UNIVERSITY OF CALIFORNIA
SANTA CRUZ

EXPLORING MULTIVARIATE EXTREME VALUE THEORY WITH APPLICATIONS TO ANOMALY DETECTION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Peter Trubey

March 2025

The Dissertation of Peter Trubey
is approved:

Professor Bruno Sansó, Chair

Professor Juhee Lee

Professor Robert Lund

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Exploring Multivariate Extreme Value Theory with Applications to Anomaly Detection

by

Peter Trubey

Significant work has been done in the field of extreme analysis in the form of generalization of the univariate generalized Pareto distribution to a multivariate setting. We consider the constructive definition of the multivariate Pareto that factorizes a Pareto random vector into independent radial and angular components; the former following a Pareto distribution, the latter following a distribution with no closed form with support on the surface of the positive orthant of the $\mathcal{L}_\infty$-norm unit hypercube. In this document, we propose a method of inferring this angular distribution, as a realization of a Bayesian non-parametric mixture of independent random gamma vectors, projected onto an arbitrary $\mathcal{L}_p$-norm unit hypersphere; the support of which will approach the support of the angular component as $p \to \infty$.

We explore applications of this BNP mixture of projected gammas in characterizing the dependence structure of extremes; the motivating example of such we examine is the integrated vapor transport, data pertaining to an atmospheric river transporting moisture from the Pacific ocean across California. We observe clear but heterogeneous geographic dependence. Second, we consider the application of the BNP mixture of projected gammas to a novelty detection setting, developing novelty scores appropriate to the support. To expand the applicability of our methods, we develop a categorical data model, and consider the extension of the angular novelty scores to categorical, and mixed data settings. We find that our model and scores compare favorably to canonical novelty scores on canonical novelty detection datasets. Finally,

we seek to understand the limitations of BNP mixture of projected gammas, by attempting to apply the model at a large scale—applied to storm surge data at specified locations, as simulated under the Sea, Lakes, and Overland Surges due to Hurricanes (SLOSH) model. We observe issues in model fidelity, in terms of recovering the marginal distributions, or capturing the dependence structure in a highly multivariate setting. We observe that as dimensionality increases, the number of extant clusters decreases. To ameliorate this loss of granularity, a regression model is proposed, that invokes a low-dimensional representation of the output space. We use these models to explore storm surge at sites of critical infrastructure in the Delaware Bay watershed.

Chapter 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 including 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 modeling 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 modeling 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. (2017), that consider Bayesian non-parametric approaches. Yet, our approach differs in that it is established in the peaks-over-threshold regime, and uses a constructive definition of the multivariate GP, based on the infinity norm. The approach explored in the following chapters adds to the growing literature on Bayesian models for multivariate extreme value analysis (see, for example, Boldi and Davison (2007), Guillotte et al. (2011), Sabourin and Naveau (2014), Hanson et al. (2017)), providing a model that has strong computational advantages due its structural simplicity, achieves flexibility using a mixture model, and scales well to moderately large dimensions. We consider applications of this model to inference on multivariate extremes, as well as anomaly detection, and evaluate limitations of this model as well.

Throughout this document, 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 $\boldsymbol{a} \wedge \boldsymbol{b} = (a_1 \wedge b_1, a_2 \wedge b_2, \ldots)$. Similarly, we apply inequality and arithmetic operators to vectors, for example $\boldsymbol{a} \leq \boldsymbol{b}$, and interpret them component-wise. We use uppercase to indicate random variables, lowercase to indicate points, and bold-face to indicate vectors or matrices thereof.

1.1 A multivariate PoT model

To develop a multivariate PoT model for extreme values, consider a D-dimensional random vector $\boldsymbol{W} = (W_1, \ldots, W_D)$ with cumulative distribution F. A common assumption on $\boldsymbol{W}$ is that it is in the so called domain of attraction of a multivariate max-stable distribution, G. Thus, following Rootzén et al. (2018), assume that there exists sequences of vectors $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$, such that $\lim_{n\to\infty} F^n(\boldsymbol{a}_n \boldsymbol{w} + \boldsymbol{b}_n) = G(\boldsymbol{w})$. G is a D-variate generalized extreme value distribution. Notice that, even though the univariate marginals are obtained from a three parameter family,

there is no parametric form to represent G. Taking logarithms and expanding, we have that

$$\lim_{n\to\infty} n(1 - F(\boldsymbol{a}_n\boldsymbol{w} + \boldsymbol{b}_n)) = -\log G(\boldsymbol{w}) \quad \forall\, \boldsymbol{w} \in \mathbb{R}^D,$$

such that $G(\boldsymbol{w}) > 0$. It follows that

$$\begin{aligned}\lim_{n\to\infty} \Pr\left[\boldsymbol{a}_n^{-1}(\boldsymbol{W} - \boldsymbol{b}_n) \leq \boldsymbol{w} \mid \boldsymbol{W} \not\leq \boldsymbol{b}_n\right] &= \frac{-\log G(\boldsymbol{w} \wedge \boldsymbol{0}) - (-\log G(\boldsymbol{w}))}{-\log G(\boldsymbol{0})} \\ &= \frac{\log G(\boldsymbol{w} \wedge \boldsymbol{0}) - \log G(\boldsymbol{w})}{\log G(\boldsymbol{0})} \;=\; H(\boldsymbol{w})\end{aligned}$$

where $\boldsymbol{a}_n^{-1}$ indicates element-wise inversion, and $\{\boldsymbol{W} \not\leq \boldsymbol{b}_n\}$ denotes the set where at least one component of $\boldsymbol{W}$ is above the corresponding component of $\boldsymbol{b}_n$. Rootzén et al. (2018) define H as a multivariate Pareto distribution.. It corresponds to a joint distribution conditional on exceeding a multivariate threshold, defined by a non-parametric function governing the multivariate dependence, and two D-dimensional vectors of parameters that control the shapes and scales of the marginals. We denote these as $\boldsymbol{\xi}$ for the shapes and $\boldsymbol{\sigma}$ for the scales. Rootzén et al. (2018) provides a number of stochastic representations for H. Throughout this work, we focus on a particular representation proposed in Ferreira and de Haan (2014). To this end, we denote as $\boldsymbol{Z}$ a random variable with distribution H where $\boldsymbol{\xi}$, and $\boldsymbol{\sigma}$ both equal $\boldsymbol{1}$. Then, $\boldsymbol{Z} = R\boldsymbol{V}$, where R, the radial component, and $\boldsymbol{V}$, the angular component, are independent. $R = \|\boldsymbol{Z}\|_\infty = \vee_{d=1}^{D} Z_i$ is distributed as a univariate standard Pareto random variable, and $\boldsymbol{V} = \boldsymbol{Z}/\|\boldsymbol{Z}\|_\infty$ is a random vector in $\mathbb{S}_\infty^{D-1}$, the positive orthant of the unit sphere under $\mathcal{L}_\infty$ norm, with distribution Φ. This representation is central to the methods proposed in this work. R and $\boldsymbol{V}$ are referred to, respectively, as the *radial* and *angular* components of H. The angular measure controls the dependence structure of $\boldsymbol{Z}$ in the tails. In view of this, to obtain a PoT model we seek a flexible model for the distribution of $\boldsymbol{V} \in \mathbb{S}_\infty^{D-1}$, based on a Bayesian non-parametric model.

The approach considered in Rootzén et al. (2018) focuses on the limiting conditional distribution H. An alternative approach to obtaining a limiting PoT distribution consists of assuming that regular variation (see, for example, Resnick, 2008) holds for the limiting distribution

of $\boldsymbol{W}$, implying that

$$\lim_{n\to\infty} n\Pr\left[n^{-1}\boldsymbol{W} \in A\right] = \mu(A),$$

for some measure μ that is referred to as the exponent measure. μ has the homogeneity property $\mu(tA) = t^{-1}\mu(A)$. Letting $\rho = \|W\|_p, p > 0$ and $\boldsymbol{\theta} = \boldsymbol{W}/\rho$, define

$$\Psi(B) = \mu(\{\boldsymbol{w} : \rho > 1,\ \boldsymbol{\theta} \in B\}),$$

, which is referred to as the angular measure. After some manipulations, we obtain that

$$\lim_{r\to\infty} \Pr\left[\boldsymbol{\theta} \in A | \rho > r\right] = \frac{\Psi(A)}{\Psi(\mathbb{S}_p^{D-1})}. \tag{1.1}$$

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 μ correspond to a standard Pareto distribution leads to the so called moment constraints on Ψ, consisting of

$$\int_{\mathbb{S}_p^{D-1}} w_d \, \mathrm{d}\Psi(\boldsymbol{w}) = \frac{1}{D}, \quad d = 1, \ldots, D.$$

Inference for the limiting distribution of the exceedances needs to account for the normalizing constant in Equation (1.1), as well the moment constraints. Because of these issues, we prefer to follow the limiting conditional distribution approach. An example of the application of the regular variation approach, using $p = 1$ is developed in Sabourin and Naveau (2014).

A brief rundown of this document is as follows: This chapter has established the theoretical basis to separate the independent radial and angular measures of extreme random vectors. Chapter 2 builds on this separation of radial and angular components by creating an angular distribution on an arbitrary p-norm unit sphere, and using it to model the angular component of extreme random vectors. In particular, Section 2.1 establishes an angular distribution based on a projection of independent gamma variables as a base of a BNP mixture model; Section 2.2

introduces an efficient means of evaluating model fidelity in $\mathbb{S}_{\infty}^{D_1}$; and Section 2.3 explores the application of this model to extreme values in the integrated vapor transport, an atmospheric river carrying moisture off the Pacific ocean over California and inland. Chapter 3 explores the application of the aforementioned model and kernel metric in a novelty detection setting. In particular, Section 3.3 establishes the novelty detection setting and assumptions used, and the methods applied; Section 3.4, expands those methods to include categorical and multinomial data; along with the unification of both scoring regimes. Chapter 4 attempts to explore limitations of the model developed in Chapter 2, by employing it on a greater scale than has been previously attempted. In particular, it considers a variational approximation of the model, and evaluates model fidelity under a significantly increased scale in terms of number of observations, and number of dimensions; Section 4.3 expands on previous work by developing a regression model by using a BNP mixture of projected gammas as its base; and Section 4.4 applies the model to extreme values in storm surge, as simulated using the SLOSH model. Finally, Chapter 5 summarizes the conclusions we reached in Chapters 2–4, and explores possible ameliorations of the shortcomings we observed in Chapter 4.

Chapter 2

Bayesian Non-Parametric Inference for Multivariate Peaks-Over-Threshold Models

2.1 Estimation of the angular measure

To infer the PoT distribution we consider 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 $\boldsymbol{V}$. Consider $\boldsymbol{w}_n,\ n = 1, \ldots, N$ a collection of realizations of $\boldsymbol{W}$. We start by setting a large threshold. We define the threshold using the empirical $(1 - 1/t)$-quantile of the dth marginal, for a large t. Then t corresponds to the so called "return period", and we take the threshold as $b_{t,d} = \hat{F}_d^{-1}(1 - 1/t)$ for the dth marginal, $d = 1, \ldots, D$. Then, the distribution of W_d, conditional on exceeding the threshold, can be approximated with a generalized univariate

Pareto. Thus,

$$\Pr[W_d > w_{nd} \mid W_d > b_{t,d}] = \left(1 + \xi_d \frac{w_{nd} - b_{t,d}}{\sigma_d}\right)_+^{-1/\xi_d}$$

where $(\cdot)_+$ indicates the positive part function. We then estimate ξ_d and σ_d, for each d, using likelihood based methods. To estimate the angular distribution, we standardize each of the marginals. The standardization yields

$$z_{nd} = \left(1 + \xi_d \frac{w_{nd} - b_{t,d}}{\sigma_d}\right)_+^{1/\xi_d} ; \qquad \boldsymbol{z}_n = [z_{n1}, \ldots, z_{nD}]^T . \tag{2.1}$$

Note that $z_{nd} > 1$ implies that $w_{nd} > b_{t,d}$, meaning that the observation $\boldsymbol{w}_n$ is extreme in the dth dimension. Thus, $r_n = \|\boldsymbol{z}_n\|_\infty > 1$ 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 $r_n > 1$. These provide a sub-sample of the standardized original sample. We define $\boldsymbol{v}_n = \boldsymbol{z}_n / r_n \in \mathbb{S}_\infty^{D-1}$. These vectors are used for the estimation of Φ.

2.1.1 Projected gamma family

At the core of our PoT method is the development of a distribution on

$$\mathbb{S}_p^{D-1} = \left\{\boldsymbol{y} : \boldsymbol{y} \in \mathbb{R}_+^d,\ \|\boldsymbol{y}\|_p = 1\right\},$$

where, for $p > 0$, $\|\cdot\|_p$ is the $\mathcal{L}_p$-norm, of a vector $\boldsymbol{x} \in \mathbb{R}^D$, defined as

$$\|\boldsymbol{x}\|_p = \left(\sum_{d=1}^S |x_d|^p\right)^{\frac{1}{p}} .$$

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,

$$\|\boldsymbol{x}\|_\infty = \lim_{p \to \infty} \|\boldsymbol{x}\|_p = \bigvee_{d=1}^D x_d.$$

To obtain a distribution on $\mathbb{S}_p^{D-1}$ we start with a vector in $\boldsymbol{x} \in \mathbb{R}_+^D$, and normalize it to obtain $\boldsymbol{y} = \boldsymbol{x}/\|\boldsymbol{x}\|_p \in \mathbb{S}_p^{D-1}$. A natural distribution to consider in $\mathbb{R}_+^D$ is given by a product of independent univariate Gamma distributions. Let $\boldsymbol{X} \sim \prod_{d=1}^D \mathcal{G}(X_d \mid \alpha_d, \beta_d)$, where α_s and β_s are the

shape and scale parameters, respectively. For any finite $p > 0$, letting $y_d = (1 - \sum_{d=1}^{D-1} y_d^p)^{1/p}$, the transformation

$$T(x_1, \ldots, x_d) = \left(\|\boldsymbol{x}\|_p, \frac{x_1}{\|\boldsymbol{x}\|_p}, \ldots, \frac{x_{D-1}}{\|\boldsymbol{x}\|_p} \right) = (r, y_1, \ldots, y_{D-1}) \tag{2.2}$$

is invertible with

$$T^{-1}(r, y_1, \ldots, y_{D-1}) = \left(ry_1, \ldots, ry_{D-1}, r\left(1 - \textstyle\sum_{d=1}^{D-1} y_d^p\right)^{\frac{1}{p}} \right). \tag{2.3}$$

The Jacobian of the transformation takes the form

$$r^{D-1}\left[\left(1 - \textstyle\sum_{d=1}^{D-1} y_d^p\right)^{\frac{1}{p}} + \sum_{d=1}^{D-1} y_d^p \left(1 - \sum_{l=1}^{D-1} y_d^p\right)^{\frac{1}{p}-1} \right]. \tag{2.4}$$

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 (2.2)–(2.4) we have the joint density

$$f(r, \boldsymbol{y}) = \prod_{d=1}^{d} \left[\frac{\beta_d^{\alpha_d}}{\Gamma(\alpha_d)} (ry_d)^{\alpha_d - 1} \exp\{-\beta_d r y_d\} \right] \times r^{D-1} \left[y_D + \textstyle\sum_{d=1}^{D-1} y_d^p y_D^{1-p} \right]. \tag{2.5}$$

Integrating out r yields the resulting *Projected Gamma* density

$$\mathcal{PG}(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{d=1}^{D} \left[\frac{\beta_d^{\alpha_d}}{\Gamma(\alpha_d)} y_d^{\alpha_d - 1} \right] \times \left[y_D + \textstyle\sum_{d=1}^{D-1} y_d^p y_D^{1-p} \right] \times \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)}{\left(\sum_{d=1}^{D} \beta_d y_d\right)^{\sum_{d=1}^{D} \alpha_d}}, \tag{2.6}$$

defined for $\boldsymbol{y} \in \mathbb{S}_p^{D-1}$, 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 Geneyro (2019) obtain the density in Equation (2.6) for $p = 2$ as a multivariate distribution for directional data, using spherical coordinates. For $\boldsymbol{y} \in \mathbb{S}_1^{D-1}$, and $\beta_d = \beta$ for all d, the density in Equation (2.6) 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 Φ models. To build a flexible family of distributions in $\mathbb{S}_p^{D-1}$ we consider mixtures of projected gamma densities defined as

$$f(\boldsymbol{y}) = \int_{\Theta} \mathcal{PG}(\boldsymbol{y} \mid \boldsymbol{\theta}) dG(\boldsymbol{\theta}), \tag{2.7}$$

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 $\boldsymbol{y}_n$, is given by

$$\boldsymbol{y}_n \sim \mathrm{PG}(\boldsymbol{y}_n \mid \boldsymbol{\theta}_n); \qquad \boldsymbol{\theta}_n \sim G; \qquad G \sim \mathcal{DP}(\eta, G_0) \tag{2.8}$$

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

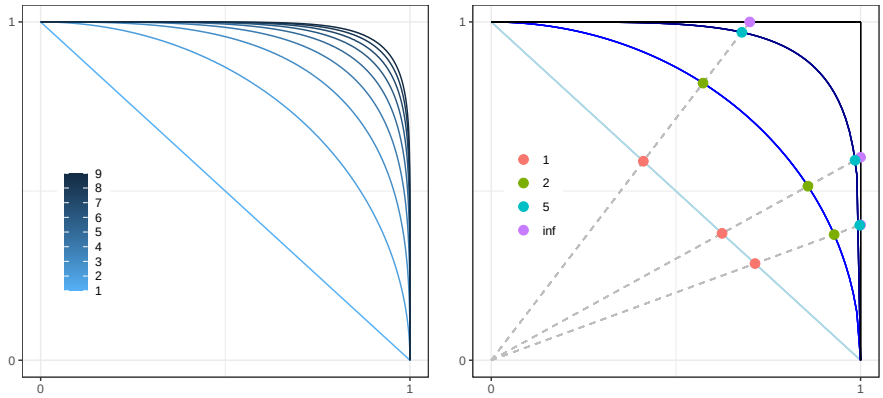
Unfortunately, in the limit when $p \to \infty$, the normalizing transformation is not differentiable. Thus, a closed form expression like Equation (2.6) for the projected gamma density on $\mathbb{S}_{\infty}^{D-1}$ is not available. Instead, we observe that for a sufficiently large p, $\mathbb{S}_{p}^{D-1}$ will approach $\mathbb{S}_{\infty}^{D-1}$. With that in mind, our strategy consists of describing the angular distribution Φ using a sample based approach with the following steps: (i) Apply the transformation in Equation (2.1) 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 (2.7) 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.

2.1.2 Tail probabilities for the PoT model

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):

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

χ_{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



(a) $\mathbb{S}_p^1$ for $p = 1, \ldots, 10$ (b) Projecting data on $\mathbb{S}_\infty^1 \to \mathbb{S}_p^1$

Figure 2.1: The positive orthant of the p-norm sphere for $D = 2$.

they are asymptotically independent. The following result provides the asymptotic dependence coefficient between two components of $\boldsymbol{Z}$ for our proposed PoT model.

Proposition 1. *Suppose that* $\boldsymbol{Z} = R\boldsymbol{V}$ *with* $R = ||Z||_\infty \sim Pa(1)$, $Pr[V_d > 0] = 1$ *and* $E[V_d]$ *exists, for* $d = 1, \ldots, D$, *then*

$$\chi_{de} = E\left[\frac{V_d}{E[V_d]} \wedge \frac{V_e}{E[V_e]}\right] \tag{2.9}$$

Proof: Denote as F_d the marginal distribution of Z_d. Observe that

$$\Pr(Z_d > z_d) = \Pr(RV_d > z_d) = \Pr\left(R > \frac{z_d}{V_d}\right) = \mathrm{E}\left[\frac{V_d}{z_d} \wedge 1\right],$$

where the expectation is taken with respect to V_d. Recall that $0 < V_d \leq 1$, almost surely. We are looking at the limiting behavior as $z_d \to \infty$, thus

$$\Pr(Z_d > z_d) = \frac{\mathrm{E}[V_d]}{z_d} \implies F_d(z_d) = \Pr(Z_d \leq z_d) = 1 - \frac{\mathrm{E}[V_d]}{z_d} = u$$
$$\implies z_d = F_d^{-1}(u) = \frac{\mathrm{E}[V_d]}{1-u}.$$

To obtain χ_{de} we need $Pr(Z_d > z_d, Z_e > z_e)$, where $z_d = \mathrm{E}[V_d]/(1-u)$, $d = 1, \ldots, D$. Using the fact that $V_d > 0$, $\forall\, d$ almost surely, we have that the former is equal to

$$\Pr\left[R > \frac{z_d}{V_d} \vee \frac{z_e}{V_e}\right] = \mathrm{E}\left[1 \wedge \left(\frac{z_d}{V_d} \vee \frac{z_e}{V_e}\right)^{-1}\right] = \mathrm{E}\left[\frac{V_d}{z_d} \wedge \frac{V_e}{z_e}\right] = (1-u)\mathrm{E}\left[\frac{V_d}{\mathrm{E}[V_d]} \wedge \frac{V_e}{\mathrm{E}[V_e]}\right]$$

where the second identity is justified by the fact that V_d, V_e are bounded and $z_d, z_e \to \infty$. The proof is completed by noting that, for the denominator in the conditional probability, and $\forall d$, $\Pr\left[F_d(Z_d) > u\right] = 1 - u$. □

Equation (2.9) implies that $\chi_{de} > 0$, 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 $\alpha \subset \{1, \ldots, D\}$ be a collections of indexes. Then*

$$\Pr\left[\bigcap_{d\in\alpha} Z_d > z_d \mid \bigcap_{d\notin\alpha} Z_d > z_d\right] = \frac{E\left[\bigwedge_{d=1}^{D} 1 \wedge \frac{V_d}{z_d}\right]}{E\left[\bigwedge_{d\notin\alpha} 1 \wedge \frac{V_d}{z_d}\right]}. \tag{2.10}$$

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

Equations (2.9) and (2.10) 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.

2.1.2.1 Inference for the projected Gamma mixture model

To perform inference for our proposed PoT model we develop a iterative sample-based approach. We implement a Markov chain Monte Carlo method that, for a given iteration, 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 (2.8), let $C_j^{(-n)}$ be the number of observations in cluster j not including observation n. Let $J^{(-n)}$ be the number of extant clusters, not including any singleton containing observation n. Under this model, the probability of cluster membership for a given observation is proportional

to

$$\Pr\left[\gamma_n = \jmath \mid \ldots\right] \propto \begin{cases} C_\jmath^{(-n)} \mathcal{PG}\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath\right) & \text{for } \jmath = 1, \ldots, J^{(-n)} \\ \eta \int \mathcal{PG}\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath\right) dG_0(\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath) & \text{for } \jmath = J^{(-n)} + 1, \end{cases}$$

where the top case is iterating over extant clusters, 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, $\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath$ for $\jmath = J^{(-n)} + 1, \ldots, J^{(-n)} + m$ from G_0. Then, we sample the cluster indicator γ_n from extant or candidate clusters, where the probability of cluster membership is proportional to

$$\Pr\left[\gamma_n = \jmath \mid \ldots\right] \propto \begin{cases} C_\jmath^{(-n)} \mathcal{PG}\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath\right) & \text{for } \jmath = 1, \ldots, J^{(-n)} \\ \frac{\eta}{m} \mathcal{PG}\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath\right) & \text{for } \jmath = J^{(-n)} + 1, \ldots, J^{(-n)} + m. \end{cases} \tag{2.11}$$

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 $\gamma_n = J = J^{(-n)} + 1$, and the associated cluster parameters are saved.

A key feature of the the projected Gamma distribution is its computational properties. We augment $\mathcal{PG}(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_{\gamma_n}, \boldsymbol{\beta}_{\gamma_n})$ by introducing a latent radial component r_n, for each observation. Using Equation (2.5) we observe that the full conditional of r_n is easy to sample from, as it is given as

$$r_n \mid \boldsymbol{\alpha}_{\gamma_n}, \boldsymbol{\beta}_{\gamma_n} \sim \mathcal{G}\left(r_n \,\middle|\, \sum_{d=1}^{D} \alpha_{\gamma_n d}, \sum_{d=1}^{D} \beta_{\gamma_n d} y_{nd}\right). \tag{2.12}$$

Moreover, the full conditional for $\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath$ is then proportional to

$$f(\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath \mid \boldsymbol{Y}, \boldsymbol{r}, \boldsymbol{\gamma}, \ldots) \propto \prod_{n:\gamma_n = \jmath} \prod_{d=1}^{D} \mathcal{G}\left(r_n y_{nd} \mid \alpha_{\jmath d}, \beta_{\jmath d}\right) \times dG_0(\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath). \tag{2.13}$$

Note that the ordering of the products can be reversed in Equation (2.13), indicating that with appropriate choice of centering distribution, the full conditionals for $\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath$ can become separable

by dimension, and thus inference on $\alpha_{\jmath d}, \beta_{\jmath d}$ can be done in parallel for all $\jmath$, d. We first consider a centering distribution given by a product of independent Gammas:

$$G_0(\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath \mid \boldsymbol{\xi}, \boldsymbol{\tau}, \boldsymbol{\zeta}, \boldsymbol{\sigma}) = \prod_{d=1}^{D} \mathcal{G}(\alpha_{\jmath d} \mid \xi_d, \tau_d) \times \prod_{d=2}^{D} \mathcal{G}(\beta_{\jmath d} \mid \zeta_d, \sigma_d). \tag{2.14}$$

This model is completed with independent Gamma priors on ξ_d, τ_d, ζ_d, σ_d. We also assume a Gamma prior on η, 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 $\beta_{\jmath d}$ can easily be integrated out for inference on $\boldsymbol{\alpha}_\jmath$. Then, the full conditional for $\alpha_{\jmath d}$ takes the form

$$\begin{aligned} \pi(\alpha_{\jmath d} \mid \boldsymbol{r}, \boldsymbol{Y}, \boldsymbol{\gamma}, \xi_d, \tau_d, \zeta_d, \sigma_d) \propto & \left[\frac{\left(\prod_{n:\gamma_n=\jmath} r_n y_{nd} \right)^{\alpha_{\jmath d}-1} \alpha_{\jmath d}^{\xi_d - 1} e^{-\tau_d \alpha_{\jmath d}}}{\Gamma^{C_\jmath}(\alpha_{\jmath d})} \right] \\ & \times \left[\frac{\Gamma\left(C_\jmath \alpha_{\jmath d} + \zeta_d\right)}{\left(\sum_{n:\gamma_n=\jmath} r_n y_{nd} + \sigma_d \right)^{C_\jmath \alpha_{\jmath d} + \zeta_d}} \right] \end{aligned} \tag{2.15}$$

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

$$\pi(\alpha_{\jmath 1} \mid \boldsymbol{r}, \boldsymbol{Y}, \boldsymbol{\gamma}, \xi_1, \tau_1) \propto \frac{\left(\prod_{n:\gamma_n=\jmath} r_n y_{n1} \right)^{\alpha_{\jmath 1}-1} \alpha_{\jmath 1}^{\xi_1 - 1} e^{-\tau_1 \alpha_{\jmath 1}}}{\Gamma^{C_\jmath}(\alpha_{\jmath 1})}. \tag{2.16}$$

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

$$\beta_{\jmath d} \mid \boldsymbol{r}, \boldsymbol{Y}, \alpha, \zeta_d, \sigma_d \sim \mathcal{G}\left(\beta_{\jmath d} \,\middle|\, C_\jmath \alpha_{\jmath d} + \zeta_d, \sum_{n:\gamma_n=\jmath} r_n y_{nd} + \sigma_d \right), \tag{2.17}$$

for $d = 2, \ldots, D$. Updating $\beta_{\jmath d}$ is done via a Gibbs step. The hyper-parameters $\xi_d, \tau_d, \zeta_d, \sigma_d$ follow similar gamma-gamma update relationships. We also explore a restricted form of this model, where $\beta_d = 1$ for all d. Under this model, we use the full conditional in Equation (2.16) for all d, and omit inference on $\boldsymbol{\zeta}, \boldsymbol{\sigma}$. We refer to this model as the *projected restricted gamma-gamma* (PRG-G) model.

