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

Atmospheric rivers are temporary events, where large elongated regions of high concentrations of water vapor are developed in the atmosphere and carry huge amounts of water potentially thousands of miles. The amount of water in transit during these events dwarfs that of terrestrial rivers. For the targetted region, the atmospheric river can represent a significant portion of the precipitation the region will experience. Such events are thus of great interest to meterologists, as well as farmers, expand and cite.

One metric by which we might identify and declare atmospheric rivers is the integrated water vapor transport, or IVT. This value represents the total amount of water vapor being transported in an atmospheric column–that is, a column of the troposphere of particular size. These values can be measured by dropsondes, but the values we are using are estimated as part of a data productneeds citation. An observation from these data includes a reading (estimated or measured) at grid cell at a period in time; where a grid cell represents the surface of the earth associated with the atmospheric column. Our specific data comes from the coast of California, with daily readings of IVT covering 30 years, omitting leap days.

We have two such datasets, in differing spatial resolutions. The lower resolution splits the coast of california into 8 grid cells, while the higher resolution does so into 46 grid cells. We will be looking at relative model performance on these two datasets as a means of evaluating how well any model we propose scales. We are specifically interested in extremal dependence–the relationship between the upper tails of dimensions of the distribution. In this case, that means the relationship between extreme values in different grid cells. We are going to be looking at point-in-time behavior rather than considering the time series nature of the data–that relationship may come later.

As we are interested in the extremal dependence, it makes sense that we would choose to represent this data using lessons from extreme value theory. Extreme value theory, or 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[2], climatology[15], 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. Extreme value theory offers us a tool set for making inference about the tails of a distribution, without having observed said tails. For instance, with an extreme weather event like flooding, we can make predictions about return levels–the average time until an observation of a particular magnitude occurs–without having seen an observation of that magnitude, or having observed that long.

Is this part of the introduction? Anyway, something that is important to highlight is the fact that extreme value theory offers theoretical tool to make inference about the tails of a distribution without actually observing the tails. The use of asymptotic theory allows you to say things about the probability of events that happen every 100 years even if you only have 50 years worth of data.I tried to incorporate that

1.1 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. Maurice Fréchet[6] originates the field, identifying what would become known as the Fréchet and Weibull distributions. Wallodi Weibull, et al, [16] expands the analysis of the Weibull distribution; the results of that work giving it its current name. Fisher and Tippet[5] identify the Fréchet and Weibull distributions, along with a third form, as the three limiting forms of the distribution of the maxima of a sample. Emil Gumbel[9] offered an analysis of that third form, what is now known as the Gumbel distribution. Later works, including that of Arthur Jenkinson[10] reparameterize all three forms as special cases of a single unifying form, the generalized extreme value distribution, GEV:

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

As we specified earlier, distributions can be in the domain of attraction of one of the aforementioned extreme value distributions. If they are, regardless of which one, they will be in the domain of attraction of the GEV. Now, one characteristic aspect of max stable distributions is that they feature the homogeneity property,

$$\mu(tx) = \frac{1}{t}\mu(x).$$

Use of this property is endemic in EVT.

As this distribution specifies asymptotic behavior for the maximum of a set of observations, inference assuming this distribution requires we specify some block of data that we would take the maximum in, and report only that maximum. A series of these blocks yielding a series of maxima allows us to conduct inference about the parameters of the distribution. Taking only the maximum in a block of observations necessitates reducing our sample size by a factor of 1/block size. In some data where there might occur a natural block, such as an hourly time series where a natural block might be a day, this might

be appropriate. There is an implicit violation of the assumption of independence within a block, but that violation is generally ignored. In data without a natural block, this data reduction might be considered wasteful, as it limits our ability to conduct inference.

This is fine, but we need more formality. I am assuming that in the introduction, that is not written yet, your would have described some of the history of the field, and provided some basic references. Even so, you should provide some references here.I've tried to provide a history of the field, at least with respect to the relevant distributions. I'm sure I still need to provide some more recent history.

