Anomaly Detection using Extreme Value Theory

Peter Trubey

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Contents

1 Introduction

2 Background Information

Extreme value theory (EVT) seeks to model and assess probability of observing extreme events. Such a topic is applicable generally, but it finds particularly strong use among such fields as finance, climatology, and insurance. [Find Citation]In these fields, extreme events may represent significant loss to the body commissioning the study. For instance, an insurance company might commission a study on extreme weather events, as a localized extreme event could cause a spike in claims from that region.

2.1 Univariate EVT - Maxima

Extreme value theory describes the asymptotic behavior of extreme events. For a sample $\mathbf{x}$ where $\mathbf{x} = (x_1, \ldots, x_n)$ represents a sequence of independent random variables from a distribution function F, the distribution of the maximum M_n of this sequence can be derived as:

$$\begin{aligned}\Pr(M_n \leq z) &= \Pr(X_1 \leq z, \ldots, X_n \leq z) \\ &= \Pr(X_1 \leq z) \times \ldots \times \Pr(X_n \leq z) \\ &= F(z)^n\end{aligned}$$

Where F is unknown, we seek to approximate the behavior of F^n as $n \to \infty$. To ensure this doesn't degenerate to a point mass, we select a sequence of constants $a_n > 0$, b_n and define M_n' as $M_n' = (M_n - b_n)/a_n$, where b_n represents the location, and a_n the scale. These sequences stabilize as n increases, which creates a limiting distribution for M_n'. To summarize, if there exists some sequence of constants $a_n > 0$, b_n such that:

$$\Pr\left[\frac{M_n - b_n}{a_n} \leq z\right] \xrightarrow{d} G(z)$$

as $n \to \infty$, then G is a max-stable distribution, and F is in the domain of attraction of that max stable distribution. The literature identifies 3 known max-stable distributions (Frechet, Weibull, and Gumbel) corresponding to different domains of attraction, and identifies them all as special cases of the *Generalized Extreme Value* Distribution (GEV).

$$F(m \mid \mu, \sigma, \xi) = \exp\left\{-\left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

Thus, if F is in the domain of attraction of any EVD, it will be in the domain of attraction of the GEV. Max-stable distributions feature the homogeneity property:

$$G^n(a_n z + b_n) = G(z)$$

For estimation, generally it will occur that we identify some natural block for the data. For example, for temperature data taken hourly, we might identify a day (24 hours) as a block. There's an implicit violation of the assumption of independence within a block, but that violation is generally ignored. In data without a natural block, we are forced to specify a block size. This has the effect of reducing our sample size by a factor of 1/block size. In data with a natural block, this might be appropriate, but without a natural block this is extremely wasteful of data.

2.2 Univariate EVT - Thresholding

If F is in the domain of attraction of an EVD, then for a random variable X that follows F, exceedances over a large threshold u can be said to follow a Pareto distribution. Again, let X follow some distribution function F. Then let us regard observed values that exceed some threshold u as extreme. It follows that:

$$\Pr\left[X > u + y\right] = \frac{1 - F(u+y)}{1 - F(u)}$$

for $y > 0$. Let $X_1, \ldots, X_n$ be a sequence of random variables with the distribution function F. Let $M_n = \max[x_1, \ldots, x_n]$. Suppose that F is in the domain of attraction of the GEV, such that for large n, $\Pr[M_n \leq z] \approx G(z)$. Then, for large enough u, $\Pr[X > u+y \mid x > u]$ approximates to

$$H(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-\frac{1}{\xi}}.$$

This defines the generalized Pareto family of distributions. Thus, if block maxima have a limiting distribution G within the EVD family, then threshold exceedances for a sufficiently high threshold have a limiting distribution H within the Generalized Pareto (GP) family. Furthermore, the extremal index χ will be the same for these two limiting distributions.

2.3 Multivariate EVT

Within multivariate EVT, we observe the joint behavior between extreme events. It is useful, at this juncture, to to standardize each variable X_i according to its marginal distribution. Note, as has been stated, that threshold exceedances for a high threshold have a limiting distribution in the GP family. Therefore, we estimate parameters for those marginal distributions considering only those observations that exceed the threshold. Standardization occurs as:

$$z_j = \left(1 + \xi \frac{x_j - b_{t,j}}{a_{t,j}}\right)_+^{1/\xi} \tag{1}$$

Note that $Z_j > 1$ implies that $x_j > b_{t,j}$, meaning that the observation x is extreme in the j'th dimension. Note also that $\sup_j Z_j$ follows a simple Pareto distribution.

2.3.1 Limit Measures

We assume the existence of a probability measure μ on $\mathbf{Z}$ such that

$$\lim_{n\to\infty} n\Pr\left[\frac{1}{n}\mathbf{Z} \geq \mathbf{z}\right] = \mu\left([\mathbf{0}, \mathbf{z}]^c\right) \tag{2}$$

μ is thus the asymptotic distribution of $\mathbf{Z}$ in extreme regions. It exhibits the homogeneity property, such that for any region A, $\mu(rA) = r^{-1}\mu(A)$. By this property we can thus factorize $\mathbf{Z}$ into two components:

$$\begin{aligned} R &= \|\mathbf{Z}\|_\infty \in [1, \infty), \\ \mathbf{V} &= \frac{\mathbf{Z}}{R} \in S_\infty^{d-1}. \end{aligned} \tag{3}$$

That is, to say, we factorize $\mathbf{Z}$ into a radial component R, and an angular component, $\mathbf{V}$, which is the projection of $\mathbf{Z}$ onto the positive orthant of the d-dimensional unit hypersphere defined by the infinity norm, S_∞^{d-1}. The radial component R, as we have stated, follows a simple pareto distribution, and by homogeneity property, is independent of the angular component $\mathbf{V}$.

