text{Categorical}(\pi)$ 
      Generate  $\mathbf{X}_i \sim \mathcal{LN}(\mu_{[\delta_i]}, \Sigma_{[\delta_i]})$ 
    end for
    for  $n_{\text{col}}$  in  $[2, 4, 8, 16, 24, 32]$  do
      Project columns 1 to  $n_{\text{col}}$  of  $\mathbf{X}$  onto
       $\mathbb{S}_{\infty}^{n_{\text{col}}-1}$  and save.
    end for
  end for
end for

```

---

## 5.1 Simulation Study

To evaluate our proposed approach for angular measure estimation we consider simulated datasets on  $\mathbb{S}_{\infty}^{d-1}$ , for values of  $d$  ranging between 2 and 32. We generated each dataset as a mixture of multivariate log-normal distributions, projected onto  $\mathbb{S}_{\infty}^{d-1}$ . The generation procedure is detailed in Algorithm 1. We produced ten replicates of each configuration. We consider two gamma-centered and two log-normal centered DP mixture models, with and without restrictions in each case. To perform a comparative analysis we fitted the pairwise betas model proposed in Cooley et al (2010). We chose this model for comparison as it is implemented in the readily available package `BMamevt`



**Fig. 2:** Average energy score rise over baseline (on  $\mathbb{S}_{\infty}^{d-1}$ ) for various models fitted to simulated data, with ascending count of mixture components (indicated by plot heading) and number of dimensions (indicated by horizontal axis). Note that pairwise betas is a moment-restricted model.

in R (Sabourin, 2023). `BMamevt` provides samples from the predictive posterior distribution. These are needed for the calculation of the energy scores that are at the basis of our comparison. In addition, `BMamevt` can be fitted to moderately large multivariate observations. For the DP mixture models, the data are projected onto  $\mathbb{S}_{10}^{d-1}$ . For the other two models they are projected on  $\mathbb{S}_1^{d-1}$ . We sampled each model for 50,000 iterations, dropping the first 40,000 as burn-in, and thinning to keep every 10th iteration after. These settings were intended to provide a consistent sampling strategy that would work with every model, even if inefficient for some.

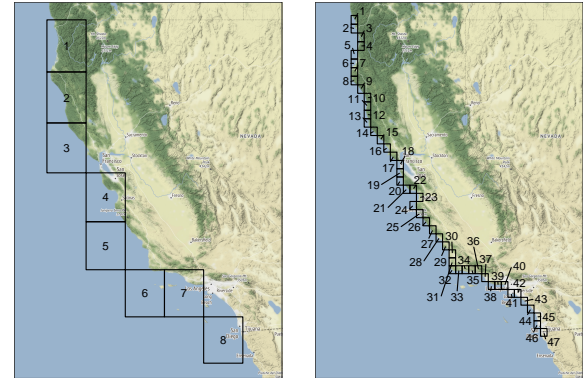
Figure 2 shows the average rise over baseline in energy score as calculated on  $\mathbb{S}_{\infty}^{d-1}$  using the kernel metric described in Proposition 3, for models trained on simulated data. After training a model, a posterior predictive dataset is generated, and the energy score is calculated as a Monte Carlo approximation of Equation (20). In our analysis, after burn-in and thinning, we had 1,000 replicates from the posterior distribution, and generated 10 posterior predictive replicates per iteration. The *baseline* value is the energy score of a new dataset from the same generating distribution as the training dataset, evaluated against the training dataset. For the simulated data, we observe that the projected gamma models dominate the other two options considered,

regardless of the choice of centering distribution. The projected restricted gamma models with a multivariate log-normal centering distribution appears to be dominated by the models based on the alternative centering distributions. Moreover, the performance deteriorates with the increase in dimensionality. Additionally, models centered around the log-normal distribution incur in the computational cost of multivariate normal evaluation and parallel tempering, taking approximately six times longer to sample relative to the gamma models. We also note that the computational cost of the pairwise betas model grows combinatorially, with a sample space of dimension  $\binom{d}{2} + 1$ . In our testing, `BMamevt` is substantially faster than any of our proposed DP mixture models for low dimensions, however for examples with high number of dimensions, the computational time for `BMamevt` was comparable or greater than that for `DP-PG-G`. We compare computing times in our data analysis in Table 1b.