1.2 Univariate EVT - Thresholding

Another way we can approach the problem that is less wasteful of data, is to specify a threshold, and consider only those observations extreme that are in excess of the threshold. From the start, if F is in the domain of attraction of an EVD, then for a random variable $X \sim F$, 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:

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

for $y > 0$. If we consider the limit of the above as $u \to \infty$, then if F is in the domain of attraction of an EVD, then $\lim_{u\to\infty} \Pr\left[X > u + y \mid X > u\right]$ has a functional form–the survival function of a Pareto distribution. 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. Not clear where the extremal index comes from.I need to add some sources for Pareto distribution, as well as further explanation. minor explanation of the existence for various estimators beyond ML for chi as well.

One other point of note is that for time series, the implicit assumption of independence for those observations in excess of a threshold is violated. One means of dealing with this violation is to consider a string of observations in excess of a threshold as correlated, and only keep one of the string.Need citation of paper that uses this approach...as we do as well.

1.3 Multivariate EVT

At this point you need to say that in this work we will focus on the methods based on threshold exceedance in a multivariate setting. You should introduce the topic (although

you might have already done that in the introduction) indicating the difficulties that there exist in the generalization of a Pareto distribution in the multivariate case. Cite some papers, of course. You can then say that the traditional approach consists on focusing on the estimation of the joint tail behavior, separate from the marginal behavior.

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.

1.3.1 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 as a consequence of the 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

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

This needs t be sharpened. I don't see why $R(\mathbf{z}) > 1$ is needed. Also

$$\mu([0, 1/\boldsymbol{z}]^c) = \int_{\mathbb{S}_\infty^{d-1}} \left(\bigvee_j \theta_j z_j\right) d\Omega(\theta) \,,$$

which is the reason Ω is a spectral measure. 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

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

Again, this needs sharpening, I don't get the $r > 1$. What you have is that, for any value of r.

$$\lim_{r\to\infty} \Pr\left(\mathbf{V} \in B \mid R > r\right) = \frac{\Omega(B)}{\Omega(S_\infty^{d-1})}, \tag{7}$$

So, the point here is what we do in practice. We start by doing the univariate analysis for each component. Take $b_{t,j} = F^{-1}(1 - 1/t)$ as the threshold. So

$$\lim_{t\to\infty} \Pr\left(\mathbf{V} \in B \mid R_t > 1\right) = \frac{\Omega(B)}{\Omega(S_\infty^{d-1})}, \tag{8}$$

. What this means in practice is that we take high thresholds in each dimension, and then take the standardized vectors for which $R > 1$. 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

$$\mathcal{C}_\nu = \{\mathbf{z} \geq 0 : \|z\|_\infty \geq 1, z_j \geq 0 \forall j \in \nu, z_j = 0 \forall j \notin \nu\}. \tag{9}$$

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},

$$\Omega_\nu = \left\{\mathbf{v} \in S_\infty^{d-1} : x_i > 0 \forall i \in \nu, x_i = 0 \forall i \notin \nu\right\}, \tag{10}$$

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.

2 Review of State of the Art

2.1 Extreme Value Theory

2.1.1 Asymptotic Dependence

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

$$\chi_{ij} = \lim_{z\to\infty} \Pr(Z_i > z \mid Z_j > z) \tag{11}$$

[14] defines the multivariate Generalized Pareto distribution, which is expanded on in [4], introducing the multivariate Generalized Pareto process.need to go through rootzen 2018

2.1.2 Generalized Pareto Process

2.2 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.

2.3 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,

2.4 Density based methods

complete rewrite of this paragraph. it sucked.

Local outlier factor–both camps

2.4.1 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.