As the angular component is now independent of the radial component, we can establish a distribution on the angular component. For $B \subset S_\infty^{d-1}$, We define the spectral (or angular) measure, $\Omega(B)$, as

$$\Omega(B) = \mu[\mathbf{z} : R(\mathbf{z}) > 1, \mathbf{V} \in B]. \tag{4}$$

Then we can think of the spectral measure in terms of the limit measure μ, and:

$$\mu[\mathbf{z} : R(z) > t, \mathbf{V} \in B] = t^{-1}\Omega(B). \tag{5}$$

Thus we see a one-to-one correspondance between the limit measure μ and the spectral measure ϕ, and by factoring out the Pareto distributed radial component, we can establish a distribution on the angular component

$$\Pr\left(\mathbf{V} \in B \mid r > 1\right) = \frac{\Omega(B)}{\Omega(S_\infty^{d-1})}, \tag{6}$$

Paraphrasing from Goix et. al., figure if/how to cite conditioned on at least one of the components being extreme in the marginal sense. It is using this property that we will establish our method.

For each $\nu \subset \{1, \ldots, d, \nu \neq \emptyset$, we define the *truncated cone* $\mathcal{C}_\nu$, where

$$\mathcal{C}_\nu = \left\{\mathbf{z} \geq 0 : \|z\|_\infty \geq 1, z_j \geq 0 \forall j \in \nu, z_j = 0 \forall j \notin \nu\right\}. \tag{7}$$

That is, ν identifies a set index specifying components of the standardized data for which the observation is greater than 0. The observation is greater than 0 for columns within that index, and 0 outside that index. By construction, we're also requiring that the observation is greater than 1 in at least one of those dimensions. By this definition, we observe that each $\mathcal{C}_\nu$ is distinct and disjoint from any other $\mathcal{C}_\nu$. Now, defining Ω_ν as the projection of C_ν onto S_∞^{d-1},

$$\Omega_\nu = \left\{\mathbf{v} \in S_\infty^{d-1} : x_i > 0 \forall i \in \nu, x_i = 0 \forall i \notin \nu\right\}, \tag{8}$$

we can clearly see $\mu(\mathcal{C}_\nu) = \Phi(\Omega_\nu)$ for all $\alpha \subset \{1, \ldots, d\}$. Where we call $\mu(\dot{})$ the limit measure, we refer to $\Phi(\dot{})$ as the *spectral* or *angular measure*. Our goal is establishing statistical inference on this angular measure.

3 Review of State of the Art

3.1 Extreme Value Theory

3.1.1 Asymptotic Dependence

A simple measure of asymptotic dependence between two variables sharing the same marginal distribution is the coefficient of asymptotic dependence, χ

$$\chi_{ij} = \lim_{z \to \infty} \Pr(Z_i > z \mid Z_j > z) \tag{9}$$

[6] defines the multivariate Generalized Pareto distribution, which is expanded on in [1], introducing the multivariate Generalized Pareto process.need to go through rootzen 2018

3.1.2 Generalized Pareto Process

3.2 Anomaly Detection

Anomaly detection is a broadly defined term, but in this context we choose it to mean finding observations that are *different* in some capacity from the rest of the observations in the data. The lion's share of anomaly detection algorithms can be summarized into two main categories: distance based methods, and density based methods.

3.3 Distance based methods

Distance based methods work with some notion of distance to define how unique an observation is by how far it is from other observations in the data. A simple illustrative example might be minimum distance to neighbor, for each observation. The algorithm would proceed as follows: first scale the data, then calculate pairwise distances between each observation. Those observations with the highest minimum distance to a neighbor are the most unique, and therefore seem the most anomalous. Insert summaries and citations of distance methods This basic idea is extended in several ways through clustering methods insert citation of k-means, decision treesinsert citation of isolation forests, and so on. A basic assumption required we can gather from this example is that we expect non-anomalous data to behave in a consistent manner, and anomalous data to behave in unique and different manners. These methods generally make little to no assumptions regarding the underlying distribution of the data.

isolation forests, DBSCAN,

3.4 Density based methods

complete rewrite of this paragraph. it sucked.

Local outlier factor–both camps

3.4.1 Observation Density

Under this approach, an illustrative example would be to look at the contribution to the marginal likelihood for each observation, and identify as anomalies those observations which are least likely. expand

something something posterior probability of observation, lowest contribution to log-likelihood, etc.

The defining characteristic of this category is we model all the data, and don't model anomalous data as having come from a separate distribution. Then, we look at how likely is that observation. The observations that are least likely given the model are considered anomalous. We might consider studentized residuals from linear regression as being something akin to this.

Our method falls somewhat into the density based method idea. We are asserting a parametric model upon the data, but our method requires one additional assumption as

to what it means to be anomalous: we assert that anomalies must be *extreme* in at least one margin.