The second form of centering distribution we explore is a multivariate log-normal distribution on the shape parameters $\boldsymbol{\alpha}_\jmath$, with independent gamma $\beta_{\jmath d}$ rate parameters.

$$G_0\left(\boldsymbol{\alpha}_\jmath, \boldsymbol{\beta}_\jmath \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \zeta, \sigma\right) = \mathcal{LN}\left(\boldsymbol{\alpha}_\jmath \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \times \prod_{d=2}^{D} \mathcal{G}\left(\beta_{\jmath d} \mid \zeta_d, \sigma_d\right). \tag{2.18}$$

This model is completed with a normal prior on $\boldsymbol{\mu}$, an inverse Wishart prior on $\boldsymbol{\Sigma}$, and Gamma priors on ζ_d, σ_d, and η. 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 $\beta_d = 1$ for all d. This is denoted as the *projected restricted gamma–log-normal* (PRG-LN) model. Updates for $\boldsymbol{\alpha}$ can be accomplished using a joint Metropolis step, where $\beta_{\jmath d}$ for $d = 2, \ldots, D$ have been integrated out of the log-density. That is,

$$\begin{aligned}\pi(\boldsymbol{\alpha}_\jmath \mid \boldsymbol{Y}, \boldsymbol{r}, \boldsymbol{\gamma}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\zeta}, \boldsymbol{\sigma}) \propto \exp&\left\{-\frac{1}{2}(\log \boldsymbol{\alpha}_\jmath - \mu)^T \boldsymbol{\Sigma}^{-1} (\log \boldsymbol{\alpha}_\jmath - \mu)\right\} \\ &\times \frac{\left(\prod_{n:\gamma_n=\jmath} r_n y_{n1}\right)^{\alpha_{\jmath 1}-1}}{\prod_{d=1}^{D} \alpha_{\jmath d} \Gamma^{C_\jmath}(\alpha_{\jmath d})} \\ &\times \prod_{d=2}^{D} \frac{\Gamma\left(C_\jmath \alpha_{\jmath d} + \zeta_d\right)}{\left(\sum_{n:\gamma_n=\jmath} r_n y_{nd} + \sigma_d\right)^{C_\jmath \alpha_{\jmath d} + \zeta_d}}\end{aligned} \tag{2.19}$$

The inferential forms for $\beta_{\jmath d}$ and its priors are the same as for PG-G. The normal prior for μ is conjugate for the log-normal $\boldsymbol{\alpha}_\jmath$, and can be sampled via a Gibbs step. Finally, the inverse Wishart prior for Σ is again conjugate to the log-normal $\boldsymbol{\alpha}_\jmath$, 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 i, the posterior density is exponentiated to the reciprocal of temperature t_i. For the *cold* chain, $t_1 := 1$. Let E_i be the log-posterior density under the current parameter state for chain i, and $\boldsymbol{\theta}_i$ the current *state* of chain i. Then states between chains i, k are exchanged

via a Metropolis step with probability

$$\Pr\left[\,\boldsymbol{\theta}_i \leftrightarrow \boldsymbol{\theta}_k\,\right] = \min\left[1, \exp\left\{(t_i^{-1} - t_k^{-1})(E_i - E_k)\right\}\right].$$

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.

2.2 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 2.1.1, 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

$$\mathrm{S}_{\mathrm{ES}}\left(P, \boldsymbol{x}_n\right) = \mathrm{E}_p\left[g\left(\boldsymbol{X}_n, \boldsymbol{x}_n\right)\right] - \frac{1}{2}\mathrm{E}_p\left[g\left(\boldsymbol{X}_n, \boldsymbol{X}_n'\right)\right], \tag{2.20}$$

where g is a kernel function. The score defined in Equation (2.20) 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 $\mathrm{S}(P, \boldsymbol{x})$ 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 $\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m g(x_n, x_m) \leq 0$ for all positive integers N, and any collection $a_1, \ldots, a_N \in \mathbb{R}$ such that $\sum_{n=1}^{N} a_n = 0$.

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

$$\mathbb{C}_d^{D-1} = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{S}_\infty^{D-1}, x_d = 1\}$$

comprise the dth *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 $p \to \infty$ 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

$$g(\boldsymbol{a}, \boldsymbol{b}) = \min_{\boldsymbol{c} \in \mathbb{C}_d^{D-1} \cap \mathbb{C}_e^{D-1}} \left\{ \|\boldsymbol{c} - \boldsymbol{a}\|_2 + \|\boldsymbol{b} - \boldsymbol{c}\|_2 \right\}. \tag{2.21}$$

where $\boldsymbol{a} \in \mathbb{C}_d^{D-1}$, and $\boldsymbol{b} \in \mathbb{C}_e^{D-1}$, for $d, e \in \{1, \ldots, D\}$. Evaluating g as described requires the solution of a $(D-2)$-dimensional optimization problem. The following proposition provides a straightforward approach.

Proposition 3. *Let $\boldsymbol{a} \in \mathbb{C}_d^{D-1}$, and $\boldsymbol{b} \in \mathbb{C}_e^{D-1}$, for $d, e \in \{1, \ldots, D\}$. For $d \neq e$, the transformation $P_{de}(\cdot)$ required to rotate the eth face along the dth axis produces the vector $\boldsymbol{b}'$, where*

$$\boldsymbol{b}'_n = P_{de}(\boldsymbol{b}) = \begin{cases} b_i & \text{for } i \neq d, e \\ 1 & \text{for } i = d \\ 2 - b_d & \text{for } i = e \end{cases}. \tag{2.22}$$

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

Proof: Notice that for $\boldsymbol{c} \in \mathbb{C}_d^{D-1} \cap \mathbb{C}_e^{D-1}$, $\|\boldsymbol{b} - \boldsymbol{c}\|_2 = \|\boldsymbol{b}' - \boldsymbol{c}\|_2$. We then have that

$$\begin{aligned} g(\boldsymbol{a}, \boldsymbol{b}) &= \min_{\boldsymbol{c} \in \mathbb{C}_d^{D-1} \cap \mathbb{C}_e^{D-1}} \left\{ \|\boldsymbol{c} - \boldsymbol{a}\|_2 + \|\boldsymbol{b} - \boldsymbol{c}\|_2 \right\} \\ &= \min_{\boldsymbol{c} \in \mathbb{C}_d^{D-1} \cap \mathbb{C}_e^{D-1}} \left\{ \|\boldsymbol{c} - \boldsymbol{a}\|_2 + \|\boldsymbol{b}' - \boldsymbol{c}\|_2 \right\} \\ &= \|\boldsymbol{a} - \boldsymbol{b}'\|_2 . \end{aligned}$$

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

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 $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N \in \mathbb{S}_\infty^{D-1}$, and $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$, such that $\sum_{n=1}^N \alpha_n = 0$. Then

$$\sum_{n,m} \alpha_n \alpha_m g(\boldsymbol{a}_n, \boldsymbol{a}_m) = \sum_{n,m} \alpha_n \alpha_m \|\boldsymbol{a}_n - \boldsymbol{a}'_m\|_2 \leq 0,$$

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

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

2.3 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 40 000 iterations. We use that as a burn-in, and thereafter sample 10 000 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 $t_i = 1.3^i$, for $i \in \{0, 1, \ldots, 5\}$. 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

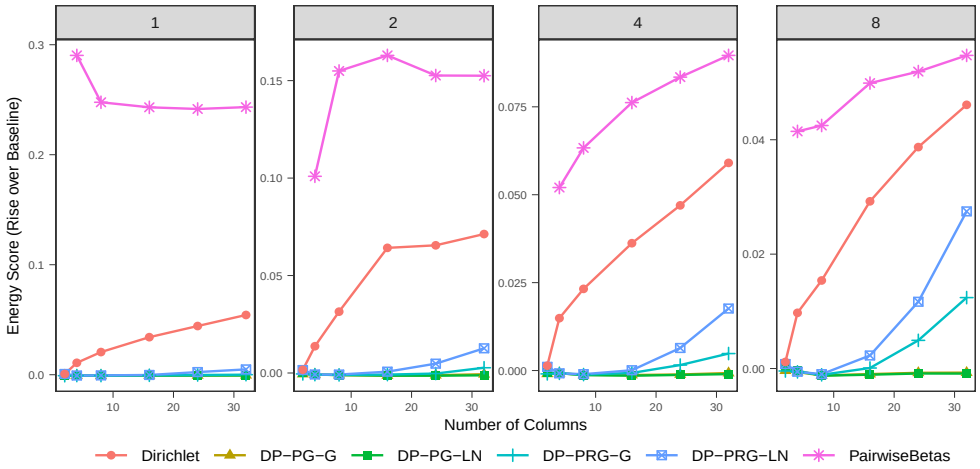


Figure 2.2: Average energy score rise over baseline (on $\mathbb{S}_{\infty}^{D-1}$) for various models fitted to simulated data, with ascending count of mixture components (indicated by plot heading) and number of dimensions (indicated by horizontal axis). Note that pairwise betas is a moment-restricted model.

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 $\xi_d \sim \mathcal{G}(1,1)$, and rate parameter $\tau_d \sim \mathcal{G}(2,2)$. For the log-normal centered models (PG-LN, PRG-LN), the centering distribution's log-mean $\boldsymbol{\mu} \sim \mathcal{N}_D(0, \boldsymbol{I}_D)$, and covariance matrix $\Sigma \sim \mathcal{IW}(D+10, (D+10)\boldsymbol{I}_D)$. These values are intended such that draws from the prior for Σ will weakly tend towards the identity matrix. For models learning rate parameters β_{jd} (PG-G, PG-LN), the centering distribution's shape parameter $\zeta_d \sim \mathcal{G}(1,1)$ and rate parameter $\sigma_d \sim \mathcal{G}(2,2)$ for $d = 2, \ldots, D$. 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. $\mu_\jmath$, $\Sigma_\jmath$ are the parameters of the mixture component distribution; π is the probability vector assigning weight mixture components; γ_n is the mixture component identifier for each simulated observation.

for s_{iter} in $[1, \ldots, 10]$ **do**

  for s_{mix} in $[1, 2, 4, 8]$ **do**

    for $\jmath$ in $1, \ldots, s_{\text{mix}}$ **do**

      Generate $\mu_\jmath \sim \mathcal{N}_{32}\left(\mathbf{0}, \boldsymbol{I}\right)$

      Generate $\Sigma_\jmath \sim \mathcal{IW}_{32}\left(70, 70\boldsymbol{I}\right)$

    end for

    Generate $\pi \sim \text{Dirichlet}(\mathbf{10}_{s_{\text{mix}}})$

    for n in $1, \ldots, 1000$ **do**

      Generate $\gamma_n \sim \text{Categorical}(\pi)$

      Generate $\boldsymbol{X}_n \sim \mathcal{LN}\left(\mu_{[\gamma_n]}, \Sigma_{[\gamma_n]}\right)$

    end for

    for D_{col} in $[2, 4, 8, 16, 24, 32]$ **do**

      Project columns 1 to D_{col} of $\mathbf{X}$ onto $\mathcal{S}_\infty^{D_{\text{col}}-1}$ and save.

    end for

  end for

end for

2.3.1 Simulation Study

The challenging problem in multivariate EVT is to capture the dependence structure of the limiting distribution. To this end, we focus our simulation study specifically on the angular component. 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 similarly works to capture a complex dependence structure on an $\mathbb{S}_p^{D-1}$ sphere, albeit with $p = 1$, and is implemented in the readily available package `BMAmevt` in R (Sabourin, 2023), which can provide samples from the posterior predictive distribution. These samples 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 Dirichlet and pairwise betas 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 2.2 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 (2.20). 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 combinatorically, with a sample space of dimension $\binom{D}{2} + 1$. By comparison, the sample space for PG-G and PRG-G are $2(J+1)d$ and $(J+1)d$ respectively, where J is the number of extant clusters, with much of that inference able to be done in parallel. In our testing, for low-dimensional problems, `BMAmevt` was substantially faster than any of our proposed DP mixture models. However, for examples with high numbers of dimensions, the computational time for `BMAmevt` was greater than that for PG-G. We compare computing times in our data analysis in Table 2.1b.

2.3.2 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

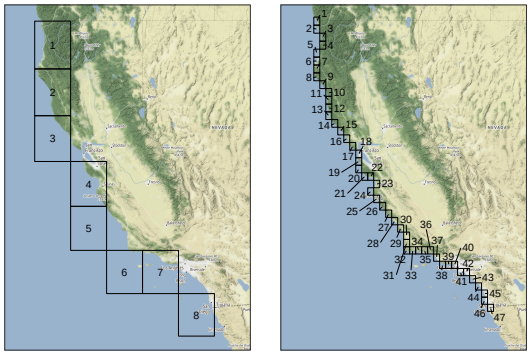


Figure 2.3: Grid cell locations for ERA-Interim (left) and ERA5 (right).

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 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 2.3 provides a visual representation of the area these grid cells cover.

Fitting our models to the IVT data requires some pre-processing. First, we subset the data to the rainy season, which in California runs roughly from November to March. Following the approach described in Section 2.1 we estimate the shape and scale parameters of a univariate

Algorithm 2 Data preprocessing to isolate and transform data exhibiting extreme behavior. r_n represents the radial component, and $\boldsymbol{v}_n$ the angular component. The declustering portion is relevant for data correlated in time.

for $d = 1, \ldots, D$ **do**

Set $b_{t,d} = \hat{F}_d^{-1}\left(1 - \frac{1}{t}\right)$.

With $\boldsymbol{x}_d > b_{t,d}$, fit σ_d, ξ_d via MLE according to generalized Pareto likelihood.

end for

for $n = 1, \ldots, N$ **do**

Define $z_{nd} = \left(1 + \xi_d \frac{x_{nd} - b_{t,d}}{\sigma_d}\right)_+^{1/\xi_d}$; then $r_n = \|\boldsymbol{z}_n\|_\infty$, $\boldsymbol{v}_n = \frac{\boldsymbol{z}_n}{\|\boldsymbol{z}_n\|_\infty}$

end for

Subset $\boldsymbol{r}, \boldsymbol{v}$ such that $r_n \geq 1$

if declustering **then**

for $n = 1, \ldots, N$ **do**

If $r_n \geq 1$ and $r_{n-1} \geq 1$, drop the lesser (and associated v_n) from data set.

end for

end if

GP, in each dimension, using maximum likelihood. We set the threshold in each dimensions d as $b_{t,d} = \hat{F}_d^{-1}(1 - t^{-1})$, where $\hat{F}$ is the empirical CDF and $t = 20$, that corresponds to the 95 percentile. We then use the transformation in Equation (2.1) 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_n such that $\|z_n\|_\infty > 1$, 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

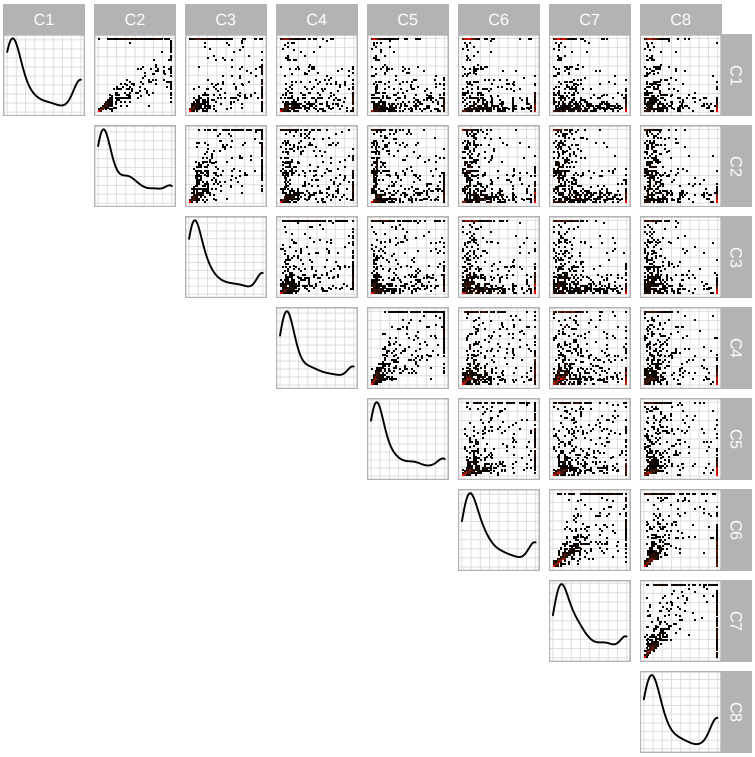


Figure 2.4: Pairwise plots from ERA-Interim data after transformation and projection to $\mathbb{S}_\infty^7$. Down the diagonal are marginal kernel densities, with two-dimensional histograms on the off-diagonal. In those plots, red indicates a higher density. All data are between 0 and 1.

observations. A pairwise plot of the transformed data after processing and declustering is presented in Figure 2.4. 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 45°, while other site combinations such as $(3, 6)$, $(2, 7)$, or $(1, 8)$ the data are 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

Table 2.1: 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	**0.7966**	0.7970
ERA5	2.0311	1.6404	1.5576	**1.4349**	1.5051

(a) Energy score criterion from fitted models against the IVT data. Lower is 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 models

PRG-LN models to both datasets.

Table 2.1a 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 2.2, the preferred model is the projected restricted gamma models, though for the lower-dimensional ERA-Interim data, all models perform comparably. Table 2.1b 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.

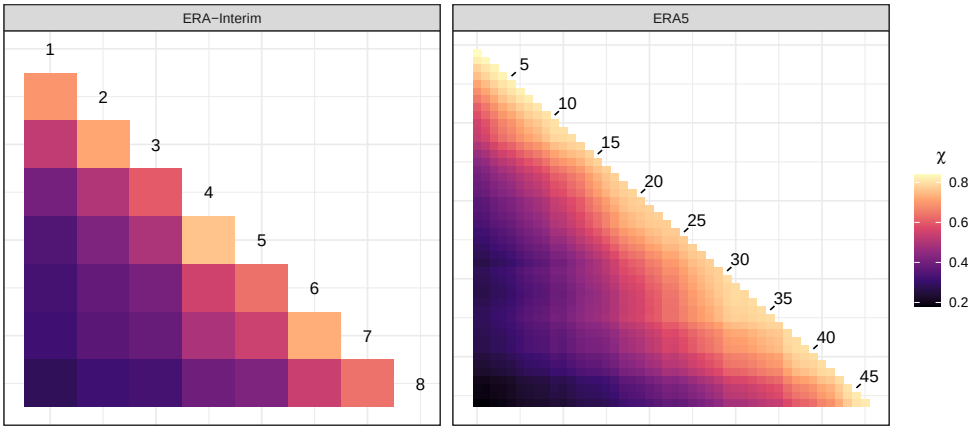


Figure 2.5: Pairwise extremal dependence coefficients for IVT data using the PRG-G model.

We consider an exploration of the pairwise extremal dependence using Monte Carlo estimates of the coefficients in Equation (2.9). For this we use samples obtained from the PRG-G model. Figure 2.5 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 dependency 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 2.6 shows, for the ERA-Interim data under the PRG-G model, the conditional survival curve defined in Equation (2.10), for one dimension, conditioned on all other dimensions being greater than their (fitted) 90th percentile. Figure 2.7 presents the bi-variate conditional

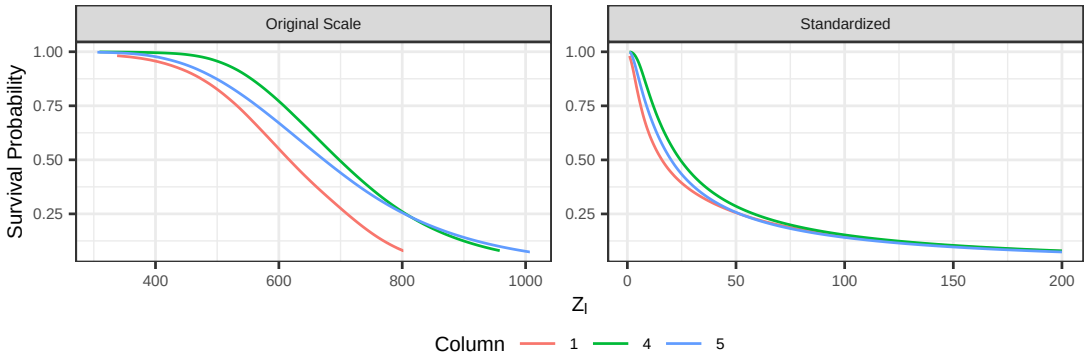


Figure 2.6: Conditional survival curves for selected locations, using ERA-Interim, and PRG-G model, conditioning on all other dimensions at greater than 90th percentile (fitted). The left panel uses original units. Right panel uses standardized units.

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 2.5 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 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 $p = 15$.

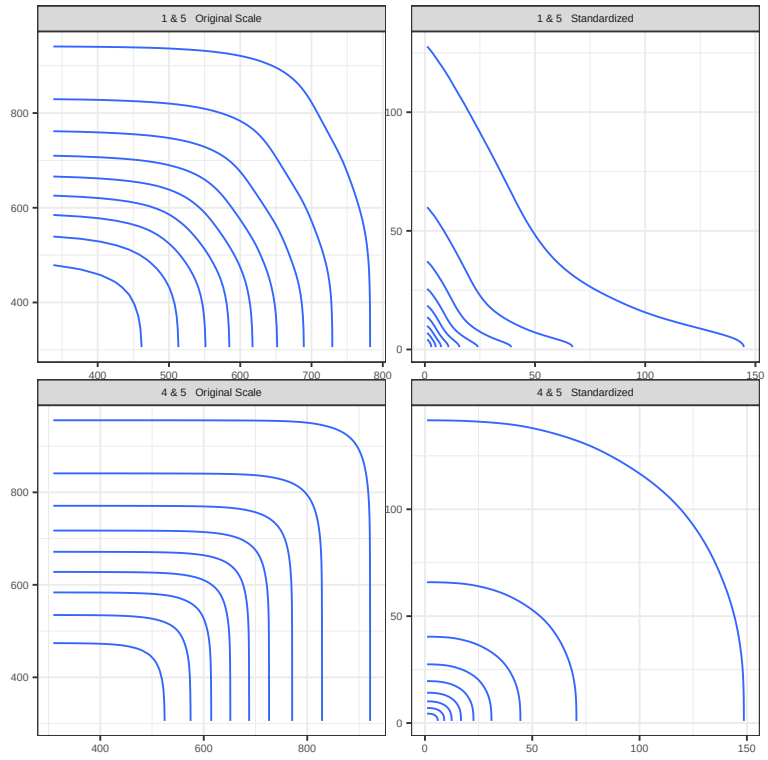


Figure 2.7: Pairwise conditional survival curves for selected locations, using ERA-Interim, and PRG-G model, conditioning on all other dimensions at greater than 90th percentile (fitted).

2.4 Conclusion

In this chapter, 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 data to $\mathbb{S}_p^{D-1}$, fits them 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 $p \to \infty$, 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 chapter 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.

Developing an 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. Recall that Equation (2.3) gives y_d as a function of $y_1, \ldots, y_{D-1}$. An analogous expression can be obtained for any y_d. This indicates that there are D equivalent transformations, each yielding a different Jacobian and, for $p > 1$, potentially resulting in a different density. Second, evaluation of geodesic distances 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 chapter 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.

Chapter 3

Anomaly Detection in Peaks-over-Threshold Settings and Angular Representations of Categorical Data

3.1 Introduction

Anomaly detection, describes a field of methods for identifying observations as *anomalous*; a term that requires defining. For this chapter following the general trend in the literature, we define anomalies as observations that are in some manner different than non-anomalous data. We interpret this to say that anomalies are data that were not produced by the same generating distribution as non-anomalous data, and as such, we would expect observations found in regions of relative data sparsity to be more likely to be anomalous than those observations found in regions of high data abundance. We characterize this assumption as *anomalies stand apart*. In

the literature as here, the term *normal data* is used to refer to data which are not anomalous. Normal data tend to cluster into homogenous groups, but anomalous data are heterogenous in their differences.