## 5.2 Integrated Vapor Transport

The *integrated vapor transport* (IVT) is a two component vector that tracks the flow of the total water volume in a column of air over a given area (Ralph et al, 2017). IVT is increasingly used in the study of atmospheric rivers because of its direct relationship with orographically induced precipitation (Neiman et al, 2009). Atmospheric rivers (AR) are elongated areas of high local concentration of water vapor in the atmosphere that transport water from the tropics around the world. AR can cause extreme precipitation, something that is usually associated with very large values of the IVT magnitude over a whole geographical area. In spite of this, AR are fundamental for the water supply of areas like California. Thus the importance of understanding the extreme behavior of IVT, including extreme tail dependence. We consider datasets that correspond to IVT estimated at two different spatial resolutions. The coarse resolution dataset is obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim reanalysis (ERA-Interim) (Berrisford et al, 2011; Dee et al, 2011). The high resolution dataset corresponds to the latest ECMWF observational product, ERA5 (Hersbach et al, 2020).

Our data correspond to daily average values for the IVT magnitude along the coast of California. The ERA-Interim data used covers the



**Fig. 3:** Grid cell locations for ERA-Interim (left) and ERA5 (right).

time period 1979 through 2014 (37 years) omitting leap days, and eight grid cells that correspond to the coast of California. The ERA5 data cover the time period 1979 through 2019 (42 years) with the same restriction, and 47 grid cells for the coast of California. This gives us the opportunity to illustrate the performance of our method in multivariate settings of very different dimensions. Figure 3 provides a visual representation of the area these grid cells cover.

---

**Algorithm 2** Data preprocessing to isolate and transform data exhibiting extreme behavior.  $r_i$  represents the radial component, and  $v_i$  the angular component. The declustering portion is relevant for data correlated in time.

---

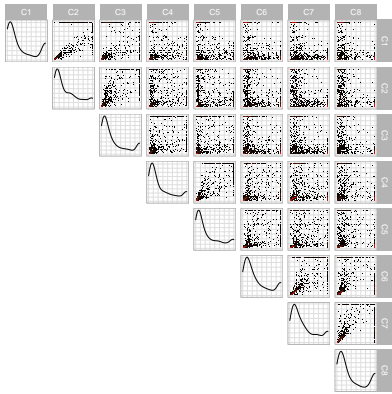
```

for  $\ell = 1, \dots, d$  do
  Set  $b_{t,\ell} = \hat{F}_\ell^{-1}(1 - \frac{1}{t})$ .
  With  $\mathbf{x}_\ell > b_{t,\ell}$ , fit  $a_\ell, \xi_\ell$  via MLE according
  to generalized Pareto likelihood.
end for
for  $i = 1, \dots, n$  do
  Define  $z_{i,\ell} = \left(1 + \xi_\ell \frac{x_{i,\ell} - b_{t,\ell}}{a_\ell}\right)_+^{1/\xi_\ell}$ ; then
   $r_i = \|\mathbf{z}_i\|_\infty, \quad v_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_\infty}$ 
end for
Subset  $\mathbf{r}, \mathbf{v}$  such that  $r_i \geq 1$ 
if declustering then
  for  $i = 1, \dots, n$  do
    If  $r_i \geq 1$  and  $r_{i-1} \geq 1$ , drop the lesser
    (and associated  $v_i$ ) from data set.
  end for
end if

```

---

Fitting our models to the IVT data requires some pre-processing. First, we subset the data to



**Fig. 4:** Pairwise plots from ERA-Interim data after transformation and projection to  $\mathcal{S}_\infty^7$ . Down the diagonal are marginal kernel densities, with two-dimensional histograms on the off-diagonal. In those plots, red indicates a higher density. All data is between 0 and 1.