4 The Problem

5 Methodology

Assuming that X_j is in the domain of attraction of a max stable distribution for each j, then the standardization of X to GPD process occurs as follows:

$$Z_j = \left(1 + \gamma_j \frac{X_j - b_{tj}}{a_{tj}}\right)_+^{1/\gamma_j} \tag{10}$$

where $b_{tj} = F^{-1}(1 - 1/t)$, and a_{tj} and γ_j are evaluated via maximum likelihood (need to fix). The quantity $\left(1 + \gamma_j \frac{X_j - b_{tj}}{a_{tj}}\right)_+$ is left truncated at 0. Note that $Z_j > 1$ indicates that $X_j > b_{tj}$, meaning that the observation was *extreme* in that dimension. Recognize here that Z_i exists on the positive orthant in Euclidean space, $\mathcal{R}_+^d$, and that $\max_j Z_j$ follows a simple Pareto distribution. Additionally, we can transform $\mathbf{Z} \to (R, \mathbf{V})$, where:

$$\begin{aligned} R &= \|\mathbf{z}\|_\infty = \max_{j \in (1,\dots,d)} z_i \\ \mathbf{V} &= \left(\frac{z_1}{R}, \dots, \frac{z_d}{R}\right) \in S_\infty^{d-1}. \end{aligned} \tag{11}$$

Thus $\mathbf{V}$ is the projection of the $\mathbf{Z}$ vector onto the unit hypersphere defined by the L_∞ norm, alternatively called the unit hypercube. As stated before by homogeneity property, a spectral measure Φ on some space in S_∞^{d-1} is independent of R, which follows the standard Pareto distribution by construction.

We are interested in the angular distribution of $\mathbf{V}$, the projection of the standardized observations $\mathbf{Z}$ onto the unit hypercube. As we are unaware of any distribution that operates natively this space (the positive orthant of the unit hypercube), or can be effectively coerced into this space, we construct distributions on other spaces for which there exists a one to one mapping from the distribution space and target space. That means, the space S_∞^{d-1} can be projected onto some other S_p^{d-1} using another norm p. We project onto the unit simplex using the L_1 norm, and onto the unit hypersphere using the L_2 norm. There exists a one-to-one mapping between these spaces and the unit hypercube.

For model comparison, as the projection induces its own distortion, we will conduct model comparison in S_∞^{d-1}, using a scoring rule appropriate to that space.

5.1 Dirichlet

As a distribution defined on the unit hypersphere using the L_1 norm, or simplex, Dirichlet is a natural choice for our purpose. The dirichlet random variable can be decomposed as a vector of independent gamma random variables with a constant rate parameter, divided by their sum (or L_1 norm). That is,

$$\mathbf{x} \sim \text{Dir}(\mathbf{x} \mid \zeta) = \int_0^\infty \prod_{l=1}^d \text{Ga}(rx_l \mid zeta_l, 1)|J|\text{d}r \tag{12}$$

The value of the shared rate parameter is irrelevant to the distribution of $\mathbf{x}$, so by convention we set it to one. The Jacobian for this transformation is r^{d-1}.

We consider two mixture models under this family; a finite mixture model of fixed dimension, and an infinite mixture model using a Dirichlet process prior for ζ.

5.1.1 Finite Mixture of Dirichlets

The finite dirichlet mixture model, *MD*, attempts to represent the distribution of data, projected onto the simplex, using a finite mixture of Dirichlet parameters ζ. That is, for each mixture component j, we have a vector ζ_j detailing the parameters of the Dirichlet distribution under which observations under that component are distributed.

$$\begin{aligned}(r_i, x_i) \mid \delta_i &\sim \prod l = 1^d \mathrm{Ga}(rx_{il} \mid \zeta_{jl}, 1) \\ \zeta_{jl} \mid \alpha_l, \beta_l &\sim \mathrm{Ga}(\zeta_{jl} \mid \alpha_l, \beta_l) \\ \alpha_l &\sim \mathrm{Ga}(\alpha_l \mid 0.5, 0.5) \\ \beta_l &\sim \mathrm{Ga}(\beta_l \mid 2, 2) \\ \lambda &\sim \mathrm{Dir}(0.5)\end{aligned} \tag{13}$$

Let i denote indexing over the data set. Let j denote indexing over mixture components. Let Let l denote indexing over columns. Thus, δ_i denotes the mixture component associated with the i'th observation. ζ_{jl} denotes the shape parameter of the Dirichlet distribution associated with mixture component j, and column l. The purpose of the rate parameter hyperpriors being being somewhat informative, is to ensure for numerical stability reasons that the rate parameters do not approach 0.

We perform data augmentation, generating r_i, and recovering the original product of independent gammas interpretation of the Dirichlet RV. This enables us to do posterior learning about ζ_{jl_1} independent of ζ_{jl_2}. The augmented variable, r_i, can be generated as

$$r_i \mid x, \delta, \zeta \sim \mathrm{Ga}(r_i \mid \sum_{l=1}^{d} \zeta_{jl}, 1) \tag{14}$$

We assign a prior distribution for probability of mixture component membership, λ, as a Dirichlet RV with a relatively weak 0.5 symmetric shape parameter.