Alternative names for the field of anomaly detection include *outlier* detection, and *novelty* detection, though these terms have their own nuances. Outliers are characterized as observations that are in some manner far from normal data. In a regression context, they may have large fitted residuals, or exert large influence on model fits. Novelties in contrast are data coming from a distribution that has not been seen before. A novelty detection application will then assume a clean training data set containing no anomalies, and identify observations not belonging to the distribution as trained. Chandola et al. (2009) refer to this practice as semi-supervised anomaly detection. For our purpose, we do not assume the existence of labels in the training dataset, and seek an algorithm that can produce anomaly scores in the absence of class labels. As such, we will offer a brief overview of unsupervised anomaly detection methods, as well as discussion of the methods we are proposing here as competing models.

The complete field of anomaly detection is vast. However, most methods can be roughly grouped into three core ideas: statistical model approaches, non-statistical model approaches, and clustering methods. Common to all approaches is the assumption that anomalous data will tend to stand apart from normal data.

Statistical models for anomaly detection attempt to model the distribution of data, with the goal of estimating the data density around an observation. In specific applications, one might make assumptions about the parametric form of the generating distribution of the data, but for general application, a non-parametric density estimator is frequently used. This might include algorithms such as k-Nearest Neighbors *k-NN* (Kramer, 2013); kernel density estimation approaches such as the Parzen-Rosenblatt windowing method (Rosenblatt, 1956; Parzen, 1962); or even semi-parametric density estimation methods, such as Gaussian mixture

models (McNicholas, 2010). Local Outlier Factor (Breunig et al., 2000) is an example of an anomaly score using a non-parametric density estimator.

Clustering methods group data into clusters of similar observations. The grouping methods tend to rely on distance metrics and generally make no assumptions regarding the underlying distribution of the data. We can further sub-divide this sub-field into types of clustering methods: linkage-based, centroid-based, and density-based. These methods as applied to the field of anomaly detection assume that anomalous observations tend to stand apart from non-anomalous data.

Linkage-based clustering methods group data based on pairwise distance point-to-point, or between elements of clusters. Ackerman et al. (2010) offers a review of the topic. An illustrative example is single linkage, where the distance between two clusters is defined as the minimum distance between a point in each set. Similarly, complete linkage defines the metric to be the maximum pairwise between a point in each set. The goal of the linkage-based clustering algorithm is to maximize the total distance between clusters under whatever metric of distance is used, along with minimizing distance within clusters. An observation's anomaly score might be a function of distance to its nearest neighbor within its assigned cluster.

Centroid based clustering methods instead generate cluster centroids according to some metric. The algorithm used to find the cluster centroids depends on a chosen metric. The very popular k-Means (Hartigan and Wong, 1979) is an example of this approach. Under k-Means, cluster assignment is determined by minimizing within-cluster distance among k clusters, which simultaneously maximizes between-cluster distance. For each observation, and anomaly score may be obtained as a function of its distance to the nearest cluster centroid.

Density based clustering methods use pairwise distances between observations to establish a measure of local density, then establish local modes as clusters. *DBSCAN* (Ester et al., 1996) follows this approach, forming neighborhoods of observations and assigning labels based

on the neighborhood.

Non-statistical—or algorithmic—models beyond clustering are generally adaptations of general classification methods, applied to unsupervised learning. The Isolation Forest (Liu et al., 2008), adapted from random forests (Breiman, 2001), uses decision trees to isolate observations. Those observations that are more easily isolable are regarded as more anomalous. One-class Support Vector Machines (Chang and Lin, 2011) is a variant of the support vector machine classification system, optimized for anomaly detection. One-class SVM uses support vectors to describe a decision boundary in kernel space around *normal* behavior. A higher distance to that decision boundary on the anomalous side is regarded as more anomalous.

The intersection of extreme value theory and anomaly detection is a current topic of research. Some methods employ univariate EVT on estimated densities calculated via other means, such as Clifton et al. (2011) using a Gaussian Mixture model, and Gu et al. (2021) using a Gaussian process. Both then employ EVT on the estimated densities to establish a decision threshold theoretically, avoiding the process of determining said threshold heuristically. Beyond these applications, the applicability of extreme value theory to anomaly detection is predicated on the assumption that extreme observations are more likely to be anomalous. A discussion on this point is provided by Goix et al. (2017), stating that extreme observations exist at the border between anomalous and non-anomalous regions. Indeed, for most datasets in our testing, the probability an individual observation is anomalous is higher for data in the tails of the distribution. This relative abundance of anomalies among extremes might cause a naive classifier that does not take into account the dependence structure of extremes to classify all extremes as anomalous. If we follow the assumption that anomalies stand apart, then extreme observations that cluster into a homogenous group should not be considered anomalous. For this reason, we desire a classifier that considers the dependence structure of the extremes as well. Goix et al. (2017) offers one such example. Their method is based on transforming the data to a

standard Pareto suing the transformation $T(x) = 1/(1 - \hat{F}(x)) \in [1, \infty)$, where $\hat{F}$ corresponds to the empirical distribution function. Then the the space $[1, \infty)^d$ is partitioned into α-cones, defined as subsets where in each dimension the observations are in excess of a α. α-cones with few observations correspond to lower-density regions, so observations falling into these cones are considered more likely to be anomalous.

A central result of multivariate EVT is that, conditional on an observation being extreme, its radial component—or magnitude—is independent of its angular component. In this chapter, following Trubey and Sansó (2024), we fit a Bayesian non-parametric mixture of projected gammas to the angular component, and use samples from its posterior predictive distribution to compute an estimate of the density of the angular component. Direct estimation of density via a fitted model is difficult, owing to the bounded nature of the angular distribution. Instead, we employ non-parametric density estimators including k-nearest neighbors and kernel density estimation to produce estimates of angular density. Further, to expand the applicability of this algorithm, we produce an extension of the BNP projected gamma model to include categorical data. Standing alone, this component represents a highly flexible density model for categorical data, and it efficiently pairs with the projected gamma model for angular data. We develop several anomaly scoring metrics applicable to the angular data, categorical data, and *mixed* data regimes. The major contributions of this chapter are thus three-fold: We develop an anomaly detection algorithm for extreme data that accounts for the dependence structure between extremes, approaching density estimation in a continuous space rather than discrete binning in a partition of the space. We obtain a flexible model for multivariate categorical data that efficiently captures the dependence structure between categories in multiple variables, as well as anomaly scores in this setting. Finally, we provide a model that links the scores developed in these two cases, tackling multivariate observations with components of different types.

The chapter proceeds as follows: Section 3.2 provides a reintroduction of the angular

data model adapted for novelty detection. Section 3.3 introduces our anomaly scores for angular data, describing the density estimation methods employed, as well as how radial information is incorporated. Section 3.4 introduces our flexible categorical data model, along with anomaly scores based on it. Section 3.4.3 provides a link between the two regimes; anomaly scores that include information from both categorical and angular data. Section 3.4.4 employs the same rank transformation used in Goix et al. (2017) to apply the angular data model to data not already assumed to be in excess of a threshold, widening the applicability of our metrics. Section 3.5 provides the resulting performance of our anomaly scores as applied to seven reference anomaly detection datasets, as well as comparing to three canonical anomaly scoring methods. Finally, Section 3.6 provides concluding remarks and discussion.

3.2 The Angular Data Model

In Chapter 1, we discussed the separation of the extreme vector $\boldsymbol{Z}$ into its radial and angular components, $\boldsymbol{R}$ following a standard Pareto, and $\boldsymbol{V}$ following a distribution Φ, with support on $\mathbb{S}_\infty^{D-1}$. To obtain a flexible model for Φ we use the projected gamma density as the kernel of a random measure mixture model, based on the Pitman-Yor ($\mathcal{PY}$) process introduced in Perman et al. (1992). Pitman-Yor processes are fully atomic random measures that are specified by two parameters and a centering distribution. They can be formulated, using a stick-breaking representation (Ishwaran and James, 2001a), as

$$\Pr[\,\alpha \mid \cdots] = \sum_{\jmath=1}^{\infty} \pi_\jmath \delta_{\boldsymbol{\alpha}_\jmath}; \quad \sum_{\jmath=1}^{\infty} \pi_\jmath = 1, \quad \pi_\jmath := \rho_\jmath \prod_{k<\jmath} (1-\rho_k)$$

where $\delta_{\boldsymbol{\alpha}_\jmath}$ indicates a point mass at $\boldsymbol{\alpha}_\jmath$, and $\boldsymbol{\alpha}_\jmath$ are sampled independently from G_0. The stick-breaking proportions $\rho_\jmath \sim \text{Beta}(1-\omega, \eta+\jmath\omega)$. Observe $\omega \in [0,1)$, and $\eta > -\omega$ are referred to as the discount and the concentration parameters, respectively. Pitman-Yor processes have the advantage over the more commonly used Dirichlet processes (Ferguson, 1974) of including

a discount parameter along with the concentration parameter, allowing greater control over the formation of new clusters. A hierarchical formulation of the model for observations $\boldsymbol{y}_n \in \mathbb{S}_p^{D-1}$, $n = 1, \ldots, N$, is

$$\begin{aligned} \boldsymbol{y}_n \mid \boldsymbol{\alpha}_n &\sim \mathcal{PG}_p\left(\boldsymbol{y} \mid \boldsymbol{\alpha}_n, \mathbf{1}\right) & G_0 &= \mathcal{LN}_D\left(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Sigma\right) \\ \boldsymbol{\alpha_n} &\sim G & \mu &\sim \mathcal{N}_D\left(\mathbf{0}, \mathbf{1}\right) \\ G &\sim \mathcal{PY}\left(\omega, \eta, G_0\right) & \Sigma &\sim \mathcal{IW}_D\left(\nu, \Psi\right). \end{aligned} \tag{3.1}$$

Here $\mathcal{LN}$ denotes a log-normal, $\mathcal{N}$ a normal, and $\mathcal{IW}$ an inverse Wishart. We refer to this model as a *Pitman-Yor mixture of projected gammas* ($\mathcal{PYPG}$). As a kernel density, it was observed in Chapter 2 that the unrestricted form of the $\mathcal{PG}_p$ with both shape and rate parameters offered no improvement in model fidelity on real data compared to the restricted form, where the rate parameters are fixed at 1. For a more parsimonious model, and for compatibility with the categorical model that will be developed in Section 3.4, we choose to use the restricted form.

Mixtures of Pitman-Yor processes can be used to group observations into stochastically assigned clusters, where all observations within a cluster share a set of parameters. Cluster assignment is accomplished through data augmentation, where γ_n, the cluster identifier for observation n, is sampled according to both cluster weight and kernel density of observation n given cluster parameters. We make use of the blocked-Gibbs sampler on a truncated stick-breaking representation of the Pitman–Yor model. Cluster weights are then sampled as

$$\begin{aligned} \rho_\jmath \mid \boldsymbol{n} &\sim \text{Beta}\left(1 + C_\jmath - \omega, \eta + \sum_{k>\jmath} C_k + \jmath\omega\right) \text{ for } j = 1, \ldots, J-1 \\ \pi_\jmath &:= \begin{cases} \rho_\jmath \prod_{k<\jmath}(1-\rho_k) & \text{for } j = 1, \ldots, J-1 \\ \prod_{k=1}^{J-1}(1-\rho_k) & \text{for } j = J \end{cases} \end{aligned} \tag{3.2}$$

where $C_\jmath$ is the number of observations in cluster $\jmath$. In this form, the Dirichlet process is a

special case of the Pitman-Yor process where the discount parameter $\omega := 0$. Then γ_n is sampled

$$\Pr\left[\gamma_n = \jmath \mid \boldsymbol{\rho}, \boldsymbol{\alpha}\right] = \frac{\pi_\jmath \, \mathcal{PG}_p\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \mathbf{1}\right)}{\sum_{k=1}^{J} \pi_k \, \mathcal{PG}_p\left(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_k, \mathbf{1}\right)}. \tag{3.3}$$

Within the blocked-Gibbs algorithm, $\boldsymbol{\rho} \mid \boldsymbol{\gamma}$ are mutually independent, as are $\boldsymbol{\gamma} \mid \boldsymbol{\rho}$. This conditional independence offers an opportunity for parallelization, increasing the speed of sampling.

The approach proposed in this section produces a sample of the angular measure of the distribution of the tails of the sample. The method has a number of advantages for anomaly detection: it focuses on the tails, which is where we are more likely to find anomalous behavior; it accounts for asymptotic dependence between the different components of the observation vector; it reduces the computational burden, by thinning the sample using thresholding; and it decouples the radial component to the angular component, thanks to independence.

3.3 Novelty Detection Methods

As previously stated, a novelty detection algorithm produces an anomaly score which provides a ranked ordering of observations in their likelihood of being anomalous, with higher scores indicating more likely anomalous. Building on the notion that anomalies occur in areas of low density, a general Bayesian anomaly score for observation x_n, can be defined as

$$\mathcal{S}_n = \left[\int_\Theta f(x_n \mid \theta) dG(\theta \mid \mathcal{D})\right]^{-1}$$

where $\mathcal{D}$ is the observed data and θ the distributional parameters. That is, the reciprocal of the posterior predictive density at observation x_n.

Given the independence between the angular and radial components of an extreme observation, we can consider sub-scores for the radial and angular components independently. That is,

$$\mathcal{S}_n = \mathcal{S}_{n,r} \times \mathcal{S}_{n,\boldsymbol{v}} = f_r(r_n)^{-1} \times \left[\int_\Omega f_v(\boldsymbol{v}_n \mid \boldsymbol{\alpha}) \, \mathrm{d}G(\boldsymbol{\alpha} \mid \mathcal{D})\right]^{-1}. \tag{3.4}$$

By construction r_n follows a standard Pareto distribution, so its density is $f_r(r_n) = r_n^{-2}$. As previously discussed in Section 3.2, the kernel $\mathcal{PG}_\infty$, needed for density estimation on the surface of $\mathbb{S}_\infty^{D-1}$ is not available in analytic form, thus, we resort to transforming the data to $\mathbb{S}_p^{D-1}$ for a large but finite p. This makes estimation of distributional parameters possible, but in the context of anomaly detection, a score based on $\mathcal{PG}_p$, for any p, is problematic. In fact, the transformation from $\mathbb{R}_+^D$ to $\mathbb{S}_p^{D-1}$ is not unique, as we can take any of the components of the original vector as a reference. This implies that under uniform $\boldsymbol{\alpha}$, the density can be changed by permuting the order of components. This is not appropriate for anomaly detection, because a relative ordering of density between observations is specifically what we're trying to calculate. In addition we have observed instabilities in the evaluation of (2.6) for small arguments, when the shape parameter is small. On the other hand, we notice that T_∞ is unique, as the reference is the largest value of the array. Thus, we fit the mixture model in $\mathbb{S}_p^{D-1}$, generate posterior predictive samples, and transform those samples to $\mathbb{S}_\infty^{D-1}$.

To avoid the problems of angular density evaluation in $\mathbb{S}_\infty^{D-1}$ we use a non-parametric angular density estimator based on a sample from the posterior predictive distribution of the model described in Section 3.2. Here, we consider two well-established methods: k–nearest neighbors, or *kNN* (Mack and Rosenblatt, 1979), and kernel density estimation, or *KDE* (Parzen, 1962). For both of these methods we make use of pairwise distances between observations from the dataset, and replicates from a posterior predictive sample.

As described in Chapter 2.1, geodesic distance on $\mathbb{S}_\infty^{D-1}$ is expensive to evaluate, as the computational burden grows combinatorically with the number of dimensions. As an alternative, we proposed a kernel metric, described in Equation 2.21 that serves as an upper bound to of geodesic distance that is computationally cheap to evaluate, bearing a cost equivalent to that of a Euclidean norm.

3.3.1 Nearest Neighbor Density estimation

We use this kernel metric to obtain a local posterior predictive density based on a kNN estimator on $\mathbb{S}_\infty^{D-1}$. To this end we consider a locally uniform density within a $(D-1)$-dimensional ball $\mathbb{B}$, centered on observation $\boldsymbol{v}_n$. The radius $R_k(\boldsymbol{v}_n) \coloneqq g\left(\boldsymbol{v}_n, \boldsymbol{v}^*_{N_k(n)}\right)$, where $g(\cdot,\cdot)$ is the upper bound on geodesic distance on $\mathbb{S}_\infty^{D-1}$ defined in Equation (2.21), and $\boldsymbol{v}^*_{N_k(n)}$ is the kth nearest neighbor of $\boldsymbol{v}_n$ in a sample from the posterior predictive distribution. The volume of the ball is calculated as

$$\mathrm{Vol}(\mathbb{B}_k^{D-1}) = \frac{\pi^{\frac{D-1}{2}} R_k(\boldsymbol{v}_n)^{D-1}}{\Gamma\left(\frac{D-1}{2}+1\right)}. \tag{3.5}$$

The density is thus estimated as $f_{\boldsymbol{v}}^{(k\mathrm{NN})}(\boldsymbol{v}_n \mid \boldsymbol{V}) \approx \frac{k}{N}\left(\mathrm{Vol}(\mathbb{B}_k^{D-1})\right)^{-1}$ where N is the total number of replicates of from the posterior predictive distribution. Taking the reciprocal of the estimated angular density, the angular score under the kNN estimator is then

$$\mathcal{S}_{n,\boldsymbol{v}}^{k\mathrm{NN}} = \frac{N\pi^{\frac{D-1}{2}} R_k(\boldsymbol{v}_n)^{D-1}}{k\Gamma\left(\frac{D-1}{2}-1\right)} \tag{3.6}$$

In our experience, using a large posterior predictive sample, the resulting ordering of scores was relatively robust to a choice of k between 2 and 10. We used $k = 5$ in our performance analysis.

3.3.2 Kernel Density Estimation

Kernel density estimation is an approach that makes use of kernel smoothing to produce a semi-parametric estimate of the density function for a dataset. For a scalar bandwidth parameter h,

$$f_n(x) \;=\; \int_\Omega \frac{1}{h}\mathcal{Q}\left(\frac{\boldsymbol{x}-\boldsymbol{t}}{h}\right)\mathrm{d}F_n(\boldsymbol{t}) \;\approx\; \frac{1}{Kh}\sum_{k=1}^{K}\mathcal{Q}\left(\frac{\boldsymbol{x}-\boldsymbol{x}_k^*}{h}\right)$$

where $\boldsymbol{x}_k^*$ are random replicates from F. The choice of kernel function $\mathcal{Q}$, and selection of the bandwidth parameter h are both topics that have been extensively researched. In practice the Gaussian kernel seems to be well regarded for its simplicity, flexibility, and interpretability.

The bandwidth parameter in this case corresponds to the standard deviation of the kernel function. The multivariate Gaussian kernel is more flexible, accepting a matrix as the bandwidth parameter. A larger bandwidth serves to smooth the resulting density estimate, where a lower bandwidth is more responsive to individual observations of data. Optimization of h is application and data specific, but there do exist various *rules of thumb* based on summary statistics of the data. For our analysis, we are making use of a distance analogue on $\mathbb{S}_{\infty}^{D-1}$ described in Equation (2.22), which precludes the ability to describe bandwidth using a matrix. We therefore consider the univariate case of f in kernel space, where $\|x-x^*\|$ has been replaced with $g(\boldsymbol{v}, \boldsymbol{v}^*)$.

For selection of the bandwidth parameter h, we employ Silverman's rule of thumb (Silverman, 2018), estimating $\hat{h} = \left(\frac{4}{D+2}\right)^{\frac{1}{D+4}} N^{-\frac{1}{D+4}} \hat{\sigma}$. This then requires the estimation of $\hat{\sigma}$, which in this case we calculate from pairwise distances. Recall that for a random variable X, $\mathrm{E}\left[\|X_j - X_k\|_2\right] = 2\mathrm{Var}(X)$. In that case,

$$\hat{\sigma} = \sqrt{\frac{1}{2N(N-1)} \sum_{j \neq k} g(\boldsymbol{v}_j^*, \boldsymbol{v}_k^*)},$$

where $\boldsymbol{v}_j^*, \boldsymbol{v}_k^*$ are replicates from the posterior predictive distribution. Then $\hat{\sigma}$ is used in the aforementioned rule of thumb for h. Finally, the angular score under KDE is then calculated as

$$\mathcal{S}_{n,\boldsymbol{v}}^{\mathrm{kde}} = \mathrm{E}_{\boldsymbol{v}^*}\left[\exp\left\{-\left(\frac{g(\boldsymbol{v}_n, \boldsymbol{v}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K}\sum_{k=1}^{K} \exp\left\{-\left(\frac{g(\boldsymbol{v}_n, \boldsymbol{v}_k^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{3.7}$$

where $\boldsymbol{v}_k^*$ are again replicates from the posterior predictive distribution. We investigated other methods of calculating bandwidth, as well as searched the neighborhood around our bandwidth estimate for example datasets. The estimator following Silverman's rule of thumb as described consistently produced the most performant rank ordering of angular anomaly scores on tested datasets.

Algorithm 3 Workflow for anomaly detection on $\mathbb{S}_\infty^{D-1}$.

1: Take r_n, $\boldsymbol{y}_n$ according to Algorithm (2), substituting $b_{q,d} = \hat{F}_d^{-1}(q)$

2: Fit $\mathcal{PYPG}(\boldsymbol{y})$ from Equation (3.1)

3: From $\boldsymbol{\alpha} \mid \boldsymbol{y}$, sample $\boldsymbol{\varrho}_k^* \mid \boldsymbol{\alpha} \sim \prod_d \mathcal{G}(\alpha_d)$ for $k = 1, \ldots, K$

4: Take $\boldsymbol{v}_k^* = T_\infty(\boldsymbol{\varrho}_k^*)$

5: Take $\mathcal{S}_{n,v}$ as per Equations (3.6,3.7)

3.4 Binary and Categorical Data

In the previous sections we have used extreme value theory to obtain samples from the tail distribution of a given sample of observations. Unfortunately those results can only be applied to continuous random variables. Many applications of novelty detection include both real and categorical data, so here we consider an extension of the projected gamma mixture model to handle categorical observations.

Suppose $\boldsymbol{X}$ is a vector of M random categorical variables. Let $\boldsymbol{C}_m$ be a categorical random variable, encoded in one-hot, or multinomial, encoding. The length of $\boldsymbol{C}_m$, D_m, indicates the number of categories. Then $\boldsymbol{C}$ is the concatenation of M one-hot encoded categorical RV's. It is a binary vector of length $D = \sum_{m=1}^{M} D_m$, and $\sum_{m=1}^{M} \sum_{d=1}^{D_m} C_{md} = M$. To account for over-dispersion, and compatibility with our methods in Chapter 2, we consider a Dirichlet-multinomial density for $\boldsymbol{C}_m$. Recall that the Dirichlet distribution is a special case of the projected gamma, projected onto $\mathbb{S}_1^{D-1}$, with rate parameters uniformly fixed as $\beta_d = \beta = 1$ by convention. We consider a Dirichlet-multinomial densiy, $\mathcal{DM}(\cdot)$, that is obtained by integrating out the latent categorical probability vector from a multinomail density with a Dirichlet prior. That is,

$$\mathcal{DM}(\boldsymbol{c} \mid \boldsymbol{\alpha}) = \int_\pi \mathcal{M}(\boldsymbol{c} \mid \pi)\ \mathcal{D}(\pi \mid \boldsymbol{\alpha}) \mathrm{d}\pi.$$

Recalling that a categorical random variable can be considered as a multinomial with size 1, we

can further simplify the Dirichlet-multinomial to a Dirichlet-categorical, reducing the computational burden. Thus,

$$\boldsymbol{c} \mid \boldsymbol{\alpha} \sim \mathcal{DC}(\boldsymbol{c} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)}{\Gamma(1 + \sum_{d=1}^{D} \alpha_d)} \prod_{d=1}^{D} \frac{\Gamma(c_d + \alpha_d)}{\Gamma(\alpha_d)} = \frac{\prod_{d=1}^{D} \alpha_d^{c_d}}{\sum_{d=1}^{D} \alpha_d} \tag{3.8}$$

We then consider a *concatenated* Dirichlet-categorical ($\mathcal{CDC}$) as a product of Dirichlet-categorical densities. That is, $\mathcal{CDC}(\boldsymbol{c} \mid \boldsymbol{\alpha}) = \prod_{m=1}^{M} \mathcal{DC}(\boldsymbol{c}_m \mid \boldsymbol{\alpha}_m)$. Then we can define a Bayesian nonparametric categorical data model as:

$$\begin{aligned} \boldsymbol{c}_n \mid \boldsymbol{\alpha}_n &\sim \mathcal{CDC}\left(\boldsymbol{c}_n \mid \boldsymbol{\alpha}_n\right) & G_0 &= \mathcal{LN}\left(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Sigma\right) \\ \boldsymbol{\alpha_n} &\sim G & \mu &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{1}\right) \\ G &\sim \mathcal{PY}\left(d, \eta, G_0\right) & \Sigma &\sim \mathcal{IW}\left(\nu, \Psi\right). \end{aligned} \tag{3.9}$$

Note that there exists a strong negative covariance between categories within a categorical variable. To account for this in our proposed prior, the parameter Ψ is chosen as a block diagonal matrix, with each m block corresponding to the mth categorical variable. Setting the value of the diagonal to ψ_0, the off-diagonals within the m block are set to $-\psi_0 D_m^{-2}$ where D_m is the number of categories in the mth categorical variable. This value corresponds to the covariance of a categorical variable where all category probabilities are equal. In addition to the proposed log-normal model, we investigated using a product of gammas as the centering distribution in Equation (3.9), but we observed that this choice induces numerical instability. We observed that the log-normal distribution, with its squared exponential tails and ability to account for negative covariance within the prior, provided stable model fitting.