3 Methodology

You said this already 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:

$$Z_j = \left(1 + \gamma_j \frac{X_j - b_{tj}}{a_{tj}}\right)_+^{1/\gamma_j} \tag{12}$$

where $b_{tj} = F^{-1}(1 - 1/t)$, and a_{tj} and γ_j are evaluated via maximum likelihood. 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:

$$\begin{aligned} R &= \|\mathbf{z}\|_\infty = \max_{j \in (1,\dots,d)} z_i \\ \mathbf{V} &= \left(\frac{z_1}{R}, \dots, \frac{z_d}{R}\right) \in S_\infty^{d-1}. \end{aligned} \tag{13}$$

Thus $\mathbf{V}$ is the projection of the $\mathbf{Z}$ vector onto the unit hypersphere defined by the L_∞ norm, S_∞^{d-1}. 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 S_∞^{d-1}. 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 $\mathcal{L}_p$. 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, and owing to the difficulty of creating a density directly on S_∞^{d-1}, we will conduct model comparison using the posterior predictive distributions in S_∞^{d-1}, using a scoring rule appropriate to that space.

3.1 Projection onto an arbitrary unit hypersphere

A hypersphere is a geometric object such that the distance from any point to the center takes a fixed, constant value. The unit hypersphere is a hypersphere where that distance is 1. We can define the hypersphere under an arbitrary distance measurement, so let's take the $\mathcal{L}_p$ norm. Let the $\mathcal{L}_p$-norm be defined as

$$\|\mathbf{s}\|_p = \left(\sum_{l=1}^{d} |s_l|^p\right)^{\frac{1}{p}}.$$

From this, we establish the $\mathcal{L}_1$ norm as $p = 1$, or the absolute sum, equivalently called Manhattan distance; the $\mathcal{L}_2$ norm, as $p = 2$, the Euclidean distance. From this we also establish the $\mathcal{L}_\infty$ norm, as $\lim_{p \to \infty} \|\mathbf{s}\|_p = \max_{l \in \{1,\dots,d} s_l$.

We are interested in the direction, or angular distribution, of vectors described in the positive orthant, $\mathcal{R}_+^d$. As we are specifically interested in direction, we can project any distribution in $\mathcal{R}_+^d$ onto the positive orthant of the unit hypersphere in a $\mathcal{L}_p$-norm, denoted as $\mathcal{S}_p^{d-1}$. That is,

$$\mathcal{S}_p^{d-1} = \left\{ \mathbf{y} : \mathbf{y} \in \mathcal{R}_+^d, \|\mathbf{y}\|_p = 1 \right\}.$$

We can project an observation onto this space by dividing said observation by its p-norm. That is, let $\mathbf{x} \in \mathcal{R}_+^d$, then $\mathbf{y} = \mathbf{x}/\|\mathbf{x}\|_p \in \mathcal{S}_p^{d-1}$. We denote the $d-1$ to indicate the loss of one degree of freedom relative to the original vector.

So $\mathcal{S}_1^{d-1}$ defines the unit simplex, $\mathcal{S}_2^{d-1}$ defines the generalization of a circle–what we would generally refer to as a hypersphere, and $\mathcal{S}_\infty^{d-1}$ the surface of the hypercube. The hyperspheres defined by $\mathcal{L}_p$ as p varies have a one to one correspondance with eachother, meaning that observations on one can be projected onto another without loss of information.

Assuming $\mathbf{y} \in \mathcal{S}_p^{d-1}$, sometimes you use bold faces to denote vectors and sometimes you don't. You need to be consistent.then for finite p, y_d can always be represented as a function of the other dimensions. That is,

$$y_d = \left(1 - \sum_{l=1}^{d-1} y_l^p\right)^{\frac{1}{p}}.$$

So the transformation

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

does not lose any information. The reverse of this transformation,

$$T^{-1}(r, y_1, \ldots, y_{d-1}) = \left(ry_1, \ldots, ry_{d-1}, r\left(1 - \textstyle\sum_{l=1}^{d-1} y_l^p\right)^{\frac{1}{p}}\right)$$

equivalently recovers the original data. The determinant of the Jacobian of this transformation takes the form

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

There is something missing in this formula Notice a factor of r^{d-1} independent of p. We refer to $\boldsymbol{y}$ and r as, respectively, the angular and radial components of $\boldsymbol{x}$. If we assume a distribution for $\mathbf{x}$, then by transforming to $r, \mathbf{y}$ and integrating out r, we are left with a distribution on solely the angular component, or, equivalently, the projection of the vector $\mathbf{x}$ onto $\mathcal{S}_p^{d-1}$.