The full conditional distribution for ζ_{jl} is not available in a known form, so sampling will require some flavor of MCMC. We employ a Metropolis-Hastings sampler on $\log \zeta_{jl}$ using a normal proposal distribution with a standard deviation of 0.3. This is also employed in the posterior sampling for α_l. The full conditional for β_l arrives in known form as a Gamma,

$$\beta_l \mid \alpha_l, \zeta \sim \mathrm{Ga}(\beta_l \mid J\alpha + 2, \sum_{j=1}^{J} \zeta_{jl} + 2) \tag{15}$$

The full conditionals for ζ_{jl} and α_l are yet to be inserted, but they are simple gamma/gamma models.

We construct the finite mixture again using data augmentation, introducing the mixture component identifier δ_i. The posterior probability that $\delta_i = j$ is constructed as:

$$p(\delta_i = j \mid r, x, \pi, \zeta) \propto \pi_j \prod_{l=1}^{d} \mathrm{Ga}(rx_{il} \mid \zeta_{jl}, 1) \tag{16}$$

Then the posterior distribution for π is formed from the cluster membership identifiers. Let $n_j = \sum 1_{\delta_i = j}$, then π is distributed as

$$\pi \mid \delta \sim \text{Dir}(n_1 + 0.5, \ldots, n_J + 0.5). \tag{17}$$

This comprises the simplest model we present for comparison.

5.1.2 Dirichlet Process Mixture of Dirichlets

The natural extension to the finite mixture of Dirichlets would be to assume the cluster parameters, ζ_j, as descending from an infinite mixture. As we are simply attempting to represent the data, and do not have a compelling interest in controlling the number of clusters, a natural choice of prior is the Dirichlet process. We denote this model as *DPD*. This model will share a great deal of construction, posterior inference, and indeed code, with the finite mixture model.

$$\begin{aligned}
(r_i, \mathbf{x}_i) \mid \zeta_i &\sim \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta_{il}, 1) \\
\zeta_i &\sim \text{DP}(\eta, G) & G &= \prod_{l=1}^{d} \text{Ga}(\zeta_{il} \mid \alpha_l, \beta_l) \\
\alpha_l &\sim \text{Ga}(\alpha_l \mid 0.5, 0.5) \\
\beta_l &\sim \text{Ga}(\beta_l \mid 2, 2) \\
\eta &\sim \text{Ga}(\eta \mid 2, \kappa) & \kappa &\in \{0.1, 1, 10\}
\end{aligned} \tag{18}$$

We denote cluster membership using δ_i, as in the previous case. Let $n_j = \sum 1_{\delta_i = j}$ denote the cluster size. In the DP literature terminology, we are using what is referred to as the collapsed sampler, where for existing clusters,

$$p(\delta_i = j \mid r, \zeta, \eta) \propto \frac{n_j}{\sum_j n_j + \eta} \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta_{jl}). \tag{19}$$

For new clusters, ostensibly we would integrate out the cluster parameters to get a true posterior predictive density. However, doing so in this case is not straitforward. Instead, we employ algorithm 8 from [4], which introduces another parameter m. We generate m new candidate clusters given α, β, then the probability of x_i belonging to any particular cluster from the candidate clusters is given as:

$$p(\delta_i = j \mid r, \zeta', \eta) \propto \frac{\eta/m}{\sum_j n_j + \eta} \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta'_{jl}) \tag{20}$$

where ζ'_j indicates the cluster parameters from a candidate cluster. If a new cluster is selected, then we append the new cluster parameters to the stack, and continue to the next observation.

This model is slightly more complex than the finite mixture of Dirichlets, but at its core it employs the same assumption–that the individual columns descend from independent gamma random variables, with a fixed rate parameter.

5.1.3 Dirichlets with log-normal Prior

A derivative model we might try is placing a lognormal prior on ζ. That is, to say, instead of having ζ descend from d independent Gamma random variables, we can have ζ instead descend from a d-dimensional log-normal random variable. The impetus for this variation comes from our implementation of the DP mixture model–that we need to generate candidate clusters for the shape parameters. In the previous model, we are generating these clusters by independently drawing from d gamma distributions. Some information may be available in terms of covariance between dimensions, so a log-normal prior on ζ may serve to capture that information between dimensions. That is, for the finite mixture model,

$$\begin{aligned}(r_i, x_i) \mid \delta_i &\sim \prod_{l=1}^{d} \mathrm{Ga}(rx_{il} \mid \zeta_{jl}, 1) \\ \zeta_j \mid \mu, \Sigma &\sim \mathcal{LN}_d\left(\zeta_j \mid \mu, \Sigma\right) \\ \mu &\sim \mathcal{N}_d\left(\mu \mid \mu_0, \Sigma_0\right) \\ \Sigma &\sim \mathrm{IW}\left(\Sigma \mid \nu, \Psi\right) \\ \lambda &\sim \mathrm{Dir}(0.5)\end{aligned} \tag{21}$$

The hope is this will allow us to better capture the relationships between dimensions, and thus more efficiently generate candidate clusters for the DP sampler. The downside, however, is that this introduces a d-dimensional normal distribution into the model, and for that we suffer the the computational complexity that induces.

The DP equivalent of this model has ζ_i as descending from a Dirichlet Process, with a log-normal kernel distribution. We place a gamma prior on the concentration parameter η, and thereafter the hyperpriors are the same as for the finite mixture model.