3.4.1 Anomaly Detection Methods for Categorical Data

Anomaly scores analogous to the ones proposed in Section 3.3 can be obtained for categorical variables by transforming the latent variables that define a Dirichlet-Multinomial

distribution on $\mathcal{S}_1^{D-1}$ to $\mathcal{S}_\infty^{D-1}$. We start by considering the cluster identifiers. Extrapolating Equation (3.3) to the categorical model, cluster identifiers γ_n are sampled with probabilities

$$\Pr[\gamma_n = \jmath \mid \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{c}_n] = \frac{\pi_\jmath \, \mathcal{CDC}\left(\boldsymbol{c}_n \mid \boldsymbol{\alpha}_\jmath\right)}{\sum_{k=1}^{J} \pi_k \, \mathcal{CDC}\left(\boldsymbol{c}_n \mid \boldsymbol{\alpha}_k\right)} \quad \text{for } \jmath = 1, \ldots, J, \tag{3.10}$$

where $\boldsymbol{\pi}$ refers to the cluster weights under the stick-breaking representation of the Pitman-Yor process model. For a given sample from the posterior for $\boldsymbol{\alpha}$, first we sample γ_n, then sample

$$\boldsymbol{\varrho}_n \mid \boldsymbol{\alpha}_{\gamma_n} \sim \prod_{d=1}^{D} \mathcal{G}\left(\varrho_{nd} \mid \alpha_{\gamma_n d}, 1\right). \tag{3.11}$$

These are the latent variables that provide the core structure to the categorical data model. In fact, the component probability vectors for the concatenated multinomial are obtained by projecting $\boldsymbol{\varrho}_n$ onto $\prod_{m=1}^{M} \mathbb{S}_1^{D_m - 1}$ to produce $\boldsymbol{\pi}_n = \prod_{m=1}^{M} T_1(\boldsymbol{\varrho}_{nm})$. Anomaly scores analogous to the ones proposed in the continuous case can then be obtained by letting $\boldsymbol{\nu}_n = T_\infty(\boldsymbol{\varrho}_n)$, the transformation of $\boldsymbol{\varrho}_n$ onto $\mathbb{S}_\infty^{D-1}$. It is important to notice that distance metrics between projections of $\boldsymbol{\varrho}_n$ and replicates of $\boldsymbol{\varrho}^*$ from the posterior predictive distribution is straightforward. This provides a distinct advantage to the approach based on the distance between the directly observed values $\boldsymbol{c}_n$ and samples of $\boldsymbol{C}$, obtained from the corresponding posterior predictive distribution (Alamuri et al., 2014).

We develop four methods based on applications of the KNN and KDE metrics previously described. Making an abuse of notation for simplicity of presentation, let

$$\tilde{\mathrm{E}}\left[\boldsymbol{\nu}_n\right] := T_\infty(\mathrm{E}\left[\boldsymbol{\nu}_n \mid \boldsymbol{c}_n\right]),$$

the projection of the expectation of $\boldsymbol{\nu}_n$ back onto $\mathbb{S}_\infty^{D-1}$. Evaluating this expectation by Monte Carlo approximation is equivalent calculating the spherical mean (Mardia et al., 1999), which takes the arithmetic mean of observations in Cartesian coordinates, then projects back onto the sphere.

The hypercube KNN *(hknn)* metric applied to the latent projected $\mathbb{S}_{\infty}^{D-1}$ space uses the negative definite kernel metric previously established to estimate distance between $\tilde{\mathrm{E}}[\boldsymbol{\nu}_n]$ and $\boldsymbol{\nu}^*$. This score takes the form:

$$\mathcal{S}_{n,\boldsymbol{\nu}}^{\mathrm{hknn}} = \frac{N\,\pi^{\frac{D-1}{2}}}{k\,\Gamma\left(\frac{D-1}{2}+1\right)}\; R_k\left(\tilde{\mathrm{E}}[\boldsymbol{\nu}_n]\right)^{D-1} \tag{3.12}$$

where $R_k\left(\tilde{\mathrm{E}}[\boldsymbol{\nu}_n]\right)$ measures the distance from $\tilde{\mathrm{E}}[\boldsymbol{\nu}_n]$ to the kth nearest replicate from a sample from the posterior predictive distribution for $\boldsymbol{\nu}^*$. This projection places all the class probabilities within the same sphere and subject to the same distance measure. Note here we are first taking the expectation of $\boldsymbol{\nu}_n$, then the expectation of the kernel metric raised to the $D-1$ power.

The *hkde* score applied to the categorical space operates in much the same way. We compute $\tilde{\mathrm{E}}[\boldsymbol{\nu}_n]$, and employ the same kernel metric to compute distance from a sample from the posterior predictive distribution. From there, however, we use kernel density estimation to compute local density for observation n. The score is then

$$\mathcal{S}_{n,\boldsymbol{\nu}}^{\mathrm{hkde}} = \mathrm{E}_{\boldsymbol{\nu}^*}\left[\exp\left\{-\frac{1}{2}\left(\frac{g(\tilde{\mathrm{E}}[\boldsymbol{\nu}_n],\boldsymbol{\nu}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\exp\left\{-\frac{1}{2}\left(\frac{g(\tilde{\mathrm{E}}[\boldsymbol{\nu}_n],\boldsymbol{\nu}_k^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{3.13}$$

We use the same previously described approach to choose h. An exploration of manually tuning h did not consistently outperform the rule of thumb estimator.

Notice that the *hkde* score depends on two expectations that are computed in sequence. A variant of the score is obtained by computing the expectations jointly:

$$\mathcal{S}_{n,\boldsymbol{\nu}}^{\mathrm{lhkde}} = \mathrm{E}_{\boldsymbol{\nu}^*,\boldsymbol{\nu}_n}\left[\exp\left\{-\frac{1}{2}\left(\frac{g(\boldsymbol{\nu}_n,\boldsymbol{\nu}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K_{\boldsymbol{\nu}_n}K_{\boldsymbol{\nu}^*}}\sum_{j=1}^{K_{\boldsymbol{\nu}_n}}\sum_{k=1}^{K_{\boldsymbol{\nu}^*}}\exp\left\{-\frac{1}{2}\left(\frac{g(\boldsymbol{\nu}_{n,j},\boldsymbol{\nu}_k^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{3.14}$$

Computing this for a given sample is more expensive than *hkde* due to the double sum. However, plugging in an estimate of $\mathrm{E}\left[\boldsymbol{v}_{\boldsymbol{c}_n}\right]$ removes a significant degree of uncertainty around the distribution of $\boldsymbol{v}_{\boldsymbol{c}_n}$, which may be relevant.

If, instead of projecting the unnormalized probability vectors onto a unified hypersphere $\mathbb{S}_{\infty}^{D-1}$, we normalize each m-component onto its associated simplex, $\mathbb{S}_{1}^{D_m-1}$. Using Manhattan distance on the simplex, we obtain the latent simplex KDE *(lskde)*.

$$\begin{aligned}\mathcal{S}_{n,\boldsymbol{\pi}}^{\text{lskde}} &= \mathrm{E}_{\boldsymbol{\pi}_i,\boldsymbol{\pi}^*}\left[\exp\left\{-\frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_n-\boldsymbol{\pi}^*\|_1}{\hat{h}}\right)^2\right\}\right] \\ &\approx \left[\frac{1}{K_{\boldsymbol{\pi}^*}K_{\boldsymbol{\pi}_n}}\sum_{j=1}^{K_{\boldsymbol{\pi}_n}}\sum_{k=1}^{K_{\boldsymbol{\pi}^*}}\exp\left\{-\frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_{nj}-\boldsymbol{\pi}_k^*\|}{\hat{h}}\right)^2\right\}\right]^{-1}\end{aligned} \tag{3.15}$$

Using the normalized latent class probabilities offers the advantage of numerical stability: diverging estimates of ϱ are isolated to the relevant m-component.

Algorithm 4 Workflow for anomaly detection for categorical data

1: Take $\boldsymbol{c}$ as the concatenation of m multinomial-encoded categorical variables.

2: Take $D := \sum_{m=1}^{M} D_m$ as the dimensionality of the process

3: Fit $\mathcal{PYCDC}(\boldsymbol{c})$ as per Equation (3.9)

4: From $\boldsymbol{\alpha} \mid \boldsymbol{c}$, sample $\boldsymbol{\varrho}_k^* \mid \boldsymbol{\alpha} \sim \prod_d \mathcal{G}(\alpha_d)$ for $k = 1, \ldots, K_\nu$; then $\boldsymbol{\nu}^* = T_\infty(\boldsymbol{\varrho}^*)$

5: From $\boldsymbol{\alpha}_n \mid \boldsymbol{c}_n$ sampled as per Equations (3.10-3.11) sample $\boldsymbol{\varrho}_{nk} \mid \boldsymbol{\alpha}_n \sim \prod_d \mathcal{G}(\alpha_d)$ for $k = 1, \ldots, K_{\nu_n}$

6: Take $\boldsymbol{v}_{nk} = T_\infty(\boldsymbol{\varrho}_{nk})$; $\boldsymbol{\pi}_{nk} = \prod_{m=1}^{M} T_1(\boldsymbol{\varrho}_{nkm})$

7: Take $\mathcal{S}_{n\boldsymbol{v}}$ as per Equations (3.12–3.15)

3.4.2 Mixed Models

To obtain a joint model for the density of a vector with mixed components we consider a product kernel, then mix over the parameters that define both kernels in order to capture the dependence between components. Thus,

$$(\boldsymbol{y}, \boldsymbol{c}) \sim \int_{\boldsymbol{\alpha}} \mathcal{PG}_p(\boldsymbol{y} \mid \boldsymbol{\alpha}_{\boldsymbol{y}}, \mathbf{1})\ \mathcal{CDM}(\boldsymbol{c} \mid \boldsymbol{\alpha}_{\boldsymbol{c}})\ dG(\boldsymbol{\alpha}) \tag{3.16}$$

with the distribution of $\boldsymbol{\alpha} = (\boldsymbol{\alpha_y}, \boldsymbol{\alpha_c})$ as defined in Equation 3.9. The dimensions are, respectively, $D_{\boldsymbol{y}}$ and $D_{\boldsymbol{c}}$. Note that for the projected gamma distribution, we restrict the rate parameters to $\beta_d := 1$. Also note that for the mixed model, the hyperparameter for the covariance matrix $\Sigma_{\boldsymbol{\alpha}}$ is taken as a blocked diagonal matrix, with the block corresponding to the angular component being a diagonal matrix.

3.4.3 Mixed Model Anomaly Scores

Let $D = D_{\boldsymbol{y}} + D_{\boldsymbol{c}}$ be the total number of dimensions. Then, for the mixed model, let $\boldsymbol{\nu}_n = T_\infty(R_n \boldsymbol{y}_n, \boldsymbol{\varrho}_{n\boldsymbol{c}})$, and $\boldsymbol{\nu} = T_\infty(\boldsymbol{\varrho})$. The *hknn* score can be adapted to the mixed model by re-projecting the angular data and the latent categorical component into the same sphere. This requires moving $\boldsymbol{y}_n$ back to $\mathbb{R}_+^{D_{\boldsymbol{y}}}$, by multiplying by the radial component R_n generated according to Equation 2.12, replacing $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha_y}$, the portion of the $\boldsymbol{\alpha}$ vector associated with the angular component. Then $\boldsymbol{\nu}_n = T_\infty(R_n \boldsymbol{y}_n, \boldsymbol{\varrho}_{n\boldsymbol{c}})$ is the latent projection of both the real component and categorical component into the same sphere. Also, let $\boldsymbol{\nu} = T_\infty(\boldsymbol{\varrho})$ be the generic $\boldsymbol{\nu}$ not specifically dependent on observation n. To obtain the corresponding anomaly scores we can proceed by using the expression in Equations (3.12)–(3.14)

All three scores seek a unifying approach for all data, projecting onto a the same sphere, and calculating a consistent distance metric. An alternative is to, instead, evaluate distances between angular data their own space, and, separately, latent posterior class probabilities in their own space, with the appropriate distance metric for each. In effect, this approach combines *hkde* from the angular component and *lskde* from the categorical component yielding:

$$\begin{aligned} \mathcal{S}_{n,\boldsymbol{v}}^{lmkde} &= \mathrm{E}_{\boldsymbol{v}^*,\boldsymbol{\pi}^*,\boldsymbol{\pi}_n}\left[\exp\left\{-\frac{1}{2}\left(\frac{g\left(\boldsymbol{v}_n, \boldsymbol{v}^*\right)}{\hat{h}_{\boldsymbol{v}^*}}\right)^2 - \frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_n - \boldsymbol{\pi}^*\|_1}{\hat{h}_{\boldsymbol{\pi}^*}}\right)^2\right\}\right]^{-1} \\ &\approx \left[\frac{1}{K_{\boldsymbol{\pi}^*} K_{\boldsymbol{\pi}_n}} \sum_{j=1}^{K_{\boldsymbol{\pi}_n}} \sum_{k=1}^{K_{\boldsymbol{\pi}^*}} \exp\left\{-\frac{1}{2}\left(\frac{g(\boldsymbol{v}_n, \boldsymbol{v}_k^*)}{\hat{h}_{\boldsymbol{v}^*}}\right)^2 - \frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_{nj} - \boldsymbol{\pi}_k^*\|_1}{\hat{h}_{\boldsymbol{\pi}^*}}\right)^2\right\}\right]^{-1} \end{aligned} \tag{3.17}$$

This choice to evaluate each component within its own space presents some loss of information

as to the dependence structure between $\boldsymbol{y}$ and $\boldsymbol{c}$ within the score. We will explore to what extent that loss of information is relevant.

Algorithm 5 Workflow for anomaly detection for *mixed* data

1: Take r_n, $\boldsymbol{y}_n$ according to Algorithm (2), substituting $b_{q,d} = \hat{F}_d^{-1}(q)$; $\boldsymbol{c}_n$ as in Algorithm 4.

2: Fit $(\boldsymbol{y}, \boldsymbol{c})$ using mixed model from Equation (3.16)

3: From $\boldsymbol{\alpha} \mid \boldsymbol{y}, \boldsymbol{c}$, sample $\boldsymbol{\varrho}_k^* \mid \boldsymbol{\alpha} \sim \prod_d \mathcal{G}(\alpha_d)$ for $k = 1, \ldots, K$

4: **if** $\mathcal{S}_{n,\boldsymbol{v}}$ is *hknn*, *hkde*, or *lhkde* **then**

5: From $\boldsymbol{\alpha} \mid \boldsymbol{y}_n, \boldsymbol{w}_n$: sample R_n according to Equation (2.12) substituting $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}_{\boldsymbol{y}}$, $\boldsymbol{\varrho}_{c,n}$ similar to Algorithm 4.

6: Take $\boldsymbol{\nu}_n = T_{infty}(R_n \boldsymbol{y}_n, \boldsymbol{\varrho}_{c,n})$; $\boldsymbol{\nu}^* = T_\infty(\boldsymbol{\varrho}^*)$.

7: Apply Score function.

8: **else if** $\mathcal{S}_{n,\boldsymbol{v}}$ is *lmkde* **then**

9: From $\boldsymbol{\alpha} \mid \boldsymbol{y}_n, \boldsymbol{c}_n$, sample $\boldsymbol{\varrho}_{c,n}$ similar to Algorithm 4.

10: Take $\boldsymbol{\pi}_n = \prod_{m=1}^M T_1(\boldsymbol{\varrho}_{m,c})$; $\boldsymbol{\pi}^* = \prod_{m=1}^M T_1(\boldsymbol{\varrho}^*)$

11: Apply Score function.

12: **end if**

3.4.4 Relaxing the assumption of independence

A valid critique of the model presented thus far is that in order to justify modeling the radial component of $\boldsymbol{Z}$ as independent to its angular component—the fundamental result of the multivariate extreme value theory presented—it is necessary to subset data to those observations $\boldsymbol{X}$ which exceeded a large threshold in at least one dimension. For some applications, this represents a very powerful data reduction with little loss of information pertaining to anomalies, as anomalies tend to be in the tails (see, for example, Table 3.1). For other applications,

this data reduction represents a significant loss of information about possible anomalies not corresponding to the tails. For this second group, one available avenue is to relax the assumption of independence between the angular and radial components.

Let $z_{nd} = 1/(1 - \hat{F}(x_{nd}))$ be the *rank-transformation* to the standard Pareto scale. The lower range of this transformation is bounded at 1. For data transformed in this manner, let $r_n = \|\boldsymbol{z}_n\|_\infty$ be the radial component, $\boldsymbol{v}_n = \boldsymbol{z}_n / r_n$ the angular component of $\boldsymbol{z}_n$, and $\boldsymbol{y}_n$ its projection onto $\mathbb{S}_p^{D-1}$. As no thresholding is performed we can no longer make the assumption that angles are independent of radius. Instead, we can include the radius within a joint model. As the radius is on the range $[1, \infty)$, we use the Pareto density, with shape parameter α_r as our choice of kernel.

$$(\boldsymbol{y}_n, \boldsymbol{c}_n, r_n) \sim \int_{\boldsymbol{\alpha}} \mathcal{PG}_p(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_{\boldsymbol{y}}, \mathbf{1})\, \mathcal{CDM}(\boldsymbol{c}_n \mid \boldsymbol{\alpha}_{\boldsymbol{c}})\, \mathcal{P}(r_n \mid \alpha_r)\, \mathrm{d}G(\boldsymbol{\alpha}) \tag{3.18}$$

As $\alpha_r > 0$, we augment the kernel parameters to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\boldsymbol{y}}, \boldsymbol{\alpha}_{\boldsymbol{c}}, \alpha_r)$, and use a joint log-normal as the center of the random measure prior for G. The scores developed previously in Section 3.4.3 remain applicable.

3.5 Results

As mentioned in 3.3, our goal is to produce novelty scores to rank observations according to how likely they are of being anomalous. This creates another problem: threshold selection—anomaly scores beyond what level are determined anomalous? We mentioned Clifton et al. (2011) and Gu et al. (2021) as examples of computing thresholds theoretically, but in general, thresholds are determined heuristically, using performance criteria. In some applications, heuristic determination can be extremely costly.

One such criteria is the receiver operating characteristics, or *ROC*, curve. For a given score threshold, one can compute the true positive rate, or TPR, as the number of anomalous

observations with scores above the threshold, divided by the total number of anomalous observations. The false positive rate, or FPR, is similarly the number of non-anomalous observations above the threshold, divided by the total number of non-anomalous observations. The ROC curve is formed as the TPR is plotted on the vertical axis against the FPR on the horizontal axis for a range of possible thresholds. The curve is non-decreasing, starting at the origin $(0, 0)$, and ending at unity $(1, 1)$. Threshold selection might include specifying an acceptable FPR, and determining the threshold that produces that FPR.

In assessing model performance, we sideline the issue of threshold selection by observing the whole ROC curve. Specifically, we look for the area under the ROC curve, *(AuROC)*. The better a classifier is, the closer its ROC curve will approach the upper left corner, and the closer its AuROC will approach 1.

In developing our model, we employ the blocked Gibbs sampler for stick-breaking priors detailed in Ishwaran and James (2001b). We set a discount factor of 0.1, and a concentration parameter of 1.0. In our testing, in the neighborhood around these values we found the resultant number of extant clusters to be relatively stable. We use $(\mu_0 = \mathbf{0}_D, \Sigma_\mu = \boldsymbol{I}_D)$ as prior parameters for μ, and $(\nu = D{+}50, \Psi = \nu \boldsymbol{I}_d)$ as prior parameters for Σ, except for the categorical components of the shape vector as described in Section 3.4. Deviations in μ_0 towards the negative direction bias the model towards asymptotic independence, which in our testing resulted in lower model fidelity. To update the cluster shape vectors, we employ a joint proposal step in log-space using a multivariate normal proposal, where the proposal covariance is informed with an adaptive Metropolis algorithm.(Haario et al., 2001). To hasten updates to the shape parameters, and speed convergence of the model, we employ a parallel tempering algorithm where parallel MCMC chains are sampled at an ascending temperature ladder, where density is exponentiated to the reciprocal of the chain temperature t: $f_t(\theta) = f(\theta)^{1/t}$. Chains with higher temperatures have flatter posteriors, and thus more readily move around the parameter space. Chain states are

Table 3.1: Characteristics of datasets used in the analysis. For a given model, N and A refer to the number of observations and anomalies in the fitting set, respectively. M identifies the number of categorical variables, with $D_{\boldsymbol{v}}$ and $D_{\boldsymbol{c}}$ identifying the total number of real and categorical columns respectively. Note $D = D_{\boldsymbol{v}} + D_{\boldsymbol{c}}$. For thresholding datasets, q is the threshold quantile such that $b_{q,d} = \hat{F}_d^{-1}(q)$. t is the time (in hours) to fit the model. Discrepancy in D between peaks–over–threshold and rank–transformation reflects differences in data transformation, as well as the additional column for the radial component in the rank–transformed model.

	Raw		Peaks over Threshold								Rank/Cat		Rank-Transform					Categorical		
name	N	A	q	N	A	$D_{\boldsymbol{v}}$	M	$D_{\boldsymbol{c}}$	D	t	N	A	$D_{\boldsymbol{v}}$	M	$D_{\boldsymbol{c}}$	D	t	M	D_c	t
annthyroid	3600	270	0.85	715	150	6	16	32	38	7.45	1200	105	6	16	31	38	4.88			
cardio	1831	176	0.85	715	152	15	10	21	36	9.17	1831	176	19	3	7	27	5.34			
cover	19070	194	0.98	5504	194	9	4	9	18	5.35	1907	20	9	4	9	19	4.31	10	30	5.02
mammography	11183	260	0.95	2390	227	5	5	11	16	5.59	1864	42	6	3	5	12	3.87			
pima	768	268	0.90	205	106	7	6	12	19	1.10	768	268	8	5	10	19	1.99	8	28	1.93
solarflare	1389	12																10	32	3.87
yeast	1484	90	0.90	343	35	6	5	11	17	1.64	1484	90	6	2	5	12	3.09	8	23	2.79

randomly exchanged via a Metropolis step with probability $p_{i,k} = \exp\left\{(t_k^{-1} - t_i^{-1})(E_k - E_i)\right\}$, where E refers to the *energy*, or log-density of the chain at its current state. The sample history of the cold chain, where $t := 1$, is preserved as draws from the posterior distribution. For each example dataset, the sampler was ran for 50 000 iterations, discarding the first 40 000 as burn-in. The resulting chain was thinned, keeping only every 10th iteration. For evaluating density under the posterior predictive distribution, we generate 10 replicates from each iteration kept.

We compared our four proposed scores against three canonical novelty detection algorithms, including isolation forest *iso* (Liu et al., 2008), local outlier factor *lof* (Breunig et al., 2000), and one-class SVM *svm* (Chang and Lin, 2011). Each dataset was subject to 5-fold cross-validation, and out-of-sample performance scores were averaged to produce the resulting performance tables seen in this section. This additional step of cross–validation turned out to be unnecessary for our model, as out–of–sample performance did not markedly differ from

Table 3.2: Area under the *ROC* curve for various anomaly detection schemes, on *strictly categorical* datasets. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

dataset	iso	lof	svm	hknn	hkde	lhkde	lskde
cover	0.384	0.515	0.424	**0.586**	0.523	0.558	0.450
pima	0.620	0.570	0.614	0.457	0.579	0.659	**0.694**
solarflare	**0.893**	0.402	0.887	0.435	0.632	0.768	0.875
yeast	0.620	0.580	0.622	0.406	**0.708**	0.650	0.702

in–sample or full–sample performance for the tested datasets. Table 3.1 provides a summary description of the datasets used in the analysis. For larger datasets, we subsetted the raw data to reduce computation time for the rank–transformation and categorical applications. Note that the categorical versions of *cover*, *pima*, and *yeast* are created from discretizing the rank–transformation subsets. First, we present score efficacy on our purely categorical data model, then mixed scoring with thresholding on continuous variables. Finally we present mixed scoring on rank-transformation data.

3.5.1 Categorical anomalies

The categorical transformation of *cover*, *pima*, and *yeast* discretized the real-valued and ordinal variables in those datasets. For *cover* in particular, it seems this transformation lost a significant amount of data. From Table 3.1, it seems a large portion of data regarding anomalies is contained within the radial component, so a categorical transformation loses that information. Likely for this reason, none of the methods offer exceptional performance on this dataset. The dataset *solarflare* was also unique in our analysis, being the only truly categorical

dataset used. Our algorithm *lskde* very slightly trailed the performance of *one-class SVM*, the best performing algorithm on this dataset. On both *pima* and *yeast*, latent-simplex KDE performed significantly better than any of the canonical methods. On this analysis, *hkde* and *hknn* both performed poorly. It seems the projection of the categorical probability vectors into a unified sphere induces some loss of information.

3.5.2 Peaks-over-Threshold anomalies

We subjected six datasets to multivariate thresholding on their numerical variables, only keeping observations that exceeded the threshold in at least one dimension. Table 3.1 indicates what quantile was used for the threshold, as well as the number of anomalies in excess of the threshold. For *cover*, we further sub-sampled the excesses to produce a more manageably sized dataset. For variables that did not exhibit properties that would allow for a peak-over-threshold model to apply, these variables were instead converted to discrete values with two or three categories. We built the mixed data model, and evaluated performance of the mixed scores, compared against the canonical methods. Of particular interest here is the *annthyroid* dataset, for which all of our scores performed comparably, and significantly better than the canonical scores. Of the other tested datasets, on *cardio*, *lmkde* approached the performance of *isolation forest* and *one-class SVM*, but all other methods performed worse. For the datasets *cover* and *mammography*, *hknn*, *lhkde*, and *lmkde* performed comparably, and each significantly better than any of the canonical methods. We see that *lmkde*, being the inheritor of the latent simplex KDE score, performs reasonably well reliably among datasets thus far in the peaks-over-threshold setting, but is outperformed by other metrics on each dataset. We may see some effect of the loss of information relating to the dependence structure between $\boldsymbol{w}$ and $\boldsymbol{v}$ on the derived performance. On that note, *lhkde* performed comparably to *lmkde* on *annthyroid*, *cover*, *pima*, and *yeast*, but slightly exceeded its performance on *mammography*. We saw in the categorical

Table 3.3: Area under the *ROC* curve for various anomaly detection schemes, on *mixed* data where the real component has undergone the *threshold* standard Pareto transformation. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

dataset	iso	lof	svm	hknn	hkde	lhkde	lmkde
annthyroid	0.458	0.512	0.640	0.691	0.692	**0.698**	0.689
cardio	**0.849**	0.610	0.836	0.590	0.812	0.804	0.823
cover	0.606	0.512	0.684	**0.832**	0.698	0.719	0.714
mammography	0.594	0.616	0.725	0.675	0.750	**0.757**	0.725
pima	0.530	**0.565**	0.511	0.525	0.525	0.524	0.522
yeast	0.427	0.579	0.560	**0.639**	0.522	0.540	0.542

datasets, *lskde* performed generally well, so the projection onto a unified sphere may induce loss of information. In that regard, it may be the case that preserving information about the dependence structure between $\boldsymbol{v}$ and $\boldsymbol{w}$ had a greater effect than a greater effect than preserving information within $\boldsymbol{w}$ specifically.