the rainy season, which in California runs roughly from November to March. Following the approach described in Section 3 we estimate the shape and scale parameters of a univariate GP, in each dimension, using maximum likelihood. We set the threshold in each dimensions  $\ell$  as  $b_{t,\ell} = \hat{F}_\ell^{-1}(1 - t^{-1})$ , where  $\hat{F}$  is the empirical CDF and  $t = 20$ , that corresponds to the 95 percentile. We then use the transformation in Equation (2) to standardize the observations. Dividing each standardized observation by its  $\mathcal{L}_\infty$  norm, we obtain a projection onto  $\mathcal{S}_\infty^{d-1}$ . As the data correspond to a daily time series, the observations are temporally correlated. For each group of consecutive standardized vectors  $z_i$  such that  $\|z_i\|_\infty > 1$ , we retain only the vector with the largest  $\mathcal{L}_\infty$  norm. The complete procedure is outlined in Algorithm 2.

After subsetting the ERA-Interim data to the rainy season we have 5587 observations. After the processing and declustering described in Algorithm 2, this number reduces to 511 observations. A pairwise plot of the transformed data after processing and declustering is presented in Figure 4. From this, we note that the marginal densities display strong similarities, with a large spike near 0 and a small spike near 1. A value of 1 in a particular axis indicates that the standardized threshold exceedance was largest in that dimension. The off-diagonal plots correspond to pairwise density plots. We observe that some site pairs, such as (1, 2), (7, 8), and especially (4, 5) have the bulk of their data concentrated in a small arc along the

**Table 1:** Model fit assessment and computation time on ERA-Interim and ERA5 data.

Source	Pairwise Betas	PG-G	PG-LN	PRG-G	PRG-LN
ERA-Interim	0.8620	0.8003	0.7986	<b>0.7966</b>	0.7970
ERA5	2.0311	1.6404	1.5576	<b>1.4349</b>	1.5051

(a) Energy score criterion from fitted models against the IVT data. L better.

Source	Pairwise Betas	PG-G	PG-LN	PRG-G	PRG-LN
ERA-Interim	1.5	16.3	66.5	14.8	52.9
ERA5	53.1	19.4	153.4	24.6	121.4

(b) Time to sample (in minutes) 50,000 iterations for various mod

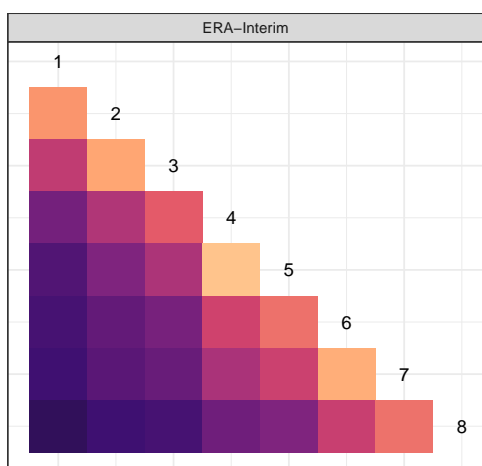
45°, while other site combinations such as (3, 6), (2, 7), or (1, 8) the data is split, favoring one side or the other of the 45° line. For the ERA5 data, after subsetting we have 6342 observations, which reduces to 532 observations after processing and declustering. We fit the PG-G, PRG-G, PG-LN, and PRG-LN models to both datasets.

Table 1a shows the values of the estimated energy scores for the different models considered. We observe that, contrary to the results in the simulation study in Figure 2, the preferred model is the projected restricted gamma models, though for the lower-dimensional ERA-Interim data, all models perform comparably. Table 1b shows the computing times needed to fit the different models to the two datasets. We see the effect of dimensionality on the various models; for gamma centered models it grows linearly; for the log-normal centered model, it will grow superlinearly as matrix inversion becomes the most costly operation. For **BMamevt**, its parameter space grows combinatorically with the number of dimensions, and thus so does computational complexity and sampling time.