5.2 Projected Gamma

Another $d-1$ dimension reduction we can use is instead of projecting onto the unit simplex, S_1^{d-1}, we can project onto the unit hypersphere formed on the Euclidean norm, S_2^{d-1}. [5] develops this idea fully into the projected gamma distribution. Again, we form the distribution as the product of d independent gammas. That is, $\mathbf{y} = (y_1, \ldots, y_d)^t$, and $y_i \sim \mathrm{Ga}(\alpha_i, \beta_i)$. We define our starting point:

$$f(\mathbf{y} \mid \alpha, \beta) = \prod_{j=1}^{d} \mathrm{Ga}(y_j \mid \alpha_j, \beta_j), \tag{22}$$

where β is specified as a rate parameter. [5] proceeds through a full spherical coordinate transformation, where $\theta_i = \cos^{-1}(y_i / \|y_{i:d})$, for $i \in \{1, \ldots, d-1$. Then $y_i = r \prod_{j=1}^{i-1} \sin\theta_j \cos\theta_i$. This results in a true $d-1$ dimensional distribution, with $\theta_i \in [0, \pi/2)$ for all $i \in \{1, \ldots, d-1$.

d-dimensional spherical coordinates $\mathbf{y} \to (r, \theta)$ as

$$\begin{aligned} y_1 &= r\cos\theta_1, \\ y_2 &= r\sin\theta_1\cos\theta_2 \\ &\vdots \\ y_{d-1} &= r\sin\theta_1 \ldots \sin\theta_{d-2} \\ y_d &= r\sin\theta_1 \ldots \sin\theta_{d-1} \end{aligned} \tag{23}$$

where $r = \|\mathbf{y}\|_2$, the euclidean norm of $\mathbf{y}$. The inverse of this transformation is:

$$\begin{aligned} \theta_1 &= \cos^{-1}\left[\frac{y_1}{\|y_{1:d}\|_2}\right] \\ \theta_2 &= \cos^{-1}\left[\frac{y_2}{\|y_{2:d}\|_2}\right] \\ &\vdots \\ \theta_{d-1} &= \cos^{-1}\left[\frac{y_{d-1}}{\|y_{(d-1):d}\|_2}\right]. \end{aligned} \tag{24}$$

The Jacobian of this transformation is

$$r^{d-1}\prod_{i=1}^{d-2}(\sin\theta_i)^{d-1-i}.$$

This creates the distribution over r, θ. The full conditional for r takes the form of a Gamma random variable, and we can integrate it out as such. This leaves the *projected gamma distribution*,

$$\mathrm{PG}(\theta \mid \alpha, \beta) = \frac{\Gamma(A)\beta_d^{\alpha_d}}{B^A\Gamma(a_d}\left(\prod_{j=1}^{d-1}\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}(\cos\theta_j)^{\alpha_j-1}(\sin\theta_j)^{(\sum_{h=j+1}^{d}\alpha_h)-1}\right)\mathcal{I}_{(0,\pi/2)^{d-1}}(\theta) \tag{25}$$

where

$$A = \sum_{j=1}^{d}\alpha_j \qquad \text{and} \qquad B = \beta_1\cos\theta_1 + \sum_{j=2}^{d-1}\left(\beta_j\cos\theta_j\prod_{i=1}^{j-1}\sin\theta_i\right) + \beta_d\prod_{j=1}^{d}-1\sin\theta_j. \tag{26}$$

As is, this model is not identifiable, as taking $\beta^{(2)} = \alpha\beta^{(1)}$ will still yield the same distribution of angles. Following [5], we have opted to place a restriction on β such that $\beta_1 := 1$, thus $\beta = (1, \beta_2, \ldots, \beta_d)^t$.

Inference on this model can take two forms: α and β in this form can not be broken down into known-form full conditionals, so we can conduct a Metropolis Hastings step for every component, or do a joint proposal Metropolis Hastings step for all components at once. Alternatively, using $f(r, \theta)$, we recognize that $\alpha_i \mid r$ is independent of $\alpha_j \mid r$, so we can sample the latent r and conduct independent Gibbs steps for each component. Further, in sampling the α_j's, we can integrate out β_j. Within the Gibbs sampler, we

sample r, then each $\alpha_j \mid r$, then each $\beta_j \mid r, \alpha_j$. This leads to fast convergence, with the only Metropolis Hastings step being for the α_j's. Both r and the β_j's are Gamma distributed.

For simplicity, let $\mathbf{y}' = r^{-1}\mathbf{y}$. That is, $\mathbf{y}'$ is a function of the angular data–from (23), $\mathbf{y}' = \mathbf{y}/r$, the projection of the $\mathbf{y}$ vector onto the unit hypersphere. We generate a latent r, and their product is the latent $\mathbf{y}$. Given $\mathbf{y}$, the posterior distributions for (α_i, β_i), (α_j, β_j), $i \neq j$ are independent.

As [5] shows, the projected gamma distribution is a flexible model for representing data on the positive orthant of the unit hypersphere. As such, given our application restricts us to this domain, one can see that this might be a natural choice of distribution for our purpose.