As to the poor performance of every method on *pima* and *yeast*, these reported *AuROC* values are conditional on the data exceeding the multivariate threshold used in building the model. As we see in Table 3.1, these datasets do not meet the assumption that anomalies are concentrated in the tails. Scores depending on r_n, the radius component of $\boldsymbol{z}_n$, or *magnitude* of the extremal observation, are going to perform poorly relative to metrics that do not make that assumption.

3.5.3 Rank Transformation anomalies

We subjected the same six datasets used in the peak-over-threshold model to rank transformation on the real and ordinal variables. We then built the mixed model including

radius described in Section 3.4.4 on the transformed datasets. Large datasets used in rank-transformation and categorical models were sub-sampled to reduce computation time. Note that rank transformation preserves the entire dataset, so we should not consider the values in Table 3.4 to be comparable to the values in Table 3.3.

Table 3.4: Area under the *ROC* curve for various anomaly detection schemes, on *mixed* data where the real component has undergone the *rank* standard Pareto transformation. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

dataset	iso	lof	svm	hknn	hkde	lhkde	lmkde
annthyroid	0.519	0.561	0.796	0.714	0.817	**0.823**	0.822
cardio	**0.887**	0.588	0.634	0.648	0.847	0.848	0.883
cover	0.898	0.680	0.931	0.833	0.960	0.960	**0.960**
mammography	0.896	0.806	**0.940**	0.700	0.928	0.930	0.845
pima	0.679	0.653	0.712	0.654	0.712	0.707	**0.714**
yeast	**0.675**	0.527	0.632	0.566	0.601	0.593	0.599

Here *lmkde* performs better than each of the canonical methods in four of six datasets, performing slightly worse than *one-class SVM* on *mammography*, and significantly worse than *isolation forest* on *yeast*. As we have stated before, *yeast* and *pima* are datasets that do not quite meet our assumptions as to how anomalies are distributed, but our methods still make a strong showing on *pima*.

3.6 Conclusion

In this chapter, we have proposed a method of *scoring* observations as anomalous based on their posterior-predictive angular density, using the result from multivariate extreme value

theory that—assuming the existence of a limiting behavior—given observations are in excess of a high threshold, after transformation their angular distribution on $\mathbb{S}_\infty^{D-1}$ is independent of the radial distribution on $\mathbb{R}_+$. In the anomaly detection setting, this independence allows us to separate anomaly scores into an angular and radial component, and treat them separately. To define an angular anomaly score, a Bayesian non-parametric model is developed on the angular data projected onto $\mathbb{S}_p^{D-1}$, and as a true density on $\mathbb{S}_\infty^{D-1}$ is not available, anomaly scores are obtained using a non-parametric estimator to that angular density built on a sample from the posterior predictive distribution of the fitted model. The non-parametric estimators we used were k-nearest neighbors, and kernel density estimation.

We then expanded the model to handle categorical data, recognizing that in the real world data does not always fit our assumption of the existence of a limiting behavior. We did this by developing a Bayesian non-parametric categorical data model that provides a general approach for the exploration of the distribution of multivariate data. This was then tied in with the previously defined angular model, providing an approach to mixed data modeling. We explored various methods of defining an anomaly score based on the categorical data, analogous to the scores considered for the angular data making use of of latent class probability vectors. We applied the categorical scores to four datasets, three of which were transformed to be categorical from mixed data. In this analysis, we observed that *lskde* performed reliably well.

In addition, the analysis of six datasets performed with the mixed model indicated that *lmkde* performed reliably well, better than canonical methods most of the time, but was itself outperformed in some cases by other methods that project the latent probability vector along with the angular vector into a unified space. Finally, as the data thresholding process may not always be applicable, we applied the mixed model to data with its angular component transformed via the standard Pareto rank ordering transformation. In this setting, we observed that the latent models—*lmkde* and *lskde*—performed reliably well, as well or better than canonical

methods in five of six tested cases.

In this chapter, we have presented a highly flexible model–based method for anomaly detection that scales to moderately large dimensions and sample sizes. However, as seen in Table 3.1, even for the dimensions and sample sizes presented, model fitting can take several hours. Scaling this model beyond some thousands of observations or tens of columns will require a paradigm shift in *how* the model is fit. For this reason, we are investigating faster means of model fitting, including a variational approach.

Chapter 4

Analysis of Extremal Dependence of Storm Surge using a PoT Model

4.1 Introduction

Storm surge, measured as water height above ground level, can produce flooding as a result of a storm pushing sea-water onto land. Its effects can be catastrophic, potentially disrupting critical infrastructure such as emergency services, logistical services, and military responsiveness. The Sea, Lake, and Overland Surges from Hurricanes (SLOSH)(Jelesnianski, 1992) is a computer model developed by the National Weather Service to simulate storm surge, and its associated inundation caused by hurricanes. Given storm characteristics, the model takes into account local topology, bathymetry, and surge management devices such as levees, to generate a spatial field of inundation—the maximum observed height of water above ground level (or above normal water level for a data point in a body of water) over the duration of the storm at a location. Storm characteristics are data pertaining to the eye of the storm when it made landfall—bearing, velocity, latitude, minimum atmospheric pressure of the storm,

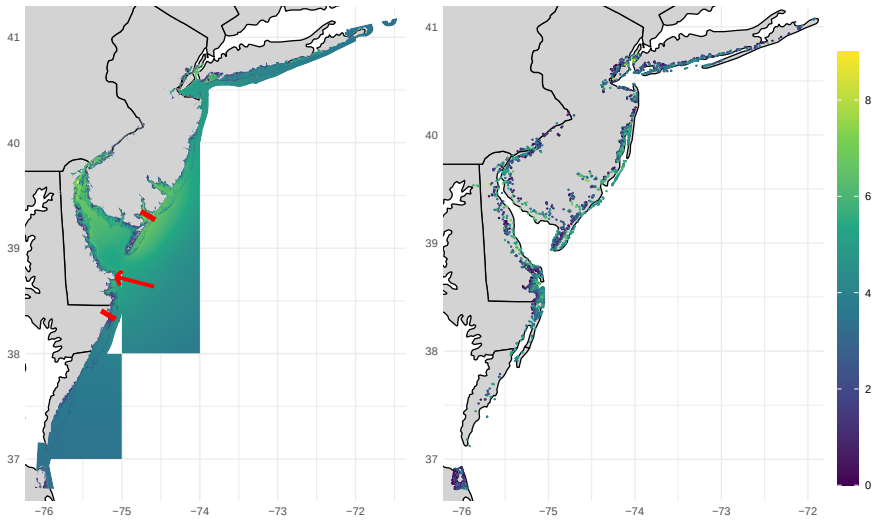


Figure 4.1: (Left) Grid output from one storm simulation in SLOSH, with values truncated at 9 feet. The bars (red) indicate the lower and upper limits on the location of the hurricane eye at landfall. The arrow indicates the direction of travel, and at the vertex, the location of the hurricane eye at landfall for this realization of the SLOSH output grid. (Right) Marginal 90th percentiles of simulated storm-surge at selected locations within the grid.

and projections of sea level rise over time. The example that motivates this work corresponds to a *simulation* from SLOSH, covering an area extending from Virginia Beach, Virginia, to Long Island, New York. The simulation output describes the surge over a grid containing some 23 119 800 elements, with a spatial resolution of 0.001 degrees, or approximately 90 meters. We have 4000 such simulations, produced from a sample of storm characteristics.

Storm parameter inputs for the SLOSH model were sampled via Latin hypercube—a space-filling technique—that attempts to evenly cover the sample space without an imposed grid. Samples thus appear marginally uniform, and lack any observable covariance structure. Figure 4.1 provides a visual depiction of the SLOSH simulation output. On the left, we have the resulting maximum storm surge of a single simulated storm using the SLOSH model. Observe there is data extending from Virginia, near the Chesapeake Bay inlet, to the Eastward tip of

Long Island, New York. In this plot, the observed surge was truncated to 9 feet. There were 45 cells in excess of this, up to 19 feet. Such phenomena are highly localized, and not visible at this scale. We also mention that cell values for a single simulation are not reported simultaneously. Each cell reports its maximum simulated value over the course of the storm. The bars bracketing Delaware Bay indicate the limits of the location of the simulated hurricane eyes when they make landfall, indicating all simulated storms approach the entrance of Delaware Bay. The arrow indicates the bearing, or direction, and location, of the hurricane eye at landfall associated with this particular realization of the storm surge grid. These values range from 200 to 380 degrees; a full 180 degrees, from South-Southwest, to North-Northeast. On the right, we have selected SLOSH grid cells that are in the vicinity of physical features, or locations, of interest, and corresponding 90th percentile of storm surge at each location.

This chapter presents a model for SLOSH simulations under an extreme value theory (EVT) framework, using a peaks-over-threshold model. We seek the joint probability that two or more locations will exhibit extreme behavior. This provides key information in the management of critical infrastructure. We stress here a caveat: EVT assumes that the originating data are independent and identically distributed. SLOSH simulations do not strictly meet this second criterion. They arise as the result of a partially stochastic simulation given a set of input parameters—the storm characteristics. That being said, application of the EVT framework to SLOSH still provides us with a great deal of information. It is with that caveat that we continue the analysis.

The chapter proceeds as follows: Section 4.2 details the theoretical background for the relevant modeling methods and dataset we will be using in this analysis. In particular Section 4.2.1 provides a brief review of variational inference which we attempt to use to speed analysis, and Section 4.2.3 further expands the discussion of SLOSH, along with detailing how the analysis will proceed. Section 4.3 expounds on our methods of posterior analysis, including a

discussion on conditional survival probability and posterior clustering. Further, Section 4.3.2 introduces a novel regression model with support $\mathbb{S}_p^{D-1}$, the positive orthant of the D-dimensional p-norm unit sphere. Section 4.4 presents the results of our analysis, first evaluating the efficacy of variational methods on simulated data as compared to MCMC, then applying our methods to the SLOSH simulation data. Finally, Section 4.5 concludes.

4.2 Review and Background

Let y_{nd} be the storm surge at location d during storm n, after thresholding, transformation, and projection onto $\mathbb{S}_p^{D-1}$. Then, the model can be specified as

$$\begin{aligned} \boldsymbol{y}_n \mid \boldsymbol{\alpha}_n &\sim \mathcal{PG}_p\left(\boldsymbol{Y} \mid \boldsymbol{\alpha}_n, \mathbf{1}\right) & G_0 &= \textstyle\prod_{d=1}^{D} \mathrm{Ga}(\alpha_d \mid \xi_d, \tau_d) \\ \boldsymbol{\alpha}_n &\sim G & \xi_d &\sim \mathcal{G}(\xi \mid a, b) \\ G &\sim \mathcal{PY}\left(\eta, \omega, G_0\right) & \tau_d &\sim \mathcal{G}(\tau \mid c, d) \end{aligned} \tag{4.1}$$

where η and ω are respectively the concentration and discount parameters of the Pitman–Yor process. Fitting this model can be accomplished via Markov-chain Monte Carlo methods. For this purpose, we introduce a latent cluster assignment variable γ_n, sampled as

$$\Pr\left[\gamma_n = \jmath \mid \boldsymbol{y}, \boldsymbol{\alpha}, \boldsymbol{\pi}\right] = \frac{\pi_\jmath \mathcal{PG}(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \mathbf{1})}{\sum_{k=1}^{J} \pi_k \mathcal{PG}(\boldsymbol{y}_n \mid \boldsymbol{\alpha}_\jmath, \mathbf{1})} \tag{4.2}$$

with $\boldsymbol{\pi}$ as described in in Equation (3.2), and the full conditional of $\rho_\jmath$ is

$$\rho_\jmath \mid \boldsymbol{\gamma} \sim \mathcal{B}\left(\zeta_{\rho_\jmath 1} = 1 + C_\jmath - \omega,\ \zeta_{\rho_\jmath 2} = \textstyle\sum_{k>\jmath} C_k + \eta + \jmath\omega\right), \tag{4.3}$$

where $C_\jmath = \sum_{n=1}^{N} \mathbf{1}_{\gamma_n = \jmath}$. We can further introduce a latent variable

$$r_n \mid \boldsymbol{y}, \gamma_n, \boldsymbol{\alpha} \sim \mathrm{Ga}\left(r \mid \sum_{d=1}^{D} \alpha_{\gamma_n d}, \sum_{d=1}^{D} y_{nd}\right)$$

such that conditional on $\boldsymbol{r}$, the likelihood of $\boldsymbol{\alpha}_\jmath$ becomes separable by dimension. Posterior updates for $\boldsymbol{\alpha}$, and $\boldsymbol{\xi}$ can be accomplished via Metropolis-within-Gibbs steps, and the full conditional of τ_d is a gamma distribution.

Such a fitting scheme can work well for a moderately sized inference problem. In Trubey and Sansó (2024), they report requiring approximately 20 minutes to run the sampler for 40 000 iterations, on a problem with 532 observations $\times$ 47 sites. If we want to conduct MCMC inference for more sites, and more observations, the sampler may need more iterations to reach convergence; each iteration will require more CPU time, and the sampler will have an increasing memory footprint. Thus, we consider, as an alternative, a variational inference approach.

4.2.1 Variational Inference - A Brief Overview

Variational inference, or variational Bayesian statistics, is an alternative method of model-fitting that proposes, if the target distribution is analytically intractable, to fit a tractable candidate distribution as *close* to the target distribution as possible (Blei et al., 2017). That is, for data $\boldsymbol{x}$ and some distributional parameter set $\boldsymbol{\theta}$, where $f(\boldsymbol{\theta} \mid \boldsymbol{x})$ is not available in closed form, we select $q(\boldsymbol{\theta})$ from a family of tractable distributions $\boldsymbol{Q}$. Variational inference selects the *optimal* variational distribution q^* by minimizing the KL divergence. Thus,

$$q^*(\boldsymbol{\theta}) = \underset{q \in \mathcal{Q}}{\arg\min} \left\{ \mathrm{KL}\left(q(\boldsymbol{\theta}) || f(\boldsymbol{\theta} \mid \boldsymbol{x})\right) \coloneqq \mathrm{E}_q \left[\log \left(\frac{q(\boldsymbol{\theta})}{f(\boldsymbol{\theta} \mid \boldsymbol{x})} \right) \right] \right\}. \tag{4.4}$$

There are myriad flavors of variational approaches, differentiating by the type and degree of structure, or dependence, they allow between the parameters in the variational distribution, or in the specific method they use to fit the variational distribution. Of particular note here is *mean field* variational Bayes, which is based on an assumption of independence between parameters. That is,

$$q_{\boldsymbol{\theta}} = \prod_{\ell \in L} q_{\theta_\ell}(\theta_\ell \mid \psi_\ell)$$

where $\ell = 1, \ldots, L$ is indexing over distributional parameters of $f(\cdot)$. If we hold each q_{θ_ℓ} to be an appropriate transformation of a normal distribution, then each q_{θ_ℓ} can be specified by a

mean and variance parameter.

Through analytic manipulation we can separate the KL divergence into two quantities: the evidence, $\log f(\boldsymbol{x})$, and the negative of the evidence lower bound, or *ELBO*,

$$\mathcal{L}(\boldsymbol{\theta}) = \mathrm{E}_q \left[\log f(\boldsymbol{x}, \boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\right] = \mathrm{E}_q[\log f(\boldsymbol{x}, \boldsymbol{\theta})] - H, \tag{4.5}$$

where H denotes the entropy of q. As the evidence is constant with respect to q, minimizing the KL divergence is equivalent to maximizing the ELBO, so restating Equation (4.4), we get

$$q^*(\boldsymbol{\theta}) = \arg\max_{q \in \mathcal{Q}} \mathcal{L}(\boldsymbol{\theta}). \tag{4.6}$$

For a given family Q, finding the optimal q^* means finding the optimal parameter set $\boldsymbol{\psi}^*$ such that $q^*(\boldsymbol{\theta}) = q(\boldsymbol{\theta} \mid \boldsymbol{\psi}^*)$. For continuous parameters, optimization of the ELBO with respect to $\boldsymbol{\psi}$ can be accomplished by analytically computing the gradient,

$$\Delta_{\boldsymbol{\psi}} = \frac{\partial}{\partial \psi_\ell} \left\{\mathrm{E}_q \left[\log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right] - H(\boldsymbol{\psi})\right\} \quad \text{for } \ell = 1, \ldots, L, \tag{4.7}$$

then iteratively moving towards the optimal point, where $\Delta_{\boldsymbol{\psi}} = \mathbf{0}$. As the above expectation is not available in closed form for our model, we make implicit use of the reparametrization gradient (Kingma and Welling, 2022), that takes samples of $\boldsymbol{\theta}$ as a function of $\boldsymbol{\psi}$ and an independent R.V. ϵ. This allows us to move the differentiation inside the expectation, and take the expectation numerically. To do the optimization, we use the well-regarded *Adam* optimizer (Kingma and Ba, 2017), a combination of *ADAGRAD* (Duchi et al., 2011) and *RMSProp* (Tieleman, 2012). Adam is stated to be well-suited to noisy and high-dimensional problems, and this problem likely fits the criteria. That said, a path-based optimization approach will be dependent upon the starting position, and if multiple local optima exist, there is no guarantee of reaching the global optimum value. In our analysis, results are highly dependent upon starting position, both in terms of model fidelity, and the resulting number of extant mixture components. One solution would be to consider a better starting position for the optimizer.

As we are considering a mixture model, there is potentially a label switching issue. If the initializing distribution of mixture component parameters provides *decent* coverage of the optimal distribution of mixture parameters, then the most important part of the initialization is the distribution of mixture weights, $\pi_\jmath$. Here we consider three strategies. First is random initialization, (*VB Random*). Second is uniform initialization—initializing the variational parameters such that, after transformation via stick-breaking, the expected probability of cluster assignment is uniform among all clusters up to the truncation point. That is, for truncation point J,

$$\mathrm{E}[\boldsymbol{\rho}] = \left(\frac{1}{J}, \frac{1}{J-1}, \ldots, \frac{1}{3}, \frac{1}{2}\right)$$

leading to $\mathrm{E}[\pi_\jmath] = \frac{1}{J}$ for all $\jmath = 1, \ldots, J$. Finally, we consider *pregaming* the variational algorithm by setting the starting position via an abridged MCMC sampler, which we describe thusly: Sampling an initial position $\boldsymbol{\alpha}_\jmath \sim \mathcal{LN}(\boldsymbol{\alpha} \mid \boldsymbol{\mu} = \mathbf{0}_s, \boldsymbol{\Sigma} = 3I_s)$ and an initial random cluster assignment, we iteratively update $\rho_\jmath$ for $\jmath = 1, \ldots, J$ according to Equation 4.3, γ_n for $n = 1, \ldots, N$ according to Equation 4.2, and $\boldsymbol{\mu}$ via a normal-normal Bayesian update routine. Every iteration, new $\boldsymbol{\alpha}_\jmath$ for empty clusters are resampled from $\mathcal{LN}(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, 3I_s)$. This abridged sampler runs for some small number of iterations; in our testing we used 1000 iterations. Then we set the starting position for the variational algorithm using the last state of the abridged MCMC sampler as follows:

$$\begin{aligned} q_{\alpha_{\jmath d}} &= \mathcal{LN}\left(\psi_{\alpha_{\jmath d\mu}} \coloneqq \log(\alpha_{\jmath d}) - 0.005,\ \psi_{\alpha_{\jmath d\sigma}} \coloneqq 0.1\right) \\ q_{\rho_\jmath} &= \mathrm{Logit}\mathcal{N}\left(\psi_{\rho_\jmath\mu} \coloneqq \psi(\zeta_{\rho_\jmath 1}) - \psi(\zeta_{\rho_\jmath 2}),\ \psi_{\rho_\jmath\sigma^2} \coloneqq \psi'(\zeta_{\rho_\jmath 1}) + \psi'(\zeta_{\rho_\jmath 2})\right), \end{aligned} \tag{4.8}$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions respectively. The distributional values for $q_{\alpha_{\jmath d}}$ are assigned via method of moments, where the standard deviation has been fixed to 0.1. The distributional values for $q_{\rho_\jmath}$ have been assigned following Aitchison and Shen (1980) as the best approximation of a logit-normal to a beta distribution as measured by minimum KL

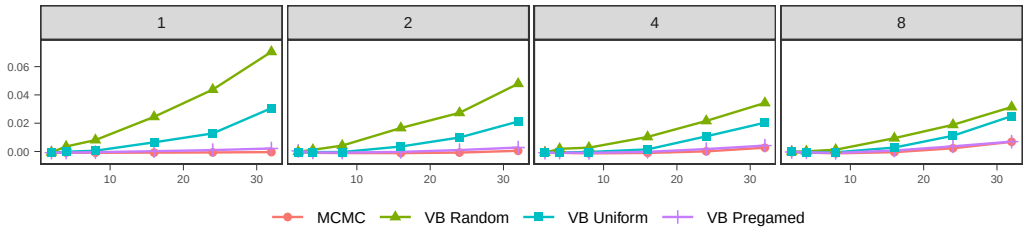


Figure 4.2: Rise in energy score over baseline by number of dimensions, on simulated data for various model fitting strategies. Faceting denotes the number of mixture components in the generating distribution.

divergence. There is a computational burden associated with establishing at least an abridged MCMC sampler, but given the limited number of iterations, and lower number of parameters being learned, that burden is relatively low compared to the full MCMC approach. This approach does have the twin benefits of moving the variational optimizer into a potentially better optima, and reaching convergence faster.

To validate the variational model, and investigate the effects of different classes of starting positions, we conduct a simulation study to evaluate various approaches for the BNP mixture of projected gammas model. The datasets are simulated from a finite mixture of projected gammas, at varying levels of dimensionality and number of mixture components. For each number of mixture components and dimensionality, 10 sets of parameters are generated, and then for each parameter set, a training dataset and testing dataset, each of 1000 replicates, are generated.

As the metric for our evaluation, we use the *energy score* criterion (Gneiting and Raftery, 2007) which is a generalization of the *continuous ranked probability score* to a multivariate setting. The energy score takes the form

$$S_{\text{ES}}(P, \boldsymbol{x}) = \text{E}_p\left[g(\boldsymbol{X}, \boldsymbol{x})\right] - \frac{1}{2}\text{E}_p\left[g(\boldsymbol{X}, \boldsymbol{X}')\right]$$

where g is a kernel function, $\boldsymbol{x}$ is an observed value, and $\boldsymbol{X}, \boldsymbol{X}'$ are posterior predictive replicates of $\boldsymbol{x}$. When the energy score is used with an appropriate negative definite kernel metric, it forms

a *proper* scoring rule. The specific negative definite kernel we use is described in Prop. 3 of Chapter 2, and leverages the fact that all faces of $\mathbb{S}_{\infty}^{D-1}$ are pairwise adjacent. By rotating the face of the second point into the same hyperplane as that of the first, the kernel metric becomes the Euclidean distance between the first point, and the rotated second point.

Figure 4.2 displays the results of our simulation, using the rise in energy score calculated from a posterior predictive sample from the fitted model against the target sample, over a baseline energy score calculated using another random sample from the same generating distribution. Thus, small values indicate high fidelity of the model in capturing the generating distribution. We investigate this recovery under a variety of conditions increasing the number of dimensions, as well as the number of mixture components in the generating distribution. With this simulation study, we see that a pure MCMC approach achieves the best model fidelity, but *VB Pregamed*, using the abridged MCMC sampler to set a starting position for the variational algorithm, achieves a model fidelity that is very nearly indistinguishable from that of the MCMC model, while running significantly faster. Note that, as the dimensionality of the problem grows, energy scores become less able to assess model fidelity, as the distance between each observation or replication will approach a constant value (Bishop and Nasrabadi, 2006).

4.2.2 The Multi-site Return Period and Conditional Survival Probability

Statistical inference of extreme values tends to focus on the *return period* as a deliverable metric. In a univariate case, this is easy to define: for event z, the return period is the average *time* it would take to observe a new event Z as extreme or more extreme than z. That is,

$$T(z) = \frac{\mu}{1 - F_z(z)} = \frac{\mu}{S_z(z)}$$

where μ is the average inter-arrival time in a series of events, and $S(\cdot)$ the survival function. Conceptual problems begin to arise in interpretation when we consider a multivariate F. Salvadori and De Michele (2010) considers a strict interpretation of the return period in a multivariate setting, defining the return period in terms of a copula—a redefinition of the joint CDF, as a distribution over marginal uniform densities, with the marginal CDF's for each site taking the place of the marginal uniforms. We say strict interpretation in interpretation of the CDF: $F(\boldsymbol{z}) = \mathrm{P}\left(\cap_{s\in S} Z_s \leq z_s\right)$, though it considers later a more general *critical region*. Salvadori et al. (2013) inverts this, defining the return period in terms of the joint survival function, $T(\boldsymbol{z}) = \frac{\mu}{S(\boldsymbol{z})}$, where $S(\boldsymbol{z}) = \mathrm{P}\left(\bigcap_{s\in S} Z_s > z_s\right)$. We observe that there exists a vast difference in interpretation between these two extremes. In generating a meaningful deliverable metric of a return period, for a given event at how many sites must the threshold be breached for us to interpret the event as over-topping the threshold? For a CDF based metric, one; for a survival function based metric, all. But as the number of locations under consideration increases, the value of such a metric decreases, as the probability of the specified scenario approaches one or zero respectively. Cho et al. (2023) sidesteps this issue of interpretation by estimating univariate return periods for each indexed location within a spatial field, along with low-dimensional multivariate extreme analysis on different summary statistics, assessed over the aggregate spatial field: flood volume, peak discharge, total rainfall depth, and maximum wind speed. Gräler et al. (2013) also follows the latter approach, establishing a multivariate return period using a copula over three dimensions of summary statistics. Salvadori and De Michele (2010) follows an approach closer to ours, in that they consider the dependence structure of extreme values of the same statistic between different points in space. They consider yearly maximum observed flow rates between 4 of the 17 available flow meters on the catchment of the river Spey, in Scotland. Beyond the conceptual issue of interpreting a multivariate F, there arises a practical issue in presenting complete results of a higher dimensional process. It is for these reasons that

in practice, as a deliverable metric, the notion of a return period is frequently tailored to the application in question. In this chapter, we tailor the survival function thresholds to describe specific scenarios.