Many of the models we present here follow this form, where we, for reasons to be elaborated, assume a d-dimensional Gamma distribution on this hypothetical $\mathbf{x}$. For finite p, this has a direct benefit in that it is easy to integrate out r. As we saw with the Jacobian computed earlier, no matter what p, the Jacobian always has a factor of r^{d-1}.

With the independent Gamma model, r easily integrates out as a gamma distribution. We can also perform data augmentation generating latent r's, and recovering the ability to do independent inference on the parameters of those gamma distributions. We investigated other unidimensional distributions with support on $\mathcal{R}_+$ in the hopes we could perform the same dimension reduction with a different parameterization, but none offered the flexibility of the Gamma while allowing r to be integrated out in closed form.

One might question why we don't use this method to construct a distribution directly on $\mathcal{S}_\infty^{d-1}$, the unit hypersphere under $\mathcal{L}_\infty$. Put simply, we encounter a problem in the transformation. If we examine the the determinant of the Jacobian under the $\mathcal{L}_p$ norm, we have a factor along the lines of

$$\left(1 - \sum_{l=1}^{d-1} y_i^p\right)^{\frac{1}{p}-1}$$

which, if we take the limit as p approaches infinity, if any other y_l than y_d is equal to 1, then that value approaches 0^{-1}–an impossibility. We see a clear breaking point between inference conducted on the finite p hypersphere, $\mathcal{S}_p^{d-1}$, and the $\mathcal{L}_\infty$ hypersphere, $\mathcal{S}_\infty^{d-1}$. Another way we can recognize this issue is from the transformation itself: under $\mathcal{L}_\infty$, $T^{-1}(r, y_1, \ldots, y_{d-1})$ will not recover $x_1, \ldots, x_d$ if $y_d \neq 1$, or equivalently $x_d \neq \max_i x_i$.

With this in mind, the way one might build a distribution on $\mathcal{S}_\infty^{d-1}$ that still operates in Cartesian coordinates geometry might be to include an equal weighting mixture model, where each component of the mixture represents the probability of an observation being on that face, times the conditional density of the other dimensions given the face. That is,

$$f(y) = \sum_{l=1}^{d} p(y_l = 1) f(y_{-l} \mid y_l = 1)$$

Under this interpretation, we can consider $p(y_l = 1) = P(x_l = \max_i x_i)$. Unfortunately, this calculation is not straitforward.

Include arguments from stackexchange thread; cite accordingly You need to sharpen the last two paragraphs because I am having a hard time understanding what you are trying to say here. I've tried to make it a little more clear.

Alternatively, one can also map $\mathbf{y} \in \mathcal{S}_p^{d-1}$ into an alternative geometry, where we can express those $d-1$ degrees of freedom in a $d-1$ dimensional vector. On the $\mathcal{L}_1$ norm, one might consider isometric or additive logratios[1] as an appropriate geometry. On the $\mathcal{L}_2$ norm, we consider spherical coordinates. This directly maps S_2^{d-1} to $[0, \pi/2]^{d-1}$. [12] follows this course, starting with the independent Gamma distribution and still integrating out r to create an angular distribution in $[0, \pi/2]^{d-1}$. But we can also construct a distribution directly in this space. Along this idea, via probit transformation we map $[0, \pi/2]^{d-1}$ to $(-\infty, \infty)^{d-1}$, and construct a multivariate normal distribution in this geometry.