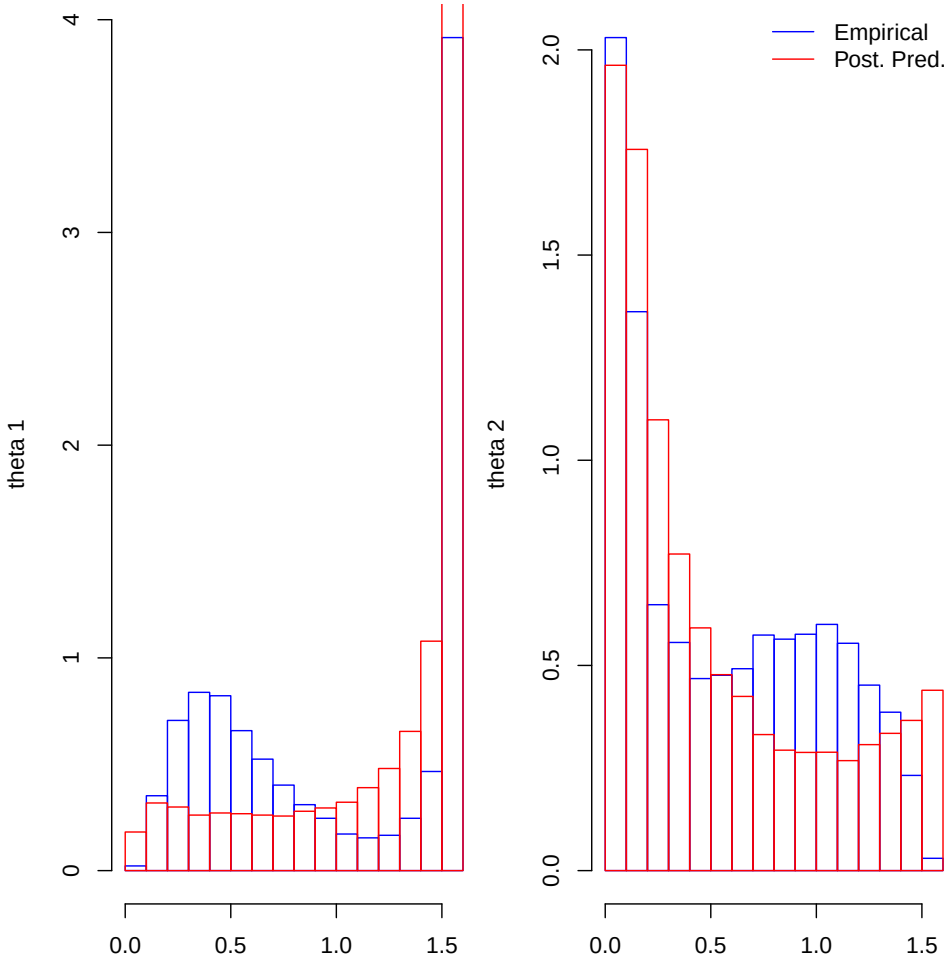

Figure 1: Histograms of Empirical vs Posterior-predictive angular data originating from a simulated 3-dimensional gamma dataset.

However, as flexible as it is, it alone is not sufficient for our purpose. Supposing

a given dataset is the result of two or more generating distributions, then using a a single distribution to represent this dataset becomes untenable. In Figure 5.2 we see the empirical distribution a 2-component mixture of projected gammas, plotted against the posterior predictive distribution of a projected gamma model fitted to that dataset. As we can see, it has trouble representing the nuances of the two component mixture.

5.3 Model Comparison on the Hypercube

It is not immediately obvious which criteria to use to judge these models and decide which best represents the data's generating distribution. We have opted to use the *posterior predictive loss* criterion of [2] and the *energy score* criterion of [3]. Both of these metrics require calculating some distance in the target space, and this section will be devoted to that end.

5.3.1 Posterior Predictive Loss Criterion

The posterior predictive loss criterion, *ppl* is introduced in [2]. When we assume a squared error loss function, then for the lth observation, the posterior predictive loss criterion is computed as

$$D_k = \mathrm{Var}(X_l) + \frac{k}{k+1}\left(\mathrm{E}[X_l] - \mathbf{x}_l\right)^2, \tag{27}$$

where X_l is a random variable from the posterior predictive distribution for x_l. The scalar k is a weighting factor by which we arbitrarily scale the importance of goodness of fit relative to precision. In our analysis, we take the limit as $k \to \infty$, and thus weight both parts equally. Interpreting this criterion, a smaller $\mathrm{Var}(\mathbf{X}_l)$ indicates a higher precision, and a smaller $(\mathrm{E}[\mathbf{X}_l] - \mathbf{x}_l)^2$ indicates a better fit. Thus, smaller is better. Note that this is defined for a univariate distribution–we are going to generalize this somewhat, to account both for the multivariate nature of our distribution, and its constrained geometry. That is, let us re-define it as

$$D'_k = \mathrm{E}\,\|\mathbf{X_l}, \mathrm{E}[\mathbf{X}_l]\|^2_\Omega + \frac{k}{k+1}\,\|\mathrm{E}[\mathbf{X}_l], \mathbf{x}_l\|^2_\Omega\,. \tag{28}$$

This generalizes the posterior predictive variance as expected squared distance from a central mean. We can numerically calculate this mean as Not sure of this. googling yields this as the result for median...at least in 1 dimension. not sure how to calculate for mean.

$$\mathrm{E}[\mathbf{X}_l] = \mathrm{argmin}_{\omega\in\Omega}\mathrm{E}\|\mathbf{X}_l, s\|_\Omega.$$

This is numerically difficult. Adjusting our interpretation somewhat, we can also project our replicates of $\mathbf{X}_l$ onto the simplex, calculate the mean on the simplex, then re-project that point onto the hypercube.