4.2.3 Extreme value analysis of SLOSH output

As the motivating example for our analysis of the extremal dependence structure, we use the aforementioned SLOSH, which simulates the storm surge resulting from hurricanes over a wide grid. Our interest is specifically in describing and exploiting the dependence structure of extremes between specific locations. As such, rather than the entire grid, it makes sense to consider data pertaining specifically to those locations. Thus, we subset the data to grid cells which are in the vicinity of such locations of interest. We gather these locations from the *point and landmark* file of the US Census Bureau's 2023 *TIGER* database (U.S. Census Bureau, 2023). We define vicinity as the nearest grid cell within 70 meters of a location—this value stemming from the grid existing on an approximately 110 meter resolution, a 70 meter radius ensures no gaps in coverage. Additionally, we select grid cells that have experienced at at least some inundation in q proportion of storm simulations. That is, such that $b_{q,d} = \hat{F}_d^{-1}(q) > 0$, for all d of interest. This restriction arises as a consequence of fitting the parameters for the marginal generalized Pareto distributions on threshold excesses. The quality of MLE estimation of the other marginal GP parameters will suffer if the threshold is not well chosen, so to ensure that excesses follow a generalized Pareto, we limit analysis to cells that meet this criteria. Here we suffer another trade-off, the implications of which are explored in Figure 4.3. Setting a higher quantile threshold allows more sites to be included in the analysis, but in turn reduces the number of storm simulations which exceed the threshold, which in turn reduces the amount of information available by which we can estimate the dependence structure. Using a quantile of 0.90, the resulting number of cells, and number of storm simulations exceeding the threshold

per slice are summarized in Table 4.1.

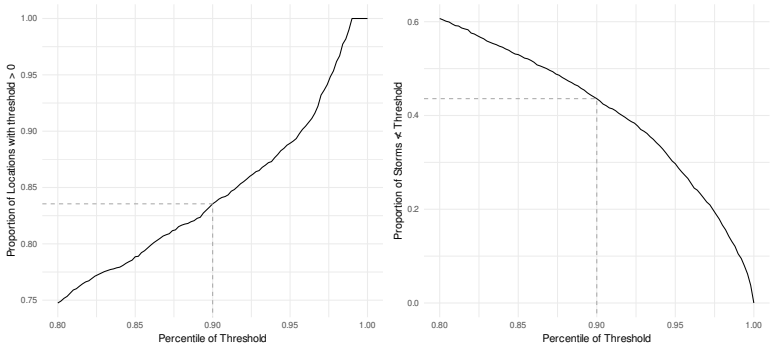


Figure 4.3: Trade-offs in threshold specification: (Left) Proportion of sites with threshold $b_{qs} > 0$ versus $(1 - q)$; (Right) Proportion of storms surviving thresholding, $\Pr[\boldsymbol{W}_n \not< \boldsymbol{b}_q]$ versus $(1 - q)$.

Dataset	Sites	Storms	$\mathrm{P}(\boldsymbol{w} \not\leq \boldsymbol{b}_q)$
Threshold	4414	1744	0.436
Delaware	950	1253	0.313
Restricted	65	1092	0.273
Critical	12	810	0.202

Table 4.1: Slices of SLOSH for analysis. *Storms* specifies the number of storms that survive thresholding, of the total 4000 storms in the sample. The probability gives that value numerically. *Sites* identifies the number of locations included in the slice. Each subsequent slice is a subset of the preceding slice.

The validity of the proposed PoT model depends on asymptotic results that consider observations above a very large threshold. To define a threshold for each location in the simulation output we take the 90th percentile of the observations per location. The resulting thresholded sample corresponds to the *Threshold* slice. Additional reduction of the sample is produced by considering the bounds of the storm approach angles, one of the SLOSH inputs. In Figure 4.1, we see the boundaries of the storm approach vectors, which we see is quite restricted in comparison to the the extent of the surge simulations. This limit in latitudes at landfall

in turn somewhat limits the relevance of the storm surge simulation data to the region where hurricanes making landfall within the bounds of our data would have the greatest effect. We are thus limited to Delaware Bay and the surrounding area. This corresponds to the *Delaware Bay* slice. Further refinement of the sample is performed by focusing on locations of particular interest, identified through the use the feature class codes that correspond to different location types of interest. These form the *Restricted* slice. Finally, we restrict the data to a few locations of critical interest, as well as a selection of locations around the bay to inform conditional analysis. These form the *Critical* slice. The resulting sizes of the different slices of the original simulation output considered in our analysis are reported in Table 4.1.

On the diagonal, Figure 4.4 shows the marginal histograms of SLOSH input parameters, for storms which exhibited extreme behavior in the *Threshold* slice. The off-diagonals are scatter-plots of the storm parameters, which reveal pairwise relationships that contribute to a particular storm exceeding the threshold. Recall that storm parameters in the original simulation were sampled via Latin hypercube, so would appear uniform, with no discernible pattern in the pairwise scatter-plots. The difference between apparently independent uniform, and the observed densities provides some indication of what characteristics are necessary for a storm to exceed the threshold. Imprimis, for sea level rise it is readily apparent that a higher sea level will make it easier for a storm to inundate larger swaths of land, and to a greater degree. So we expect and, in fact, see a higher proportion of storms exceeding the threshold, for a higher sea level rise. Similarly, a lower minimum pressure in the storm's eye corresponds to a more powerful storm. This bears out, as a lower minimum pressure has a higher probability of exceeding the threshold. The relationship to approach speed is interesting in that it is nearly linear. Perhaps, the mechanism there lies in that a higher approach speed indicates more power behind the storm. The spike in approach angle past 360 degrees is interesting as well, especially considering the lull in approach angle between 270 and 350 degrees. 360 degrees indicates due North, thus

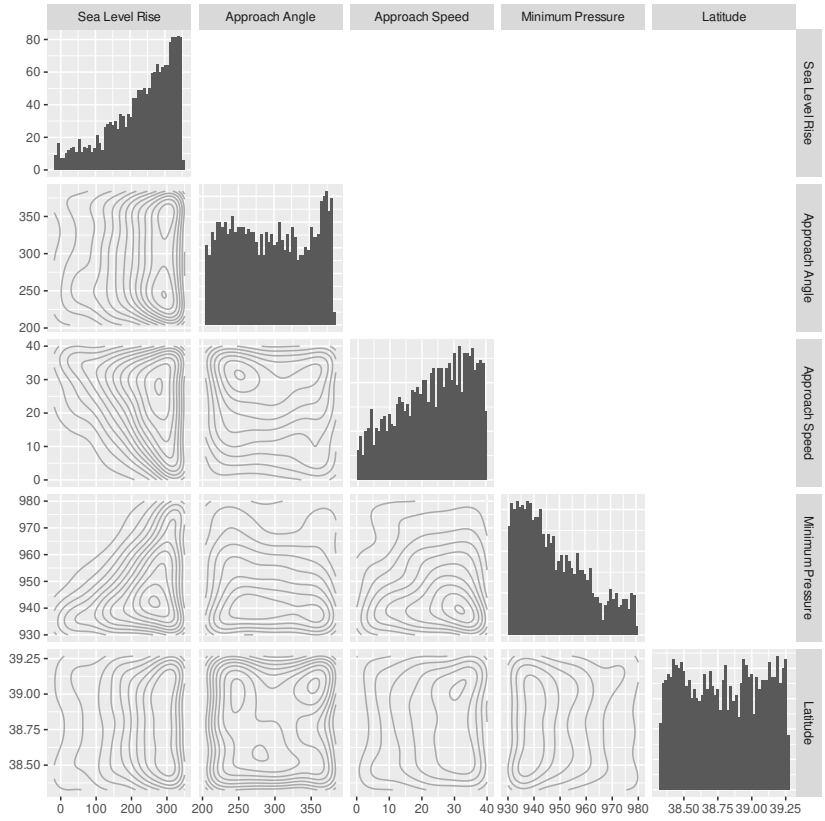


Figure 4.4: (Off-diagonal) Pairwise contour plots of SLOSH simulation inputs that survived thresholding in the *Threshold* slice. Sea level rise in mm, approach angle in degrees (360 is North), approach speed in $\mathrm{km\,h^{-1}}$, minimum pressure in mbar, latitude in decimal degrees. (Diagonal) Marginal histograms of same.

approach angles beyond 360 degrees indicate the storm is heading slightly northeast. As these approaches are on the eastern seaboard, this means a shallower approach angle relative to the land—perhaps offering a given storm more time to inundate larger swaths of land. The apparent lack of interaction between approach angle and latitude is most interesting.

4.3 Methodology

Following the example of Trubey and Sansó (2024), we can use the dependence structure we infer by fitting the model in Equation (4.1). Then, conditional on the fitted model, there are various inferences we can make. One of the more compelling applications of modeling the dependence structure between locations in storm surge lies in the reality that a storm surge occurs over a period of time, and the maximum observed values in storm surge at different sites occur asynchronously. A decision maker, interested in the storm surge at a smaller group of locations, can observe storm surge at other locations, and make inference as to the probability of catastrophic flooding at their locations to make an informed decision. Equation (4.9), from Proposition 2 of Trubey and Sansó (2024) offers us a practical means by which this can be accomplished.

$$\Pr\left[\bigcap_{d\in\alpha} Z_d \geq z_d \;\middle|\; \bigcap_{d\notin\alpha} Z_d \geq z_d\right] = \frac{\mathrm{E}\left[\bigwedge_{d=1}^{D} 1 \wedge \frac{V_d}{z_d}\right]}{\mathrm{E}\left[\bigwedge_{d\notin\alpha} 1 \wedge \frac{V_d}{z_d}\right]} \tag{4.9}$$

Letting Set α be a group of locations of interest, we obtain the probability of entering a failure region, conditional on the current state of the inundation field. Say, given current storm flooding near the mouth of the Delaware Bay, will the Philadelphia International Airport, situated on the Delaware river, experience catastrophic flooding? If one can describe the dependence structure of extremes in inundation analytically or via samples, Equation (4.9) offers a practical, actionable metric.

One point of concern: the framework and tools of multivariate EVT do not allow for the concept of negative dependence. In fact, a peaks-over-threshold approach using the projected gamma to model the dependence structure can not encompass even complete independence: the closest we can represent is weak positive dependence. As such, it is difficult to accurately model phenomena that exhibit mutually exclusive extreme behavior.

4.3.1 Posterior clustering of storms

One potential application leveraging the use of a Bayesian non-parametric prior is exploiting the clustering of observations with similar characteristics. Recall that δ_n is the cluster identifier for observation n. Within the MCMC approach considered in this work, δ_n is explicitly sampled using its full conditional in Equation (4.2). Given samples of $\boldsymbol{\alpha}$, $\boldsymbol{\pi}$, from their posterior density, δ_n can be explicitly generated using the same its full conditional. This approach makes no assumption with regard to the stability of cluster labelling. This can present a problem, as interpretation of clusters requires effective labelling—a fixed assignment of observations to clusters—to avoid label switching. A label-switch between clusters a and b occurs when the parameters of clusters a and b swap, causing the bulk of observations formerly in cluster a to be classified in cluster b, and vice-versa. This issue is not present in a variational implementation, as, once the variational distribution is obtained, the distributions of cluster parameters are fixed. In our example, we find that the label switching concern may be overstated. In the MCMC approach, after the sampler reaches convergence, we find the posterior distribution of cluster assignments and resulting cluster parameters to be relatively stable. Thus, a means of estimating the cluster label of a given observation that is consistent between MCMC and variational approaches is to take samples of $\boldsymbol{\pi}$, and $\boldsymbol{\alpha}$, and then sample $\boldsymbol{\delta}_n$ following Equation (4.2). Using this approach, we group observations by their sampled cluster identifier, and count the number of emergent clusters in the data.

4.3.2 Regression on the unit sphere

To consider storm-relevant information in the estimation of the probability of a significant storm surge event, we develop a novel regression model for angular data in $\mathbb{S}_p^{D-1}$. Consider

$$\boldsymbol{y}_n \sim \mathcal{PG}(\boldsymbol{y} \mid g(\boldsymbol{x}_n^T \boldsymbol{\theta}), \mathbf{1})$$

where $g(\cdot)$ is a link function that maps $\mathbb{R} \to \mathbb{R}_+$ to maintain the viability of inputs for the underlying gamma density. For our purpose, we use the softplus function, $g(x) = \log[1+\exp(x)]$. This asymptotically approaches identity for $x > 0$, and 0 for $x < 0$. The major reason for this choice over the more commonly used $\exp(\cdot)$ is numerical: for softplus, small deviations in inputs produce small deviations in outputs. Some discussion is necessary here as to the dimensionality of $\boldsymbol{x}_n$ and $\boldsymbol{\theta}$. Consider a vector $\boldsymbol{x}_n$ of dimension L, where we might expect each element of $\boldsymbol{x}_n$ to contribute to each dimension s of $\boldsymbol{y}_i$. We call a model *fully specified* if, for each dimension d, we have a vector $\boldsymbol{\theta}_{nd}$, with the same dimensionality as $\boldsymbol{x}$. With some abuse of notation, the fully specified model can be written as

$$\boldsymbol{y}_n \sim \int_0^\infty \prod_{d=1}^{D} \mathcal{G}\left(r_n y_{nd} \mid g(\boldsymbol{x}_n^T \boldsymbol{\theta}_{nd}), 1\right) \times J(\boldsymbol{y}_i) r^{D-1} \mathrm{d}r$$

where $J(\boldsymbol{y}_n)$ is the rest of the Jacobian of the projection. Or more succinctly,

$$\boldsymbol{y}_n \sim \mathcal{PG}\left(\boldsymbol{y}_n \mid g((\boldsymbol{x}_n \otimes \boldsymbol{I}_D)^T \boldsymbol{\theta}), \mathbf{1}\right)$$

where $\otimes$ denotes the Kronecker product. To fully realize the flexibility of this model, we feature it as the kernel density of a Bayesian non-parametric mixture.

$$\begin{aligned}
&\boldsymbol{y}_n \sim \mathcal{PG}\left(\boldsymbol{y} \mid g\left((\boldsymbol{x}_n \otimes \boldsymbol{I}_D)^T \boldsymbol{\theta}_n\right), \mathbf{1}\right) && G_0 = \mathcal{N}(\boldsymbol{\theta} \mid \mu, \Sigma) \\
&\theta_n \sim G && \Sigma \sim \mathcal{IW}(\Sigma \mid \nu, \psi) \\
&G \sim \mathcal{PY}(G \mid \eta, \omega, G_0) && \mu \mid \Sigma \sim \mathcal{N}(\mu \mid \mathbf{0}, \Sigma/\kappa)
\end{aligned} \tag{4.10}$$

Note the cardinality of $\boldsymbol{\theta}$ under the fully specified model is $L \times D$. For practical considerations, this is likely overspecified in application; we consider it here to test for model recovery. In Figure 4.5, we conduct this this simulation example, with two input dimensions, 3 output dimensions, and thus $\boldsymbol{\theta}$ has a cardinality of $2 \times 3 = 6$. For each cluster of inputs, we generate an associated $\boldsymbol{\theta}_j$, and project $y_n = g\left((\boldsymbol{x}_n \otimes \boldsymbol{I}_D)^T \boldsymbol{\theta}_j\right) + \boldsymbol{\epsilon}_n$ where ϵ_{nd} is a small jitter term onto $\mathbb{S}^2_{10}$. The center plot re-projects that onto $\mathbb{S}^2_1$ for display. In the right plot, we have the posterior

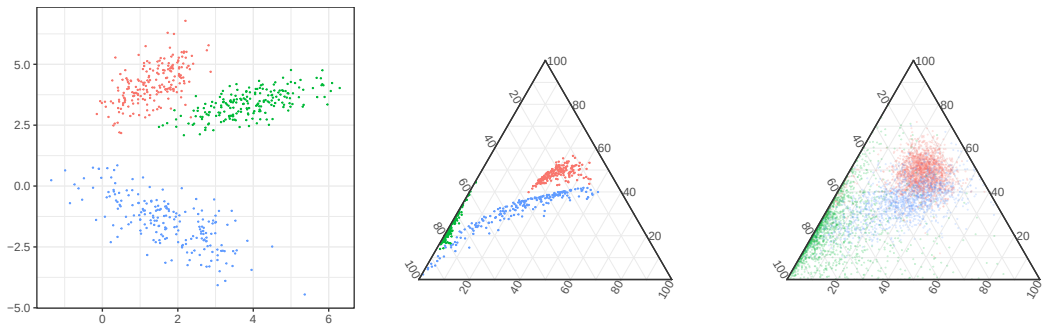


Figure 4.5: Model recovery under a *fully specified* model, colored by cluster. Left is the regressors, $\boldsymbol{X}$. Center is $g(\boldsymbol{x}_n^T\boldsymbol{\theta}_j + \boldsymbol{\epsilon}_n)$ projected onto $\mathbb{S}_1^2$. Right is a sample of the posterior predictive distribution using the same regressors, colored by their emergent cluster identity.

predictive distribution of $\boldsymbol{y}_n^* \mid \boldsymbol{x}_n$. We see that we can reasonably recover the original clusters. In truth, we've given it a hard task, as for the fully specified model and the separated nature of the inputs might mean that a single $\boldsymbol{\theta}$ vector might reasonably cover 2 or more clusters. In the simulation, we find 3 emergent clusters, with stable posterior cluster assignment that matches that of the input data.

The fully specified model is extremely inefficient. As D increases, the dimensionality of $\boldsymbol{\theta}$ increases linearly, which renders it inappropriate for modelling the SLOSH data. However, we can consider other transformations of the data to keep the dimensionality of $\boldsymbol{\theta}$ at an appropriate level. In the SLOSH data, let $\boldsymbol{x}_{nd}$, the covariates associated with observation n at location d, consist of $\boldsymbol{x}_{n,\text{obs}}$, the (scaled) parameters under which the nth storm is modelled, along with $\boldsymbol{x}_{d,\text{loc}}$, information pertaining to the dth location including (scaled) latitude and longitude along with elevation above sea level, and $\boldsymbol{x}_{nd,\text{int}}$, any interaction thereof. We include a single interaction term describing the distance between the location of the storm eye at landfall, and that of location d, in hectomiles. This results in a $\boldsymbol{\theta}$ of dimension $5 + 3 + 1 = 9$. Then we may add an additional random effect by location, ε_d. Thus for analysis of the SLOSH data we

update Equation (4.10) such that

$$\boldsymbol{y}_n \sim \mathcal{PG}\left(\boldsymbol{y} \mid g(\boldsymbol{x}_n^T\boldsymbol{\theta}_n + \boldsymbol{\varepsilon}), \mathbf{1}\right) \qquad \varepsilon_d \sim \mathcal{N}(\varepsilon \mid 0, \sigma_\varepsilon^2) \tag{4.11}$$

where $\boldsymbol{x}_n$ has been overloaded as discussed. We fit this model also under an MCMC framework.

4.4 Results

Slice	Var Bayes	Monte Carlo	Reg w/o RE	Reg w/ RE
Threshold	3	37		
Delaware	4	30		
Restricted	8	21	127	118
Critical	11	51	30	22

Table 4.2: Counts of emergent clusters identified in data slices via posterior sampling.

In Section 4.2.3 we described our criteria for narrowing the focus of our analysis, and specifically in Table 4.1 we detailed the number of observations, and number of storm simulations, for the resulting datasets at each stage of this narrowing process. In Table 4.2, we detail the number of clusters that emerge in fitting our model to these datasets. In all cases, for the Pitman-Yor process parameters, we used a concentration parameter $\eta = 0.1$, along with a discount parameter also of 0.1. For any given model and dataset, the number of emergent clusters found was relatively robust to our choice of parameters, for concentration within a range from 0.01–20, and discount ω within a range 0.001–0.2. Higher values for both generally resulted in slightly more emergent clusters, but did not change the total number by more than 10 percent. Here we encounter an issue: the model fitted via variational methods, and the model fitted via MCMC methods, are nominally the same model, and should result in a similar number

of emergent clusters. That there is such a discrepancy is likely a weakness of the variational fitting process. In the variational Bayes fitting method, we see the number of extant clusters fall significantly as the number of dimensions rises. This behavior is predicted by Chandra et al. (2023), which argues that as dimensionality increases, a BNP mixture model will degenerate to one of two possible states: every observation falling into a single cluster, or every cluster a singleton. We do not quite see that happen yet in the MCMC approach, where we see the number of extant clusters actually rise slightly between Delaware slice and the Threshold slice. However, the dimensionality of the threshold slice is approximately 14.6 times that of the Delaware slice, and the number of simulations which meet the criteria for inclusion in the Threshold slice is approximately 1.4 times that of the Delaware slice. Taking into account the associated increase in data complexity between the two slices, we should see *significantly more* clusters. That we only see marginally more appears to indicate that the BNP mixture of projected gammas will eventually degenerate towards a single cluster. Chandra et al. (2023) suggests an amelioration of this behavior: to instead base the BNP process on a lower-dimensional representation of the output space—suggesting a factor analysis. The regression model we have developed works in a similar manner, representing an D-dimensional space using a L-dimensional vector. In the regression models, we see the opposite problem occurring: there are too many clusters. The Pitman-Yor process is accounting for the lack of information in $\boldsymbol{X}$, the regressors, by producing more clusters. We see evidence to this interpretation by the addition of the random effects. By including random effects in the model specification, more information is contained within the regressors, and thus a single cluster is able to represent slightly more varied outcomes. Thus, the number of emergent clusters is marginally reduced.

Given the geographical focus inherent to the original data, we concentrate our primary analysis on Delaware Bay. Figure 4.6 gives the locations of sites we identified, along with a classification of those sites. The original feature class codes in the TIGER location data were sorted

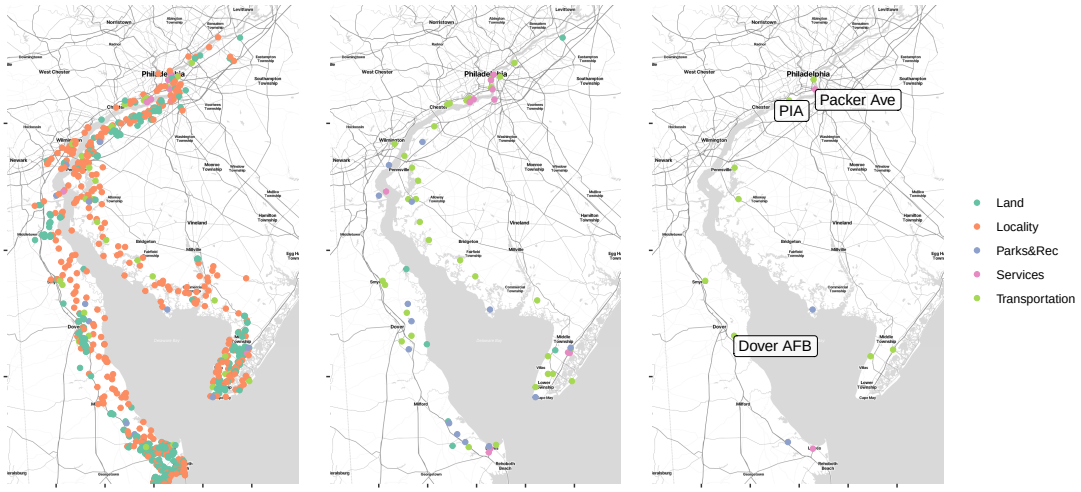


Figure 4.6: Locations of identified sites in the *Delaware* (left), *Restricted* (center), and *Critical* (right) Slices. Three locations of interest in further analysis have been specifically identified.

into five categories: "Land" includes prominent land features, as well as some road features; "Locality" includes major road intersections, communities, or populated places; "Parks&Rec" includes state and local parks, cemeteries, and places of worship; "Services" includes emergency services: police, fire, and medical services, and "Transportation" includes airports, heliports, ferry landings, and other major transportation infrastructure. We identify within the data three locations which are of particular interest, for which significant inundation can can lead to catastrophic consequences. Dover Air Force Base (Dover AFB) is a military installation on the South shore of the Delaware Bay, with a direct line of approach from the ocean. Philadelphia International Airport (PIA) is a major airport, situated on the bank of the Delaware River, near Philadelphia. PIA is much further upstream relative to Dover AFB, and would require a storm to backflow the Delaware River a significant amount to reach it. Packer Avenue Terminal is a major shipping hub, connecting transport ships to truck and rail transport services. It is situated only slightly further upstream than PIA, so we would expect outcomes for these two locations to be strongly dependent. Inundation in any of these three locations could lead to

significant negative consequences.

4.4.1 Assessing Model Fidelity

One difficulty with a high-dimensional model is evaluating its fidelity. As we saw in the simulation study, the relative rise in energy score between a bad modeling approach and a good modeling approach shrinks as data complexity increases. However, as we saw in Table 4.2, MCMC and what appears to be a similarly good modeling approach yield very different outcomes. This disconnect in assessment of model fidelity using energy score as a criterion is related to the curse of dimensionality in applications like k-nearest neighbor algorithms: as the number of dimensions increases, the ratio in average distance between the nearest replicate, and farthest replicate in a sample will tend to approach unity. In this regard, distance, and metrics based on distance will be fundamentally flawed in a high-dimensional setting.