3.2 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,

$$\mathbf{y} \sim \text{Dir}(\mathbf{y} \mid \zeta) = \int_0^\infty \prod_{l=1}^{d} \text{Ga}(ry_l \mid zeta_l, 1)|J|\text{d}r \tag{14}$$

I don't think I understand this, basically because I don't understand the meaning of $Ga(ry_l|...)$. 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}. Can you give a better explanation of what the idea behind the Dirichlet? I suppose you are mapping S_∞^{d-1} into S_1^{d-1}, right?

We consider two mixture model approaches under this family; a finite mixture model of fixed dimension, and an infinite mixture model using a Dirichlet process prior for ζ. We For comparison, we also consider a vanilla Dirichlet model.

3.2.1 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.

$$\begin{aligned}
(r_i, x_i) \mid \delta_i &\sim \prod l = 1^d \text{Ga}(rx_{il} \mid \zeta_{jl}, 1) \\
\zeta_{jl} \mid \alpha_l, \beta_l &\sim \text{Ga}(\zeta_{jl} \mid \alpha_l, \beta_l) \\
\alpha_l &\sim \text{Ga}(\alpha_l \mid 0.5, 0.5) \\
\beta_l &\sim \text{Ga}(\beta_l \mid 2, 2) \\
\lambda &\sim \text{Dir}(0.5)
\end{aligned} \tag{15}$$

Let i denote indexing over the data set, j denote indexing over mixture components, and 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_{j1} independent of ζ_{j2} within the full conditional. The augmented variable, r_i, can be generated as

$$r_i \mid x, \delta, \zeta \sim \text{Ga}(r_i \mid \sum_{l=1}^{d} \zeta_{jl}, 1) \tag{16}$$

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,

$$\beta_l \mid \alpha_l, \zeta \sim \text{Ga}(\beta_l \mid J\alpha + 2, \sum_{j=1}^{J} \zeta_{jl} + 2) \tag{17}$$

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:

$$p(\delta_i = j \mid r, x, \pi, \zeta) \propto \pi_j \prod_{l=1}^{d} \text{Ga}(r x_{il} \mid \zeta_{jl}, 1) \tag{18}$$

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

$$\pi \mid \delta \sim \text{Dir}(n_1 + 0.5, \ldots, n_J + 0.5). \tag{19}$$

This comprises the simplest model we present for comparison.

3.2.2 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.

$$\begin{aligned}
(r_i, \mathbf{x}_i) \mid \zeta_i &\sim \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta_{il}, 1) \\
\zeta_i &\sim \text{DP}(\eta, G) & G &= \prod_{l=1}^{d} \text{Ga}(\zeta_{il} \mid \alpha_l, \beta_l) \\
\alpha_l &\sim \text{Ga}(\alpha_l \mid 0.5, 0.5) \\
\beta_l &\sim \text{Ga}(\beta_l \mid 2, 2) \\
\eta &\sim \text{Ga}(\eta \mid 2, \kappa) & \kappa &\in \{0.1, 1, 10\}
\end{aligned} \tag{20}$$

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,

$$p(\delta_i = j \mid r, \zeta, \eta) \propto \frac{n_j}{\sum_j n_j + \eta} \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta_{jl}). \tag{21}$$

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 [11], which, instead of evaluating the posterior predictive

density as a single point, we generate m new candidate clusters given α, β, then the probability of x_i belonging to any particular candidate cluster is given as:

$$p(\delta_i = j \mid r, \zeta', \eta) \propto \frac{\eta/m}{\sum_j n_j + \eta} \prod_{l=1}^{d} \text{Ga}(r_i x_{il} \mid \zeta'_{jl}) \tag{22}$$

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.

3.2.3 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 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 in terms of covariance between dimensions may be available, and would be lost assuming an independent gamma prior, so a log-normal prior on ζ may serve to capture that information. That is, for the finite mixture model,

$$\begin{aligned}
(r_i, x_i) \mid \delta_i &\sim \prod_{l=1}^{d} \text{Ga}(r x_{il} \mid \zeta_{jl}, 1) \\
\zeta_j \mid \mu, \Sigma &\sim \mathcal{LN}_d\left(\zeta_j \mid \mu, \Sigma\right) \\
\mu &\sim \mathcal{N}_d\left(\mu \mid \mu_0, \Sigma_0\right) \\
\Sigma &\sim \text{IW}\left(\Sigma \mid \nu, \Psi\right) \\
\lambda &\sim \text{Dir}(0.5)
\end{aligned} \tag{23}$$

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.