5.3.2 Energy Score

The energy score of [3] is a generalization of the continuous ranked probability score, or *crps*, defined for a multi-dimensional random variable.

$$\mathrm{ES}(P, x) = \frac{1}{2}\mathrm{E}_p\|\mathbf{X}_l, \mathbf{X}'_l\|^\beta_\Omega - \mathrm{E}_p\|\mathbf{X}_l - \mathbf{x}_l\|^\beta_\Omega \tag{29}$$

where $\mathbf{X}'_l$ is another replicate from the posterior predictive distribution of $\mathbf{x}_l$. This means, rather than relying on the first and second moments of the posterior predictive distribution as in the case of posterior predictive loss, we are instead calculating pairwise distances between the observation and draws from the posterior predictive distribution, as well as pairwise distances between those replicates themselves.

Now here's the rub. We are not aware of any standardized distance metrics developed on the positive orthant of the unit hypercube. In the unit simplex, we can assume the use of Euclidean norm. on the unit hypersphere, our task would be slightly more difficult as Euclidean norm would under-report the actual distance required for travel between points a_1 and a_2. On the unit hypercube, where our task is defined, the distortion between Euclidean norm and the actual distance travelled will be even greater.

The positive orthant of the unit hypercube, defined in Euclidean geometry, is that structure for which, in a given point on the hypercube, all dimensions of that point are between 0 and 1, and at least one dimension must be 1. Developing terminology, we can consider observations for which the jth dimension is equal to 1, to be on the jth *face*. The intersection of the ith and jth face is a hypercube with $d-2$ degrees of freedom, and observations in this space have dimensions i, j equal to 1.

A *distance* in this space corresponds to a *geodesic* on this space. From geometry, we know that the geodesic, or shortest path between 2 points along the surface of a d dimensional figure corresponds to at least one *unfolding*, or *rotation* of the d-dimensional figure into a $d-1$ dimensional space. The appropriate term for the structure generated by this unfolding is a *net*. For the appropriate net, a line segment connecting the two points and staying within the boundaries of the net corresponds to the shortest path between those points needs citation!, and is thus a geodesic. The length of that line segment is properly defines the distance required for travel between those points.

Consider a 3-dimensional cube. Consider 2 points on this 3-dimensional cube, $\mathbf{a}_1 = (x_1, y_1, z_1)$, and $\mathbf{a}_2 = (x_2, y_2, z_2)$. Let's say that the two points are on the same face. Then the distance between those two points, the distance one has to travel along the space to move from one point to the other, is calculated by Euclidean norm. Now, consider two points on separate faces. All faces are pairwise adjacent, as we have stated, so in order to move to the other point, we must *at least* move to the intersection between the faces, then to the other point. Let $\mathbf{a}_1$ lie on the x face, and $\mathbf{a}_2$ lie on the y face. That is, $\mathbf{a}_1 = (1, y_1, z_1)$, $\mathbf{a}_2 = (x_2, 1, z_2)$. Then traveling between these points we must at least pass through the intersection of faces x, y. One possible net representation of this is unfolding the y face alongside the x face. We accomplish this by applying a rotation and translation to a_2, corresponding to the following:

$$a'_2 = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ 1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - x_2 \\ z_2 \end{bmatrix} \tag{30}$$

Then, if this is the appropriate net, the distance becomes $\|\mathbf{a}_1, \mathbf{a}_2\| = \|\mathbf{a}_1 - \mathbf{a}'_2\|_2$ However, there is another possible net we must consider, travelling first through the z face then to the y face. The rotation for that becomes:

$$a'_2 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_2 \\ 1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 - z_2 \\ 2 - x_2 \end{bmatrix} \tag{31}$$

Every successive rotation is relative to the last face. So, as the number of dimensions grows, the number of possible rotations grows as well.

need to finish this!

As we have d faces, if 2 observations are on different faces, then there are $\sum_{j=1}^{d-2} \binom{(}{d} - 2, j) + 1$ possible rotations to consider. There are truly $d!$ possible nets, but when we consider starting and ending faces fixed, and that portions of the net that diverge after the ending face are irrelevant, we arrive at that number of rotations that we actually need consider. While this is not insurmountable, it is numerically difficult, and developing the generalized rotation strategies for d dimensions is beyond the scope of this analysis. However, we are in luck in that all we need for a valid energy score is a *negative definite kernel*. This is defined as a function having symmetry in its arguments, $d(x_1, x_2) = d(x_2, x_1)$, and for which $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j d(x_1, x_2) \leq 0$ for all positive integers n, with the restriction that $\sum_{i=1}^{n} a_i = 0$. The Euclidean norm is one example of a negative definite kernel.

Let's go back to the first rotation–we held that as a numerically easier analogue to what we were actually calculating–the distance from the starting point, to some optimal point along the intersection between the starting and ending faces, to the ending point. At that optimal point, the total distance travelled between starting and ending points becomes symmetric. We can recognize this distance as the sum of two Euclidean norms. That is,

$$\|\mathbf{a}, \mathbf{b}\|_H = \|\mathbf{a}, \mathbf{c}\|_2 + \|\mathbf{c}, \mathbf{b}\|_2 \tag{32}$$

If we can be assured of symmetry in the functional arguments, then the requirements for a negative definite kernel are trivially proved.