As a partial amelioration of this issue, we can subjectively assess recovery of marginal empirical CDF, by observing the marginal posterior predictive CDF for various locations under our modeling approaches. Having sampled $\boldsymbol{V}^*$ from its posterior predictive density, we can get a sample of W_d^* by inverting Equation (2.1). Thus, for for $R^* \sim \text{Pareto}(1)$, $\boldsymbol{Z}^* = R^*\boldsymbol{V}^*$,

$$W_d^* = a_d \left(\frac{(Z_d^*)^{\xi_d} - 1}{\xi_d} \right) + b_d \tag{4.12}$$

where $\boldsymbol{\xi}$, $\boldsymbol{a}$, and $\boldsymbol{b}$ were previously calculated. For consistency with regard to the originating data, we truncate replicates from the posterior predictive distribution such that $W_d^* \geq 0$.

In Figure 4.7, we observe the marginal empirical and posterior predictive CDFs for storm surge at Dover AFB, Philadelphia International Airport, and Packer Avenue Terminal. With respect to Dover AFB, this location is adjacent to Delaware Bay, approximately 2 miles inland of the bay shoreline, with a direct line of sight to the mouth of the bay and open ocean. Take note in particular, that the empirical cdf of w_d shows that the storm surge does not reach

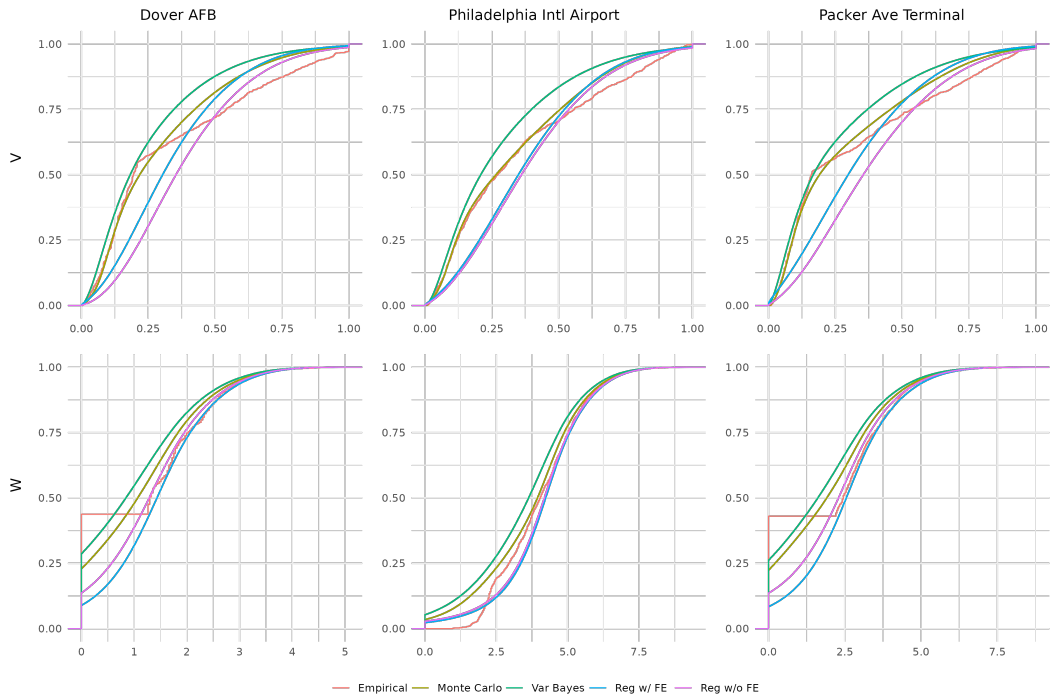


Figure 4.7: Empirical, and posterior predictive cumulative distribution functions for marginal v_d, V_d^* (top), and w_d, W_d^* (bottom) at denoted locations, under various modeling considerations.

Dover AFB in approximately 44 percent of storms, post-thresholding. With respect to PIA, the airport is situated along the banks of the Delaware River. Storm surge has much further to travel to reach this point, yet it experiences inundation much more frequently, owing to the fact it is only 4 feet above sea level, relative to the 9 feet for Dover AFB. Packer Avenue Terminal is similarly situated along the Delaware River, further upstream than PIA, but it experiences flooding significantly less than PIA.

Looking at equivalent marginal plots for all locations (including those not displayed), nearly all preserve the same ordering, from top-left to bottom-right: first the Variational Bayes fit of PYPG, then the Monte Carlo fit of the same model, then the regression models—though the specific ordering of the regression models changes. The marginal empirical CDF for each location tend to favor the regression models or the MCMC fit model. This means that the variational fit

model tends to consistently under-predict, or predict lower values than is appropriate. This fact permits us some insight to comment what effect granularity, or the number of extant clusters, has in model fidelity. In Table 4.2, we saw counts of extant clusters for each model and fitting approach. Variational methods found significantly fewer than MCMC, while the regression models found significantly more. It is perhaps enlightening to realize that the largest single cluster in the variational approaches for all datasets is a cluster with all shape parameters tending towards 0. This parameter set results in an extremely unstable distribution, with its mass concentrated near the edges of the support; specifically edges where only one component is large. Replicates drawing an angular component from this cluster would tend to result in smaller replicates of V_d^*, perhaps explaining the under-prediction of the variational approach overall. That the empirical CDF might favor the regression models should perhaps come as little surprise; with the number of extant clusters large, individual variation is easier to account for. However, in doing so, we are almost certainly over-fitting the model to the data.

4.4.2 Conditional Survival Curves

From Equation 4.9, we can obtain the conditional probability of exceeding a specified threshold for some set of components, given that other dimensions exceed their specified threshold. Using the *Critical* slice with a model fitted via MCMC, we use this equation to establish conditional survival curves for three locations: Dover Air Force Base, Philadelphia International Airport, and Packer Avenue Marine Terminal. In keeping with our goal of a practical actionable metric, we consider three scenarios where we observe extreme behavior further out in the bay than the positions of interest. In the *Lower Bay* scenario, we observe extreme behavior at sites on the south side of the bay towards the entrance of the bay. That is, a scenario in which sites 1,2, and 7 (Beebe Hospital, Henlopen Memorial Park, and Smyrna Airport respectively) experienced storm surge at or above their respective 90th percentiles. In the *Upper Bay* scenario,

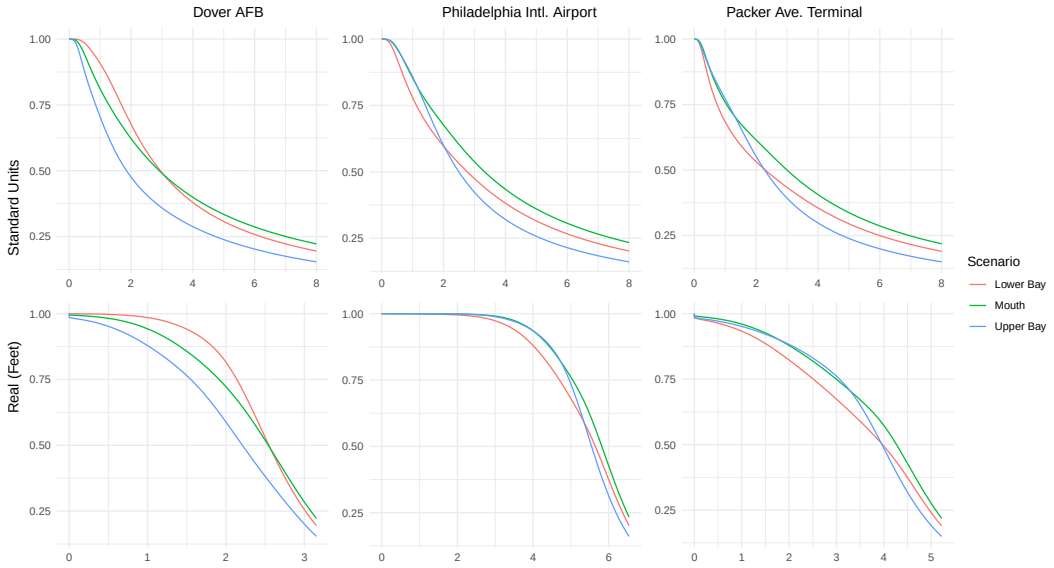


Figure 4.8: Conditional Survival Curves (standardized) for labeled locations, under three scenarios where particular *downstream* locations have already experienced significant inundation. The top row is scaled in standardized units, while the bottom row is in real units (feet).

we observe extreme behavior at sites 6, and 8 (Bay Island Fish and Wildlife Refuge, and Salem Airfield respectively), sites situated along the northern edge of Delaware Bay. In the *Mouth* scenario, we observe extreme behavior at all sites near the mouth of the bay. This includes sites 1, 2, 3, and 4 (Beebe Hospital, Henlopen Memorial Park, Paramount Airport, and the Cape Regional Medical Center).

Figure 4.8 shows the one-dimensional survival curves for these three locations, under these three scenarios. Note that a survival curve indicates $P(Z_s > z_s)$. That is, the probability of a surge event greater than the specified value, under the specific scenarios outlined. Additionally, a $\boldsymbol{z}$ score greater than 1 indicates storm surge above the 90th percentile. Perhaps unsurprisingly in interpreting these results, as Dover AFB is on the south side of the bay, we see the survival curve for the Upper Bay scenario dip below that of both Lower Bay and Mouth scenarios. What is interesting, however, is that that ordering is not uniform; we see the ordering change to that

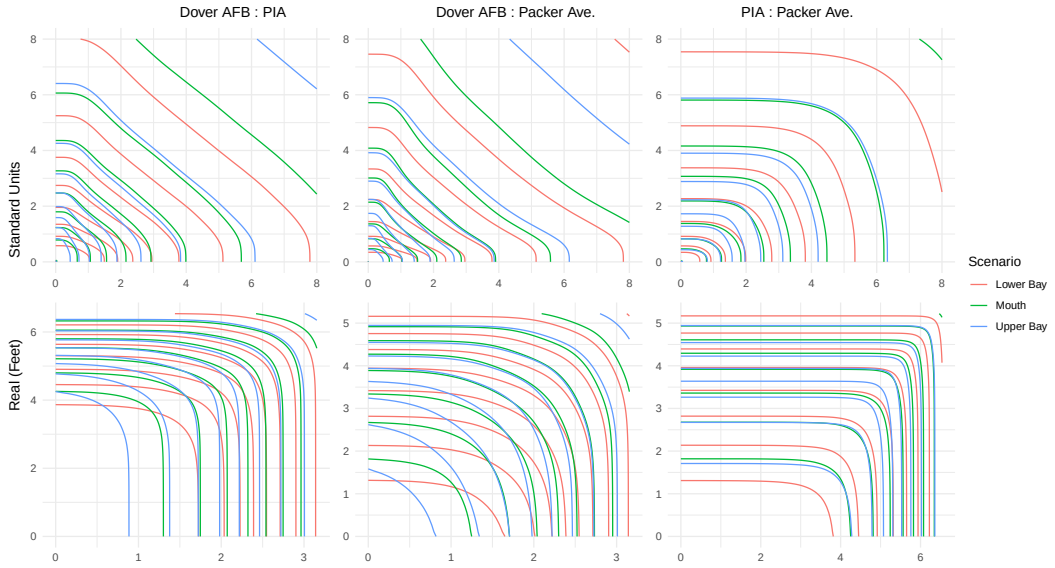


Figure 4.9: Conditional $2d$ survival curve (contour plot, standardized) of flooding, at selected pairs of locations. Note, (X axis) : (Y axis). The top row is in standardized units, while the bottom row is in real units (feet).

behavior around $z = 4$. The Mouth scenario indicates a storm that has inundated both the lower and upper portions of the Delaware Bay entrance, indicating a powerful storm that is well positioned to enter the Bay. As such, it is no surprise that the survival curve associated with that scenario indicates the highest probability of extreme surge throughout the entire curve. What is interesting is that we observe the same crossing behavior and the same ordering on all three curves, though their exact shape and the exact point at which that cross occurs differ. It is apparent that relative to the other scenarios, extreme surge on the Upper Bay sites does not strongly indicate increased surge at the other sites. It appears to be the case that a hurricane optimally aligned towards inundating the North bank is sub-optimally aligned towards inundating the rest of the bay.

Figure 4.9 provides contour plots of a two-dimensional survival surface, between flooding at pairs of locations in the *Critical* slice, still conditioning on various scenarios of flooding in-process. Considering Dover AFB : PIA, we do not expect to see a strong association be-

tween the two locations. Dover AFB is on the south edge of the bay, while PIA is far up the Delaware River. However, we observe that the survival surface in standard units is nearly linear on the transition between the two locations; neither convex nor concave. This shape indicates moderate dependence between those locations under these scenarios. We must call attention to the ordering of the contours of the survival surface. Here, the Lower Bay scenario crosses even the Mouth scenario. In contrast, in Dover AFB : Packer Ave., we see that survival surface is actually concave, which indicates an extremely weak dependence between the two locations. This is understandable, as Packer Avenue Terminal is around 4 miles upstream of PIA, even further away from Dover AFB. Given the proximity between Packer Ave., and PIA, we would expect a rather strong dependence. That relationship is borne out, as in the contour plot of the survival surface in PIA : Packer Ave. we observe strong convexity.

4.4.3 Conditional survival under a regression model

Under the regression model, providing a particular set of storm characteristics $\boldsymbol{\theta}^*$ permits us to take samples of $\boldsymbol{V}^* \mid \boldsymbol{\theta}^*$, by doing posterior prediction of $\boldsymbol{Y}^* \mid \boldsymbol{\theta}^*$ and projecting onto $\mathbb{S}_\infty^{D-1}$. With this sample of $\boldsymbol{V}^*$ following the same procedure used in Section 4.4.2, we can estimate the conditional probability of survival—the conditional probability of experiencing a surge event greater than or equal to the stated value, given both the scenario *and* the storm characteristics. Keeping the extant flooding scenarios outlined previously, we investigate the effects of storm characteristics on these scenarios. A *strong* storm indicates a storm with an approach speed approximately 1 standard deviation higher than mean, and a minimum pressure approximately 1 standard deviation lower than mean. Looking at Figure 4.4, both of these parameters serve to indicate a higher probability of the storm's surge exceeding the threshold, indicating a stronger storm. A *weak* storm indicates the opposite, on both counts. *landing + direction of approach* specifies the approach vector of the storm's eye: landing at the South end

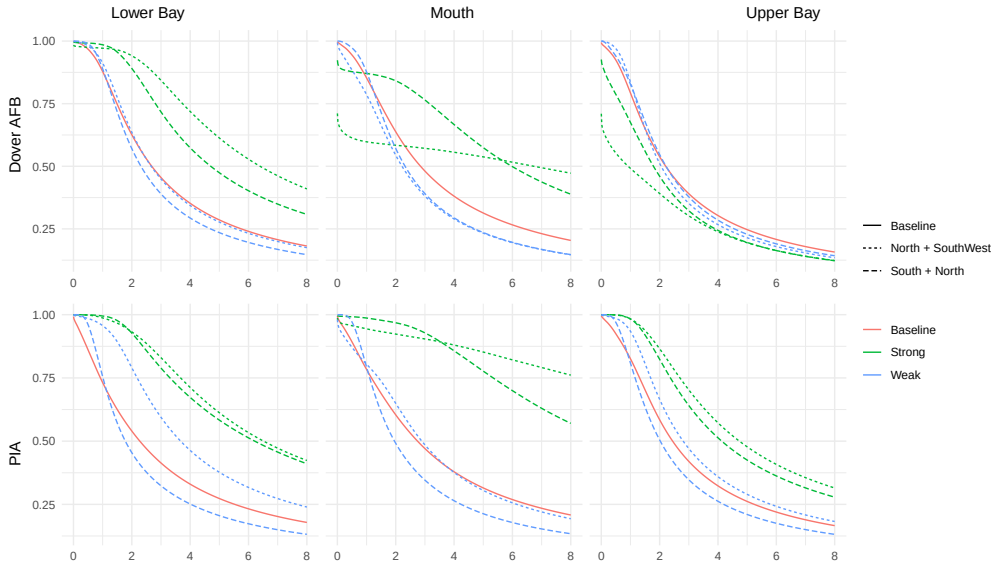


Figure 4.10: Conditional survival curves of flooding, at selected locations, under various scenarios of extant flooding and storm characteristics. A *strong* storm has both a significantly higher than average approach speed, and significantly lower than average minimum pressure. Neutral sea-level-rise was assumed. We further separate by (landing latitude) + (direction of approach).

of the range, angled North, or landing at the North end of the range, angled South.

In Figure 4.10, we see the results of these storm profiles, under the aforementioned scenarios, applied to the listed locations. The *baseline* is not controlling for storm strength, storm landing latitude, or storm direction. It *should* be comparable to the curves displayed in Figure 4.8. There are a few observations to make here. One, it is evident that while a stronger storm generally has a greater potential of inundating any given site to a greater degree, we see odd behavior under the *Upper Bay* scenario for Dover AFB: we see under this scenario the weak storm has a slightly higher potential to inundate, and of particular interest is that all investigated storm profiles fall below the baseline. Second, under the *Mouth* scenario, the landing latitude and direction dramatically affects the probability and degree of inundation. This is most apparent at Dover AFB, for which with a strong storm, the North landing latitude

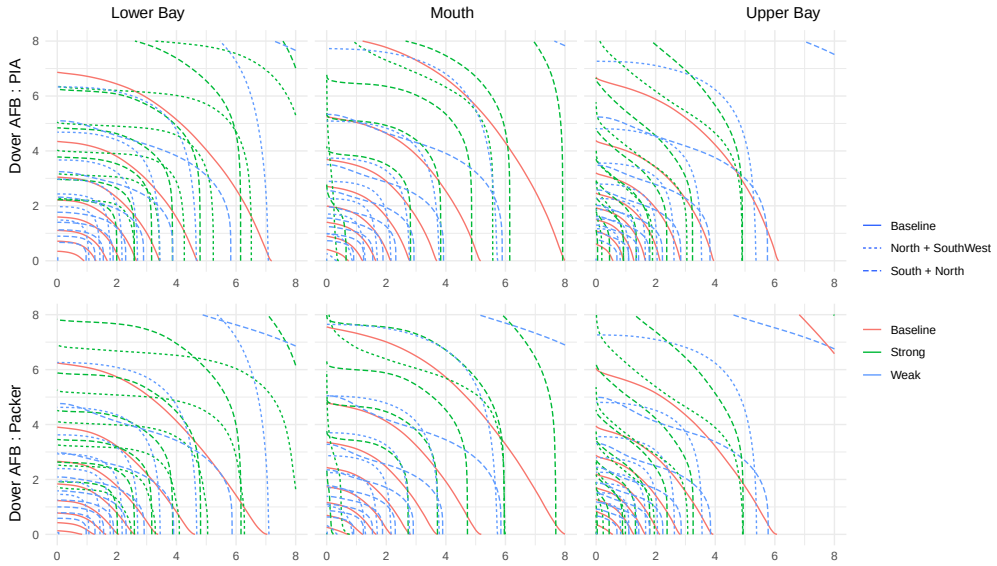


Figure 4.11: Conditional survival surfaces (contour plot, standardized) of flooding, at selected pairs of locations, under various scenarios of extant flooding and storm characteristics. A *strong* storm has both a significantly higher than average approach speed, and significantly lower than average minimum pressure. Neutral sea-level-rise was assumed. We further separate by (landing latitude) + (direction of approach).

and South-southwest direction produces almost uniformly higher probability of flooding at each site. PIA and Packer Avenue fare similarly under each scenario and storm profile. This is somewhat to be expected, considering their proximity, and our previous analysis.

In Figure 4.11, we see the results of these storm profiles, under the aforementioned scenarios, applied pairwise to the listed locations. Again, baseline is not controlling for storm profile or strength, and thus should be comparable to the results displayed in Figure 4.9. Very quickly we see a discrepancy: whereas previously, *Dover AFB : PIA* exhibited moderate dependence, and *Dover AFB : Packer* exhibited extremely weak dependence, under this model they exhibit slightly strong dependence. We no longer see the concave survival surfaces we saw previously under the MCMC-fit projected gamma model. We do still see the highly convex survival surface for *PIA : Packer*.

4.5 Conclusion

In this chapter, we extended our work on the BNP mixture of projected gammas model for angular data to include a variational approximation. Unfortunately, we found that the particular approximation that we chose suffers from a loss of model fidelity, as the number of dimensions increases, to a much greater degree than the equivalent MCMC-fit model. We have also developed a novel and flexible regression model with support on the unit p-norm sphere, in part to ameliorate an expected loss of granularity as dimensionality increases, and for the additional information gain offered by a directional regression model. Though the regression model did impart a strong bias towards pairwise dependence of sites, as seen in Figure 4.11, it nonetheless brings a great deal more information into the survival surface calculation—as we can see in the deviations from the baseline in the survival surfaces.

We explored the application of said models to the angular distribution of multivariate extremes, at a much greater scale than has been previously attempted in our knowledge. In doing so, we have encountered some weaknesses of the model in its application at scale. We discuss those issues and potential remedies below.

4.5.1 Proposed solutions

It is clear that, for our intended purpose of multivariate extreme inference on a high-dimensionality problem, the variational inference method we have chosen is inadequate. A target model where the discrete cluster identifiers have been integrated out leads to too few emergent clusters, and that lack of granularity is detrimental to the model's ability to recover the generating distribution. There are a few potential remedies that we could take. First, following the example of Loaiza-Maya et al. (2022), we might reintroduce the latent cluster identifiers via a Gibbs-sampling step. With that said, an optimal variational approximation

can only do as well as a well-tuned MCMC fitted model. As Chandra et al. (2023) shows, dimensionality increases, model fidelity in recovering the generating distribution will eventually face more challenges than merely an inadequate model fitting method. We should also consider other solutions.

One additional step we can consider is placing a restriction on the shape parameter in the projected gamma model, such that $\max_s \alpha_{js} \geq 1 \ \ \forall \ j$. It is the case, when considering the data, that nearly every observation falls close to an edge on some dimension. Recollecting the gamma distribution, if the shape parameter is less than one, then we observe a spike in density approaching zero. In a projected gamma setting, that translates to high density towards the edges of a distribution where the shape parameters are less than 1. If all dimensions are so, then we arrive at a very unstable distribution with all mass around the edges of the support, and little to no mass in the center. For a *very* high dimensional problem, a single cluster with all shape parameters approaching zero is a likely inevitable end result, and thus, in fitting the model, we learn nothing. One means of tackling such a problem would be to enforce a restriction that at least one dimension have a shape parameter greater than 1. For extreme data, which tends to exist near an edge, this would at least create clusters for edges with high density. We might consider some dimensions as *active* or *inactive* for a given observation depending on whether that observation falls near that dimension's edge. We can consider this restriction as requiring that at least one dimension is active.

Building on this notion of active or inactive dimensions, another extension we might consider would be a *zero-inflated*, or *sparse*, projected gamma, where some dimensions of the projected gamma are structurally zero. With some process controlling the structure of the zero-inflation, we can separate inference on active dimensions from that on inactive dimensions. In the current model with a product-of-gammas centering distribution on $\boldsymbol{\alpha}$, it is currently the case that a single gamma density has to cover cluster parameters for both active dimensions,

away from zero, and inactive dimensions, near zero, resulting in hyper-parameter estimates approaching zero for all dimensions. This is not optimal.

One of the remedies suggested in Chandra et al. (2023) is, rather than clustering on distributional parameters in the target high-dimensional space, to instead cluster on a lower-dimensional representation of that space. That is, a factor analysis model. We have taken a slightly different approach in this chapter by developing a regression model. This is an alternative low-dimensional representation of the high-dimensional space. But, for the regression model we have considered, the number of emergent clusters grows with the number of dimensions. In fact, during model fitting, the 950 location sample quickly consumed all available candidate cluster slots up to the truncation level. We believe this occurs because the proposed model formula is inadequate to handle information contained in the output surface, so it is being absorbed by the cluster representation. Adding additional variables—making the regression model more descriptive of the output surface—will likely result in fewer emergent clusters being necessary, and a higher-fidelity model.

Chapter 5

Conclusion

This document is primarily concerned with representing a peaks-over-threshold setting as a finite-dimensional realization of Ferreira and de Haan (2014)'s constructive definition of the generalized Pareto process as a transformation of a standard Pareto process. This action permits us some freedoms. First, by computing estimates of marginal generalized Pareto parameters, we can use them to transform data into that of a multivariate standard Pareto R.V. Second, under the multivariate standard Pareto framework, we can model the dependence structure of the extremes, expressed as the angular component on the $\mathbb{S}_{\infty}^{D-1}$ sphere, independent of the magnitude, or radial component, and without consideration of the marginal GP parameters.

Chapter 2 establishes a flexible and performant framework for modeling that angular component. It creates a distribution on an arbitrary $\mathcal{L}_p$-norm sphere as the result of projecting a vector of independent random gamma variables onto the surface of said sphere. Using this distribution as the kernel distribution of a Dirichlet process mixture model, we arrive at a highly flexible distribution for modeling angular data with support on $\mathbb{S}_{p}^{D-1}$. Additionally, we establish a negative-definite kernel metric that acts as an upper bound on geodesic distance on $\mathbb{S}_{\infty}^{D-1}$, that we use in conjunction with the energy score criterion, to evaluate model fidelity. We

evaluate our model, and it compares favorably with other possible models of angular data, both in terms of model fidelity, and in terms of computational cost. This model was used to evaluate the dependence structure of extremes in the Integrated Vapor Transport, an atmospheric river carrying moisture inland over California.