We consider an exploration of the pairwise extremal dependence using Monte Carlo estimates of the coefficients in Equation (10). For this we use samples obtained from the PRG-G model. Figure 5 provides a graphical analysis of the results. The coefficients achieve values between 0.286 and 0.759 for the ERA-Interim data and between 0.181 and 0.840 for the ERA5 data. The greater range in dependence scores observed with

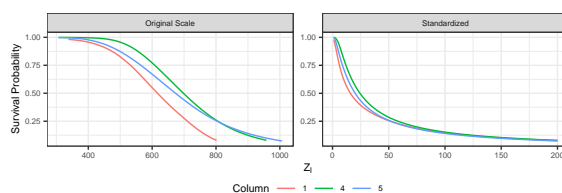
the ERA5 data versus ERA-Interim speaks to the greater granularity of the ERA5 data, indicating that distance between locations is a strong contributor to the strength of the pairwise asymptotic dependence. The highest coefficients are 0.759 for locations 4 and 5 in the ERA-Interim data and 0.840 for locations 1 and 2 in the ERA5 data. Clearly, pairwise asymptotic dependence coefficients tell a limited story, as a particular dependence may include more than two locations. We can, however, glean some information from the patterns that emerge in two dimensions. For the ERA-Interim data, we observe a possible cluster between cells 5-8, indicating a strong dependence among these cells. Analogously, for the ERA5 data, we observe three possible groups of locations.

Figure 6 shows, for the ERA-Interim data under the PRG-G model, the conditional survival curve defined in Equation (11), for one dimension, conditioned on all other dimensions being greater than their (fitted) 90th percentile. Figure 7 presents the bi-variate conditional survival function, conditioning on all other dimensions. These results illustrate quantitatively how extremal dependence affects the shape of the conditional survival curves. The two top panels represent the joint survival function between grid locations 4 and 5, which are shown in Figure 5 to exhibit strong extremal dependence. We observe that the joint survival surface is strongly convex. The bottom panels represent the joint survival surface between grid locations 1 and 5, which exhibited low extremal dependence. In this case

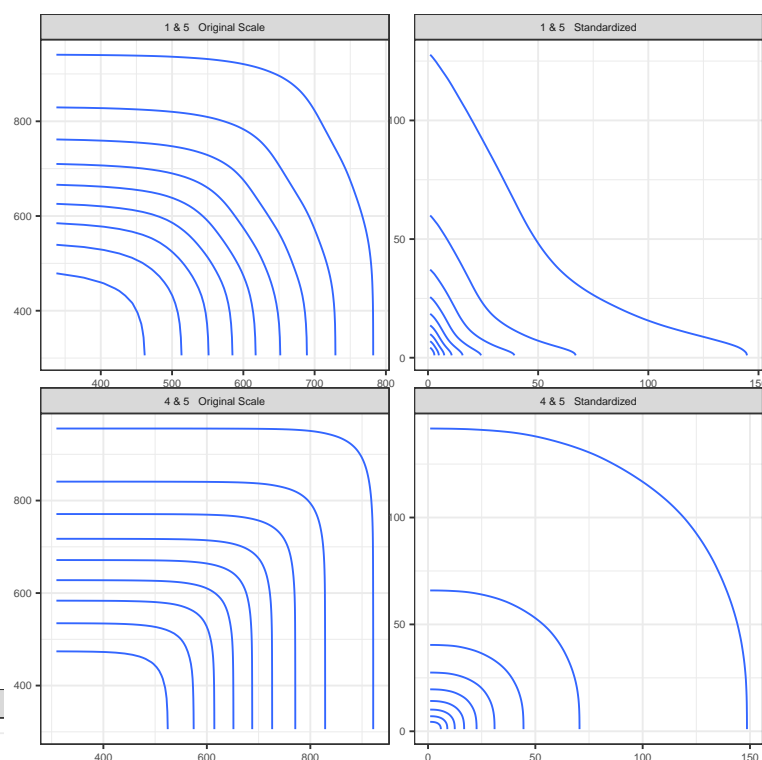


**Fig. 5:** Pairwise extremal dependence coefficients for IVT data using the PRG-G model.

*BNP inference for multivariate PoT*



**Fig. 6:** Conditional survival curves for selected locations, using ERA-Interim, and PRG-G model, conditioning on all other dimensions at greater than 90th percentile (fitted). The left panel uses original units. Right panel uses standardized units.