5.3.3 Intrinsic Energy Score

While we have at length discussed our means of comparison between models, we have limited means of comparing how our model is doing relative to the data. That is, to say, we lack a metric like R^2, the coefficient of variation from linear regression, forcing us to construct something to serve along these lines.

We introduce the *intrinsic energy score* to offer a baseline energy score *in the data*, against which we might compare candidate models. This allows us to see how well our model is doing relative to the data, rather than just relative to other models. We construct this using the energy score metric, but for any particular observation in the data, we compare that observation against all other observations in the data. That is, for a given observation, treat other observations as replicates of that observation.

5.4 Spatial Threshold Modeling

Following the work of [1], any of the above methods can be extended to the spatial domain by modification of the marginalization process. Assume a spatial process $X(\mathbf{s})$, $\mathbf{s} \in \mathbf{S}$. Then take the transformation

$$Z(\mathbf{s}) = \left(1 + \gamma(\mathbf{s}) \frac{X(\mathbf{s}) - b_t(\mathbf{s})}{a_t(\mathbf{s})}\right)_+^{1/\gamma(\mathbf{s})} \tag{33}$$

where $b_t(\mathbf{s})$ is a function corresponding to a high threshold at location $\mathbf{s}$, analogous to the role b_t played previously. $a_t(\mathbf{s})$ corresponds to a scaling function, and $\gamma(\mathbf{s})$ an extremal value index function.

6 Results

type	name	PPL_L1	PPL_L2	PPL_Linf	ES_Linf
md	results_10	23.143	83.292	190.927	0.278
md	results_20	14.961	54.877	132.154	0.213
md	results_30	11.801	44.250	108.827	0.190
dpd	results_2_1e-1	11.444	41.287	101.031	0.171
dpd	results_2_1e0	11.451	41.284	100.829	0.172
dpd	results_2_1e1	11.864	44.340	108.385	0.186
mgd	results_10	67.109	188.681	344.121	0.405
mgd	results_20	59.770	167.675	310.530	0.373
mgd	results_30	48.180	140.479	270.322	0.352
dpgd	results_2_1e-1	80.187	233.552	426.893	0.440
dpgd	results_2_1e0	76.133	217.615	397.436	0.412
dpgd	results_2_1e1	77.313	221.316	404.790	0.410
mprg	results_10	31.942	98.863	209.722	0.292
mprg	results_20	15.207	56.343	135.617	0.221
mprg	results_30	12.261	44.676	108.121	0.186
dpprg	results_2_1e-1	10.432	40.166	100.086	0.170
dpprg	results_2_1e0	10.318	39.239	97.956	0.169
dpprg	results_2_1e1	16.848	53.743	122.638	0.193
mpg	results_10	73.526	200.870	364.297	0.429
mpg	results_20	56.583	159.596	297.343	0.375
mpg	results_30	50.216	145.472	274.975	0.349
dppg	results_2_1e-1	78.643	223.581	406.923	0.425
dppg	results_2_1e0	76.558	228.945	422.949	0.433
dppg	results_2_1e1	84.384	243.056	441.553	0.456
mdln	results_10	43.429	126.159	254.331	0.322
mdln	results_20	16.544	58.340	138.082	0.221
mdln	results_30	12.977	47.172	113.800	0.191
dpdln	results_2_1e-1	12.923	45.969	110.086	0.184
dpdln	results_2_1e0	24.869	66.693	136.520	0.203
dpdln	results_2_1e1	12.754	46.683	113.024	0.190
dpgdln	results_2_1e-1	29.534	95.054	195.602	0.280
dpgdln	results_2_1e0	33.053	100.624	202.264	0.283
dpgdln	results_2_1e1	41.390	117.498	226.499	0.299
mprgln	results_10	34.192	108.854	233.089	0.294
mprgln	results_20	17.330	58.243	135.671	0.217
mprgln	results_30	13.366	47.179	112.964	0.189
dpprgln	results_2_1e-1	12.262	44.450	107.437	0.181
dpprgln	results_2_1e1	26.004	69.656	143.344	0.215
mpgln	results_10	95.587	308.977	572.978	0.472
mpgln	results_20	145.792	492.075	900.009	0.807
mpgln	results_30	53.790	152.978	289.603	0.350
dppgln	results_2_1e-1	33.675	105.547	213.656	0.288
dppgln	results_2_1e0	31.101	96.914	196.846	0.277
dppgln	results_2_1e1	39.658	112.152	215.999	0.296
dppn	results_2_1e-1	74.857	225.126	449.246	0.481
dppn	results_2_1e0	74.719	225.185	449.232	0.481
dppn	results_2_1e1	74.836	225.134	449.599	0.482

7 Conclusion

References

[1] Ana Ferreira and Laurens de Haan. The generalized pareto process; with a view towards application and simulation. *Bernoulli*, 20(4):1717–1737, 11 2014.

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

[3] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

[4] Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[5] Gabriel Núñez-Antonio and Emiliano Geneyro. A multivariate projected gamma model for directional data. *Communications in Statistics - Simulation and Computation*, pages 1–22, 05 2019.

[6] Holger Rootzén and Nader Tajvidi. Multivariate generalized pareto distributions. *Bernoulli*, 12(5):917–930, 2006.