Chapter 3 applies the angular data model developed in Chapter 2 to a novelty detection setting, taking advantage of the independence between angular and radial components of the standard Pareto R.V. to develop scores for each component independently. We develop the angular scores as the product of the clustering behavior of the Bayesian non-parametric model, for which we elected to use a Pitman-Yor process rather than a Dirichlet process for greater control over the clustering behavior. This separation of novelty scores allows us to extend the angular novelty score, and indeed the framework of the angular distribution, to a more general class of data. We consider the Dirichlet-multinomial distribution, the result of a multinomial distribution with a Dirichlet prior, with the Dirichlet R.V. integrated out, as the kernel distribution of a Bayesian non-parametric model, and extend the novelty scores previously developed to this new setting. Finally, we consider a *mixed* case, involving both angular and categorical data, and extend the scores to this setting as well. We find the scores we develop compare generally well on canonical novelty detection datasets, though they tend to perform better when novelties are extremely rare.

Chapter 4 attempts to push the limits of the projected gamma model in terms of the complexity it can represent, and our ability to evaluate model fidelity in a high-dimensional or noisy setting. It applies the projected gamma model developed in Chapter 2 with the Pitman-Yor process prior used in Chapter 3 to simulations of storm surge, gathered from the SLOSH model, at varying levels of dimensionality. In Section 4.4.2, we see the fitted model provides us a very powerful tool for predicting extreme levels of storm surge, conditional on that which is already observed. However, we observe that model fidelity, or the ability of the model to

capture the nuances of the data, is lost as dimensionality increases and the granularity of the BNP clustering decreases. This effect is most readily apparent in our variational approximation of the Pitman-Yor mixture of projected gammas, but it is seen in the MCMC approach as well. We also implement a regression model based on the projected gamma. It is obvious that including information pertaining to storm conditions would provide a more rich inference of the multivariate dependence structure, including the conditional survival probability, and the regression model gives us a means of including that data in the calculation. Further, a regression model as a low-dimensional approximation to a high dimensional process, also serves as a means of increasing the granularity in clustering of the BNP mixture model. That said, mixing on regression coefficients, there is a delicate balance to be attained. In our testing, we have too few coefficients, resulting in too many clusters.

There are some potential approaches that might hold merit for our task. First, moving away from mean-field variational Bayes, we can consider a variational approximation within Gibbs approach echoing Loaiza-Maya et al. (2022), by sampling latent cluster identifiers and weights via a Gibbs-sampling step. Second, we can consider placing a restriction on a clusters shape parameter vector such that at least one dimension's shape parameter be greater than 1. This would remove, as a possibility, a single cluster with all shape parameters approaching zero. Third, accepting that all data tends to exist near an edge of the support—that is, at least one dimension is near 0—we can consider internalizing to the model the notion of active or inactive dimensions, using a *zero-inflated* or *sparse* projected gamma. This sparseness can be used to combat the issue of density spiking near an edge, for dimensions with a shape parameter near 0.

Perhaps the most important advancement in this document is the regression model. However, the density is highly multimodal. Model coefficients would benefit from a tempered approach, but the model is already prohibitively slow to fit in a high-dimensional setting. Per-

haps in conjunction with the aforementioned variational approximation within Gibbs, there is room for a tempered variational approach (Mandt et al., 2016) to be considered. Beyond that, we can consider adding more data—a more descriptive regressor vector to include more information about the storm or local conditions. Adding additional information, making the regression model more descriptive of the output surface, will result in fewer clusters being needed, and a higher fidelity model.

Bibliography

Ackerman, M., S. Ben-David, and D. Loker (2010). Characterization of linkage-based clustering. In *COLT*, Volume 2010, pp. 270–281.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological) 44*(2), 139–160.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions: Some properties and uses. *Biometrika 67*(2), 261–272.

Alamuri, M., B. R. Surampudi, and A. Negi (2014). A survey of distance/similarity measures for categorical data. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 1907–1914.

Allen, D. E., A. K. Singh, and R. J. Powell (2013). EVT and tail-risk modelling: Evidence from market indices and volatility series. *The North American Journal of Economics and Finance 26*, 355–369.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Ascolani, F., A. Lijoi, G. Rebaudo, and G. Zanella (2022, 09). Clustering consistency with Dirichlet process mixtures. *Biometrika 110*(2), 551–558.

Bader, B. and J. Yan (2020). *eva: Extreme Value Analysis with Goodness-of-Fit Testing*.

Balkema, A. A. and L. De Haan (1974). Residual life time at great age. *The Annals of probability*, 792–804.

Beirlant, J., Y. Goegebeur, J. Segers, and J. L. Teugels (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons.

Beirlant, J., J. L. Teugels, and P. Vynckier (1994). Extremes in non-life insurance. In *Extreme value theory and applications*, pp. 489–510. Springer.

Beranger, B., S. Padoan, and G. Marcon (2023). *ExtremalDep: Extremal Dependence Models*. R package version 0.0.4-0.

Berg, C., J. P. R. Christensen, and P. Ressel (1984). *Harmonic analysis on semigroups: theory of positive definite and related functions*, Volume 100. Springer.

Berrisford, P., P. Kållberg, S. Kobayashi, D. Dee, S. Uppala, A. Simmons, P. Poli, and H. Sato (2011). Atmospheric conservation properties in ERA-interim. *Quarterly Journal of the Royal Meteorological Society 137*(659), 1381–1399.

Bishop, C. M. and N. M. Nasrabadi (2006). *Pattern recognition and machine learning*, Volume 4. Springer.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association 112*(518), 859–877.

Boldi, M.-O. and A. C. Davison (2007). A mixture model for multivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 217–229.

Boltz, S., E. Debreuve, and M. Barlaud (2009). High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing 18*(6), 1266–1283.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.

Castro-Camilo, D., M. de Carvalho, and J. Wadsworth (2018). Time-varying extreme value dependence with application to leading european stock markets. *Annals of Applied Statistics 12*, 283–309.

Chakraborty, S. and S. W. K. Wong (2021). BAMBI: An R package for fitting bivariate angular mixture models. *Journal of Statistical Software 99*(11), 1–69.

Chalapathy, R., A. K. Menon, and S. Chawla (2018). Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*.

Chandola, V., A. Banerjee, and V. Kumar (2009, jul). Anomaly detection: A survey. *ACM Compututing Surveys 41*(3).

Chandra, N. K., A. Canale, and D. B. Dunson (2023). Escaping the curse of dimensionality in bayesian model-based clustering. *Journal of machine learning research 24*(144), 1–42.

Chang, C.-C. and C.-J. Lin (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chau, D. and M. van Dijck Nemcsik (2020). *Anti-Money Laundering Transaction Monitoring Systems Implementation: Finding Anomalies*. John Wiley & Sons.

Chiapino, M., S. Clémençon, V. Feuillard, and A. Sabourin (2020). A multivariate extreme value

theory approach to anomaly clustering and visualization. *Computational Statistics 35*(2), 607–628.

Cho, E., E. Ahmadisharaf, J. Done, and C. Yoo (2023). A multivariate frequency analysis framework to estimate the return period of hurricane events using event-based copula. *Water Resources Research 59*(12), e2023WR034786.

Clifton, D. A., S. Hugueny, and L. Tarassenko (2011). Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems 65*(3), 371–389.

Coles, S. G. (2001). *An Introduction to Statistical Modelling of Extreme Values*. Springer.

Cooley, D., R. A. Davis, and P. Naveau (2010). The pairwise beta distribution: A flexible parametric multivariate model for extremes. *Journal of Multivariate Analysis 101*(9), 2103–2117.

Cooley, D., D. Nychka, and P. Naveau (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association 102*(479), 824–840.

Corbella, S. and D. D. Stretch (2012). Multivariate return periods of sea storms for coastal erosion risk assessment. *Natural Hazards and Earth System Sciences 12*(8), 2699–2708.

Davis, N., G. Raina, and K. Jagannathan (2019). Lstm-based anomaly detection: detection rules from extreme value theory. In *EPIA Conference on Artificial Intelligence*, pp. 572–583. Springer.

Davis, N., G. Raina, and K. Jagannathan (2020). A framework for end-to-end deep learning-based anomaly detection in transportation networks. *Transportation Research Interdisciplinary Perspectives 5*, 100112.

de Carvalho, M. (2016). Statistics of extremes: Challenges and opportunities. *Extreme events in finance: A handbook of extreme value theory and its applications*, 195–213.

de Carvalho, M., A. Kumukova, and G. Dos Reis (2022). Regression-type analysis for multivariate extreme values. *Extremes*, 1–28.

De Haan, L. and A. Ferreira (2006). *Extreme value theory: an introduction*, Volume 21. Springer.

Dee, D. P., S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, 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.

Deng, L., C. Ma, and W. Yang (2011). Portfolio optimization via pair copula-GARCH-EVT-CVaR model. *Systems Engineering Procedia 2*, 171–181.

Duchi, J., E. Hazan, and Y. Singer (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research 12*(7).

Earl, D. J. and M. W. Deem (2005). Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics 7*(23), 3910–3916.

Einmahl, J. H., V. I. Piterbarg, and L. De Haan (2001). Nonparametric estimation of the spectral measure of an extreme value distribution. *The Annals of Statistics 29*(5), 1401–1423.

Einmahl, J. H. and J. Segers (2009a). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics*, 2953–2989.

Einmahl, J. H. J. and J. Segers (2009b). Maximum empirical likelihood estimation of the spectral measure of an extreme-value distribution. *The Annals of Statistics 37*(5B), 2953–2989.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Escobar-Bach, M., Y. Goegebeur, and A. Guillou (2018). Local robust estimation of the pickands dependence function. *The Annals of Statistics 46*(6A), 2806–2843.

Ester, M., H.-P. Kriegel, J. Sander, X. Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, Volume 96, pp. 226–231.

Falk, M. and A. Guillou (2008). Peaks-over-threshold stability of multivariate generalized Pareto distributions. *Journal of Multivariate Analysis 99*, 715–734.

Falk, M. and R. Michel (2009). Testing for a multivariate generalized Pareto distribution. *Extremes 12*(1), 33–51.

Falk, M., S. A. Padoan, and F. Wisheckel (2019). Generalized Pareto copulas: A key to multivariate extremes. *Journal of Multivariate Analysis 174*, 104538.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics 2*, 615–629.

Fernando Ferraz do Nascimento, D. G. and H. F. Lopes (2012). A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing 22*(2), 661–675.

Ferreira, A. and L. de Haan (2014). The generalized Pareto process; with a view towards application and simulation. *Bernoulli 20*(4), 1717–1737.

Fisher, R. A. and L. H. C. Tippett (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical proceedings of the Cambridge philosophical society 24*(2), 180–190.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Math. Polon. 6*, 93–116.

Gelfand, A. E. and S. K. Ghosh (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika 85*(1), 1–11.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association 102*(477), 359–378.

Goix, N., A. Sabourin, and S. Clémençon (2017). Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis 161*, 12–31.

Gräler, B., M. J. van den Berg, S. Vandenberghe, A. Petroselli, S. Grimaldi, B. De Baets, and N. E. C. Verhoest (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation. *Hydrology and Earth System Sciences 17*(4), 1281–1296.

Green, P. J., K. Łatuszyński, M. Pereyra, and C. P. Robert (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing 25*(4), 835–862.

Gu, X., S. Yang, Y. Sui, E. Papatheou, A. D. Ball, and F. Gu (2021). Real-time novelty detection of an industrial gas turbine using performance deviation model and extreme function theory. *Measurement 178*, 109339.

Guan, B. and D. E. Waliser (2015). Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *Journal of Geophysical Research: Atmospheres 120*(24), 12514–12535.

Guillotte, S., F. Perron, and J. Segers (2011). Non-parametric Bayesian inference on bivariate extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(3), 377–406.

Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré 5*(2), 115–158.

Gumbel, E. J. (1942). On the frequency distribution of extreme values in meteorological data. *Bulletin of the American Meteorological Society 23*(3), 95–105.

Haario, H., E. Saksman, and J. Tamminen (2001). An adaptive metropolis algorithm. *Bernoulli 7*, 223–242.

Hanson, T. E., M. de Carvalho, and Y. Chen (2017). Bernstein polynomial angular densities of multivariate extreme value distributions. *Statistics and Probability Letters 128*, 60–66.

Hanson, T. E., M. de Carvalho, and Y. Chen (2017). Bernstein polynomial angular densities of multivariate extreme value distributions. *Statistics & Probability Letters 128*, 60–66.

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics) 28*(1), 100–108.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049.

Hutchings, G. C., B. Sanso, J. R. Gattiker, D. C. Francom, and D. Pasqualini (2021). Comparing

emulation methods for a high-resolution storm surge model. Technical Report UCSC-SOE-21-06, University of California, Santa Cruz, Santa Cruz, CA.

Ishwaran, H. and L. F. James (2001a). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*(453), 161–173.

Ishwaran, H. and L. F. James (2001b). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association 96*(453), 161–173.

Jelesnianski, C. P. (1992). *SLOSH: Sea, lake, and overland surges from hurricanes*, Volume 48. US Department of Commerce, National Oceanic and Atmospheric Administration.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society 81*(348), 158–171.

Jentsch, A., J. Kreyling, and C. Beierkuhnlein (2007). A new generation of climate-change experiments: events, not trends. *Frontiers in Ecology and the Environment 5*(7), 365–374.

Kingma, D. P. and J. Ba (2017). Adam: A method for stochastic optimization.

Kingma, D. P. and M. Welling (2022). Auto-encoding variational bayes.

Kiriliouk, A., H. Rootzén, J. Segers, and J. L. Wadsworth (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics 61*(1), 123–135.

Kramer, O. (2013). K-nearest neighbors. In *Dimensionality reduction with unsupervised nearest neighbors*, pp. 13–23. Springer.

Li, C., F. Zwiers, X. Zhang, G. Chen, J. Lu, G. Li, J. Norris, Y. Tan, Y. Sun, and M. Liu (2019). Larger increases in more extreme local precipitation events as climate warms. *Geophysical Research Letters 46*(12), 6885–6891.

Lijoi, A. and I. Prünster (2009). Distributional properties of means of random probability measures. *Statistics Surveys 3*, 47–95.

Liu, F. T., K. M. Ting, and Z.-H. Zhou (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422.

Loaiza-Maya, R., M. S. Smith, D. J. Nott, and P. J. Danaher (2022). Fast and accurate variational inference for models with many latent variables. *Journal of Econometrics 230*(2), 339–362.

Mack, Y. and M. Rosenblatt (1979). Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis 9*(1), 1–15.

Mackay, E. and P. Jonathan (2020). Assessment of return value estimates from stationary and non-stationary extreme value models. *Ocean Engineering 207*, 107406.

Mandt, S., J. McInerney, F. Abrol, R. Ranganath, and D. Blei (2016). Variational tempering.

Mardia, K. V., P. E. Jupp, and K. Mardia (1999). *Summary Statistics*, Chapter 2, pp. 13–24. John Wiley & Sons, Ltd.

McNicholas, P. D. (2010). Model-based classification using latent Gaussian mixture models. *Journal of Statistical Planning and Inference 140*(5), 1175–1181.

Mhalla, L., M. de Carvalho, and V. Chavez-Demoulin (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.

Müller, P., F. A. Quintana, A. Jara, and T. Hanson (2015). *Bayesian nonparametric data analysis*, Volume 1. Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Neiman, P. J., A. B. White, F. M. Ralph, D. J. Gottas, and S. I. Gutman (2009). A water vapour flux tool for precipitation forecasting. In *Proceedings of the Institution of Civil Engineers-Water Management*, Volume 162, pp. 83–94. Thomas Telford Ltd.

Núñez-Antonio, G. and E. Geneyro (2019). A multivariate projected Gamma model for directional data. *Communications in Statistics - Simulation and Computation*, 1–22.

Oscar (2015). Dynstatcov - cython library for fast dynamic statistical co-variance update.

Padoan, S. A., M. Ribatet, and S. A. Sisson (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association 105*(489), 263–277.

Pappas, T. (1989). *The joy of mathematics: Discovering mathematics all around you.* Wide World Pub Tetra.

Park, D., H. Kim, and C. C. Kemp (2019). Multimodal anomaly detection for assistive robots. *Autonomous Robots 43*(3), 611–629.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics 33*(3), 1065–1076.

Perman, M., J. Pitman, and M. Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields 92*(1), 21–39.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics 3*(1), 119–131.

Ralph, F., T. Coleman, P. Neiman, R. Zamora, and M. Dettinger (2013). Observed impacts of

duration and seasonality of atmospheric-river landfalls on soil moisture and runoff in coastal northern california. *Journal of Hydrometeorology 14*(2), 443–459.

Ralph, F. M., M. D. Dettinger, M. M. Cairns, T. J. Galarneau, and J. Eylander (2018). Defining "atmospheric river": How the glossary of meteorology helped resolve a debate. *Bulletin of the American Meteorological Society 99*(4), 837–839.

Ralph, F. M., S. Iacobellis, P. Neiman, J. Cordeira, J. Spackman, D. Waliser, G. Wick, A. White, and C. Fairall (2017). Dropsonde observations of total integrated water vapor transport within north Pacific atmospheric rivers. *Journal of Hydrometeorology 18*(9), 2577–2596.

Raquel Barata, Raquel Prado, B. S. (2020). Fast inference for time-varying quantiles via flexible dynamic models with application to the characterization of atmospheric rivers. Technical Report UCSC-SOE-20-14, Santa Cruz, CA.

Renard, B. and M. Lang (2007). Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources 30*(4), 897–912.

Resnick, S. (2008). *Extreme Values, Regular Variation, and Point Processes*. Applied probability. Springer.

Roberts, S. J. (1999). Novelty detection using extreme value statistics. *IEE Proceedings-Vision, Image and Signal Processing 146*(3), 124–129.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate peaks over thresholds models. *Extremes 21*(1), 115–145.

Rootzén, H. and N. Tajvidi (2006). Multivariate generalized Pareto distributions. *Bernoulli 12*(5), 917–930.

Rootzén, H., J. Segers, and J. L. Wadsworth (2018). Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis 165*, 117–131.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics 27*(3), 832 – 837.

Sabourin, A. (2023). *BMAmevt: Multivariate Extremes: Bayesian Estimation of the Spectral Measure*. R package version 1.0.5.

Sabourin, A. and P. Naveau (2014). Bayesian Drichlet mixture model for multivariate extremes: A re-parametrization. *Computational Statistics & Data Analysis 71*, 542–567.

Sabourin, A., P. Naveau, and A.-L. Fougères (2013, September). Bayesian model averaging for multivariate extremes. *Extremes 16*(3), 325–350.

Salvadori, G. and C. De Michele (2010). Multivariate multiparameter extreme value models and return periods: A copula approach. *Water Resources Research 46*(10).

Salvadori, G., C. De Michele, and F. Durante (2011). On the return period and design in a multivariate framework. *Hydrology and Earth System Sciences 15*(11), 3293–3305.

Salvadori, G., F. Durante, and C. De Michele (2013). Multivariate return period calculation via survival functions. *Water Resources Research 49*(4), 2308–2311.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Stephenson, A. G. (2002, June). evd: Extreme value distributions. *R News 2*(2), 0.

Suboh, S. and I. A. Aziz (2020). Anomaly detection with machine learning in the presence of extreme value-a review paper. In *2020 IEEE Conference on Big Data and Analytics (ICBDA)*, pp. 66–72. IEEE.

Thomas, A., S. Clémençon, A. Gramfort, and A. Sabourin (2017). Anomaly detection in extreme regions via empirical mv-sets on the sphere. In *Artificial Intelligence and Statistics*, pp. 1011–1019. PMLR.

Tieleman, T. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning 4*(2), 26.

Tran, M.-N., T.-N. Nguyen, and V.-H. Dao (2021). A practical tutorial on variational bayes.

Trepanier, J. C. and C. S. Tucker (2018). Event-based climatology of tropical cyclone rainfall in Houston, Texas and Miami, Florida. *Atmosphere 9*(5), 170.

Trubey, P. and B. Sansó (2024). Bayesian non-parametric inference for multivariate peaks-over-threshold models. *Entropy 26*(4).

Tuo, R. (2018). Uncertainty quantification with α-stable-process models. *Statistica Sinica*, 553–576.

U.S. Census Bureau (2023). TIGER Line Shapefiles — Point & Landmark. U.S. Department of Commerce. [Accessed 02-12-2024].

Van Engelen, J. E. and H. H. Hoos (2020). A survey on semi-supervised learning. *Machine Learning 109*(2), 373–440.

Vignotto, E. and S. Engelke (2020). Extreme value theory for anomaly detection–the gpd classifier. *Extremes 23*(4), 501–520.

Vousdoukas, M. I., L. Mentaschi, E. Voukouvalas, M. Verlaan, S. Jevrejeva, L. P. Jackson, and L. Feyen (2018). Global probabilistic projections of extreme sea levels show intensification of coastal flood hazard. *Nature communications 9*(1), 1–12.

Warner, M. and B. Sansó (2018). Comparison and assessment of the extremes of different types of climate model simulations. Technical report, University of California Santa Cruz.

Weibull, W. et al. (1951). A statistical distribution function of wide applicability. *Journal of applied mechanics 18*(3), 293–297.

Xiang, G. and R. Lin (2021). Robust anomaly detection for multivariate data of spacecraft through recurrent neural networks and extreme value theory. *IEEE Access 9*, 167447–167457.

Zhang, F., N. Lin, and H. Kunreuther (2023). Benefits of and strategies to update premium rates in the us national flood insurance program under climate change. *Risk Analysis 43*(8), 1627–1640.

Appendix A

Ancillary Material

A.1 Additional Conditional Survival Curves

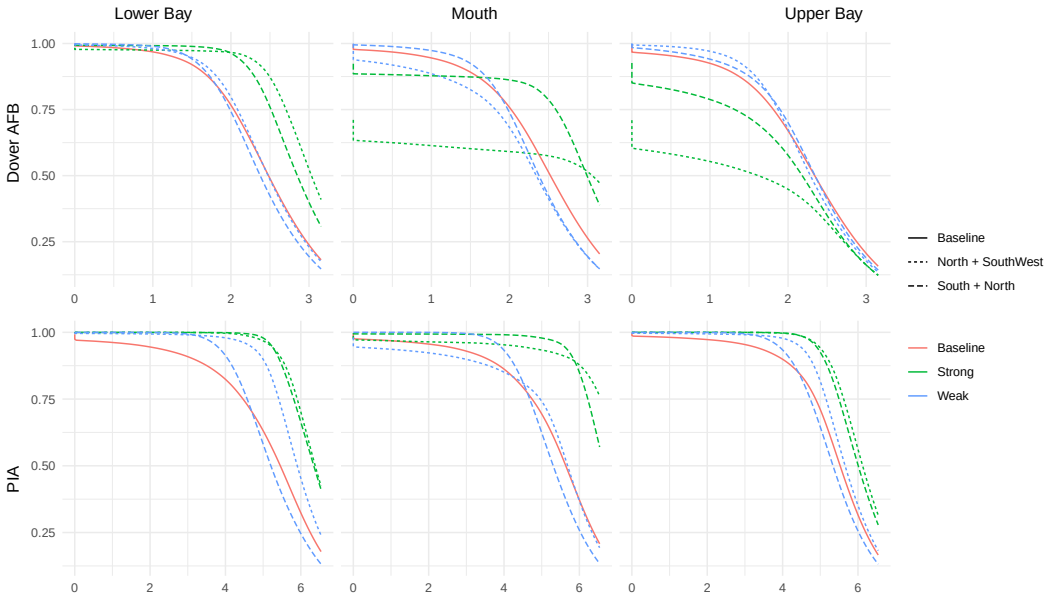


Figure A.1: Conditional survival curves of flooding, at selected locations, under various scenarios of extant flooding and storm characteristics. A *strong* storm has both a significantly higher than average approach speed, and significantly lower than average minimum pressure. Neutral sea-level-rise was assumed. We further separate by (landing latitude) + (direction of approach).

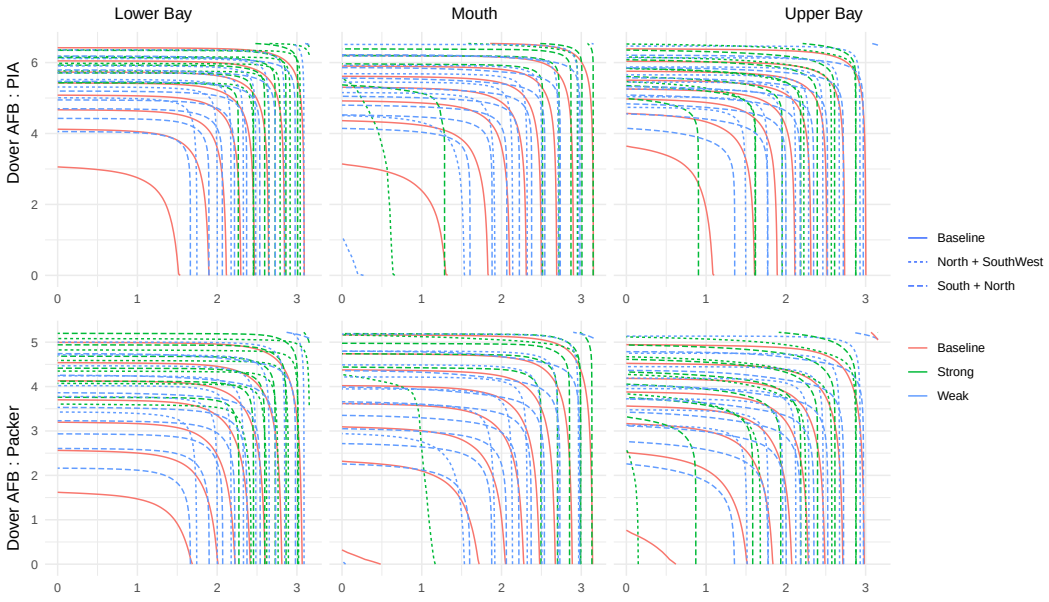


Figure A.2: Conditional survival surfaces (contour plot, standardized) of flooding, at selected pairs of locations, under various scenarios of extant flooding and storm characteristics. A *strong* storm has both a significantly higher than average approach speed, and significantly lower than average minimum pressure. Neutral sea-level-rise was assumed. We further separate by (landing latitude) + (direction of approach).