**Fig. 7:** Pairwise conditional survival curves for selected locations, using ERA-Interim, and PRG-G model, conditioning on all other dimensions at greater than 90th percentile (fitted).

the shape of the contours tend to be concave, quite different from the shapes observed in the top panels.

Using our proposed scoring criteria, we explored the effect of the choice of  $p$  on the final results. Using the simulated data, generated from a mixture of projected Gammas, we were unable to observe sizeable differences in the scores for  $p$  ranging between 1 and 15. However, for the

IVT data, we observed a drop in the energy score associated with higher  $p$ , with diminishing effect as  $p$  increased. We observed no significant differences in the performance of the model that uses  $p = 10$ , which corresponds to the analysis presented, relative to the one that uses  $p = 15$ .

## 6 Conclusion

In this paper, we have built upon the definition of the multivariate Pareto described in [Ferreira and de Haan \(2014\)](#) to establish a useful representation of its dependence structure through the distribution of its angular component, which is supported on the positive orthant of the unit hypersphere under the  $\mathcal{L}_\infty$  norm,  $\mathbb{S}_\infty^{d-1}$ . Due to the inherent difficulty of obtaining the likelihood of distributions with support on  $\mathbb{S}_\infty^{d-1}$  our method transforms the data to  $\mathbb{S}_p^{d-1}$ , fits then using mixtures of products of independent gammas, then transforms the predictions back to  $\mathbb{S}_\infty^{d-1}$ . As  $\mathbb{S}_p^{d-1}$  converges to  $\mathbb{S}_\infty^{d-1}$  as  $p \rightarrow \infty$ , we expect the proposed resampling to be efficient for large enough  $p$ . In fact, our exploration of the simulated and real data indicates that the procedure is robust to the choice of moderately large values of  $p$ . Our method includes two inferential steps. The first consists of the estimation of the marginal Pareto distributions; the second consists of the estimation of the angular density. Parameter uncertainty incurred in the former is not propagated to the latter. Conceptually, an integrated approach that accounts for all the estimation uncertainty is conceivable. Unfortunately, this leads to posterior distributions with complex data dependent restrictions that are very difficult to explore, especially in large dimensional settings. In fact, our attempts to fit a simple parametric model for the marginals and the angular measure jointly in several dimensions were not successful.

In this paper we have focused on a particular representation of the multivariate Pareto distribution for PoT inference on extreme values. To this end, our model provides a computationally efficient and flexible approach. An interesting extension of the proposed model is to consider regressions of extreme value responses, due to extreme value inputs following the ideas in [de Carvalho et al \(2022\)](#). This will produce PoT based Bayesian non-parametric extreme value regression models. More generally, models that allow for covariate-dependent extremal dependence ([Mhalla et al, 2019](#)) could be considered. In

addition, we notice that our approach is based on flexibly modeling angular distributions for any  $p$ -norm. As such, it can be applied to other problems focused on high dimensional directional statistics constrained to a cone of directions.

Establishing the angular measure specifically in  $\mathbb{S}_\infty^{d-1}$  provides two benefits over  $\mathbb{S}_p^{d-1}$ . First, the transformation to  $\mathbb{S}_\infty^{d-1}$  is unique. Observe in Equation (4), establishing  $y_d$  as a function of  $y_1, \dots, y_{d-1}$  can be replaced in form with any  $y_\ell$ , indicating there are  $d$  equivalent transformations, each yielding a different Jacobian and, for  $p > 1$ , potentially resulting in a different calculated density. Second, evaluation of geodesic distance on  $\mathbb{S}_p^{d-1}$  is not straightforward. However, we have demonstrated a computationally efficient upper bound on geodesic distance on  $\mathbb{S}_\infty^{d-1}$ . Accepting these foibles, it would be interesting to explore the distribution on  $\mathbb{S}_p^{d-1}$ ,

The computations in this paper were performed on a desktop computer with an AMD Ryzen 5000 series processor. The program is largely single-threaded, so computation time is not dependent on available core count. In each case, we run the MCMC chain for 50 000 iterations, with a burn-in of 40 000 samples. Fitting the PG-G model on the ERA5 dataset took approximately 15 minutes. Work is in progress to optimize the code, and explore parallelization where possible. We are also exploring alternative computational approaches that will make it feasible to tackle very high dimensional problems, such as variational Bayes. In fact, to elaborate on the study of IVT, there is a need to consider several hundreds, if not thousands, of grid cells over the Pacific Ocean in order to obtain a good description of atmospheric events responsible for large storm activity over California.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. SES-2050012 and DMS-2153277.

## References

- Antoniak CE (1974) Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The Annals of Statistics* 2(6):1152–1174. <https://doi.org/10.1214/aos/1176342871>

- Ascolani F, Lijoi A, Rebaudo G, et al (2022) Clustering consistency with Dirichlet process mixtures. *Biometrika* 110(2):551–558. <https://doi.org/10.1093/biomet/asac051>, <https://arxiv.org/abs/https://academic.oup.com/biomet/article-pdf/110/2/551/50311302/asac051.pdf>
- Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups: theory of positive definite and related functions, vol 100. Springer
- Berrisford P, Kållberg P, Kobayashi S, et al (2011) Atmospheric conservation properties in ERA-interim. *Quarterly Journal of the Royal Meteorological Society* 137(659):1381–1399. <https://doi.org/10.1002/qj.864>
- Boldi MO, Davison AC (2007) A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 69(2):217–229
- de Carvalho M, Kumukova A, Dos Reis G (2022) Regression-type analysis for multivariate extreme values. *Extremes* pp 1–28
- Coles SG (2001) An Introduction to Statistical Modelling of Extreme Values. Springer, <https://doi.org/10.1007/978-1-4471-3675-0>
- Cooley D, Davis RA, Naveau P (2010) The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis* 101(9):2103–2117. <https://doi.org/https://doi.org/10.1016/j.jmva.2010.04.007>
- De Haan L, Ferreira A (2006) Extreme value theory: an introduction, vol 21. Springer, <https://doi.org/10.1007/0-387-34471-3>
- Dee DP, Uppala S, Simmons A, et al (2011) The ERA-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society* 137(656):553–597. <https://doi.org/10.1002/qj.828>
- Earl DJ, Deem MW (2005) Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics* 7(23):3910–3916
- Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90(430):577–588. <https://doi.org/10.1080/01621459.1995.10476550>
- Falk M, Guillou A (2008) Peaks-over-threshold stability of multivariate generalized Pareto distributions. *Journal of Multivariate Analysis* 99:715–734. <https://doi.org/10.1016/j.jmva.2007.03.009>
- Falk M, Padoan SA, Wisheckel F (2019) Generalized Pareto copulas: A key to multivariate extremes. *Journal of Multivariate Analysis* 174:104,538. <https://doi.org/10.1016/j.jmva.2019.104538>
- Ferguson TS (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics* 2:615–629. <https://doi.org/10.1214/aos/1176342752>
- Ferreira A, de Haan L (2014) The generalized Pareto process; with a view towards application and simulation. *Bernoulli* 20(4):1717–1737. <https://doi.org/10.3150/13-BEJ538>
- Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102(477):359–378. <https://doi.org/10.1198/016214506000001437>
- Guillotte S, Perron F, Segers J (2011) Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3):377–406
- Hanson TE, de Carvalho M, Chen Y (2017a) Bernstein polynomial angular densities of multivariate extreme value distributions. *Statistics and Probability Letters* 128:60–66. <https://doi.org/10.1016/j.spl.2017.03.030>
- Hanson TE, de Carvalho M, Chen Y (2017b) Bernstein polynomial angular densities of multivariate extreme value distributions. *Statistics & Probability Letters* 128:60–66. <https://doi.org/10.1016/j.spl.2017.03.030>
- Hersbach H, Bell B, Berrisford P, et al (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*

- 146(730):1999–2049. <https://doi.org/10.1002/qj.3803>
- Jentsch A, Kreyling J, Beierkuhnlein C (2007) A new generation of climate-change experiments: events, not trends. *Frontiers in Ecology and the Environment* 5(7):365–374. [https://doi.org/10.1890/1540-9295\(2007\)5\[365:ANGOCE\]2.0.CO;2](https://doi.org/10.1890/1540-9295(2007)5[365:ANGOCE]2.0.CO;2)
- Kiriliouk A, Rootzén H, Segers J, et al (2019) Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics* 61(1):123–135. <https://doi.org/10.1080/00401706.2018.1462738>
- Li C, Zwiers F, Zhang X, et al (2019) Larger increases in more extreme local precipitation events as climate warms. *Geophysical Research Letters* 46(12):6885–6891. <https://doi.org/10.1029/2019GL082908>
- Mhalla L, de Carvalho M, Chavez-Demoulin V (2019) Regression-type models for extremal dependence. *Scandinavian Journal of Statistics* 46(4):1141–1167
- Michel R (2008) Some notes on multivariate generalized Pareto distributions. *Journal of Multivariate Analysis* 99(6):1288–1301. <https://doi.org/10.1016/j.jmva.2007.08.007>
- Müller P, Quintana FA, Jara A, et al (2015) *Bayesian nonparametric data analysis*, vol 1. Springer
- Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9(2):249–265. <https://doi.org/10.1080/10618600.2000.10474879>
- Neiman PJ, White AB, Ralph FM, et al (2009) A water vapour flux tool for precipitation forecasting. In: *Proceedings of the Institution of Civil Engineers-Water Management*, Thomas Telford Ltd, pp 83–94, <https://doi.org/10.1680/wama.2009.162.2.83>
- Núñez-Antonio G, Geneyro E (2019) A multivariate projected Gamma model for directional data. *Communications in Statistics - Simulation and Computation* pp 1–22. <https://doi.org/10.1080/03610918.2019.1612910>
- Pappas T (1989) *The joy of mathematics: Discovering mathematics all around you*. Wide World Pub Tetra
- Ralph FM, Iacobellis S, Neiman P, et al (2017) Dropsonde observations of total integrated water vapor transport within north Pacific atmospheric rivers. *Journal of Hydrometeorology* 18(9):2577–2596. <https://doi.org/10.1175/JHM-D-17-0036.1>
- Renard B, Lang M (2007) Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources* 30(4):897–912. <https://doi.org/10.1016/j.advwatres.2006.08.001>
- Resnick S (2008) *Extreme Values, Regular Variation, and Point Processes*. Applied probability, Springer
- Rootzén H, Tajvidi N (2006) Multivariate generalized Pareto distributions. *Bernoulli* 12(5):917–930. <https://doi.org/10.3150/bj/1161614952>
- Rootzén H, Segers J, Wadsworth JL (2018) Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis* 165:117–131. <https://doi.org/10.1016/j.jmva.2017.12.003>
- Rootzén H, Segers J, Wadsworth JL (2018) Multivariate peaks over thresholds models. *Extremes* 21(1):115–145. <https://doi.org/10.1007/s10687-017-0294-4>
- Sabourin A (2023) *BMamevt: Multivariate Extremes: Bayesian Estimation of the Spectral Measure*. URL <https://CRAN.R-project.org/package=BMamevt>, r package version 1.0.5
- Sabourin A, Naveau P (2014) Bayesian Dirichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis* 71:542–567. <https://doi.org/10.1016/j.csda.2013.04.021>
- Vousdoukas MI, Mentaschi L, Voukouvalas E, et al (2018) Global probabilistic projections of extreme sea levels show intensification of coastal flood hazard. *Nature communications* 9(1):1–12. <https://doi.org/10.1038/s41467-018-04692-w>