The three models that you consider in this section are essentially the same model with three variations. You should write the section by presenting a model that consists of a mixture of Dirichlet with two options: a finite mixture and a DP mixture, the a submodel for the DP misture where you consider two choices fot he hyperparameter priors. Now, is it really worth presenting all three different models?

3.3 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}. [12] 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 $y_i \sim \text{Ga}(\alpha_i, \beta_i)$. We define our starting point:

$$f(\mathbf{y} \mid \alpha, \beta) = \prod_{j=1}^{d} \text{Ga}(y_j \mid \alpha_j, \beta_j), \tag{24}$$

where β is specified as a rate parameter. [12] 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

$$\begin{aligned} y_1 &= r\cos\theta_1, \\ y_2 &= r\sin\theta_1\cos\theta_2 \\ &\vdots \\ y_{d-1} &= r\sin\theta_1 \ldots \sin\theta_{d-2} \\ y_d &= r\sin\theta_1 \ldots \sin\theta_{d-1} \end{aligned} \tag{25}$$

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

$$\begin{aligned} \theta_1 &= \cos^{-1}\left[\frac{y_1}{\|y_{1:d}\|_2}\right] \\ \theta_2 &= \cos^{-1}\left[\frac{y_2}{\|y_{2:d}\|_2}\right] \\ &\vdots \\ \theta_{d-1} &= \cos^{-1}\left[\frac{y_{d-1}}{\|y_{(d-1):d}\|_2}\right]. \end{aligned} \tag{26}$$

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,

$$\text{PG}(\theta \mid \alpha, \beta) = \frac{\Gamma(A)\beta_d^{\alpha_d}}{B^A \Gamma(a_d} \left(\prod_{j=1}^{d-1} \frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)} (\cos\theta_j)^{\alpha_j - 1} (\sin\theta_j)^{(\sum_{h=j+1}^{d} \alpha_h) - 1} \right) \mathcal{I}_{(0,\pi/2)^{d-1}}(\theta) \tag{27}$$

where

$$A = \sum_{j=1}^{d} \alpha_j \qquad \text{and} \qquad B = \beta_1 \cos\theta_1 + \sum_{j=2}^{d-1} \left(\beta_j \cos\theta_j \prod_{i=1}^{j-1} \sin\theta_i \right) + \beta_d \prod_{j=1}^{d} -1 \sin\theta_j. \tag{28}$$

As is, this model is not identifiable, as taking $\beta^{(2)} = \alpha\beta^{(1)}$ will still yield the same distribution of angles. Following [12], 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 (25), $\mathbf{y}' = \mathbf{y}/r$, 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 [12] 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.

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 3.3 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.

What happened to the mixture of projected gammas?

3.4 Generalized Dirichlet

The other natural extension we can form from the Dirichlet model would be to assume data as descending from a Generalized Dirichlet distribution. This model has the same support as the Dirichlet, but it is more general in that the rate parameters are not assumed to be the same. That is,

$$\mathbf{x} \sim \text{GD}(\mathbf{x} \mid \zeta, \sigma) = \int_0^\infty \prod_{l=1}^{d} \text{Ga}(rx_l \mid \zeta_l, \sigma_l) |J| \text{d}r \tag{29}$$

Again, the Jacobian is r^{d-1}. As the model is not identifiable otherwise, some restriction must be placed on the rate parameters σ, and the restriction most often used is that the first rate parameter σ_1 is set to 1.

We again consider 2 models under this family: A finite mixture model, and a dirichlet process mixture model, and we again also consider a vanilla generalized Dirichlet model.

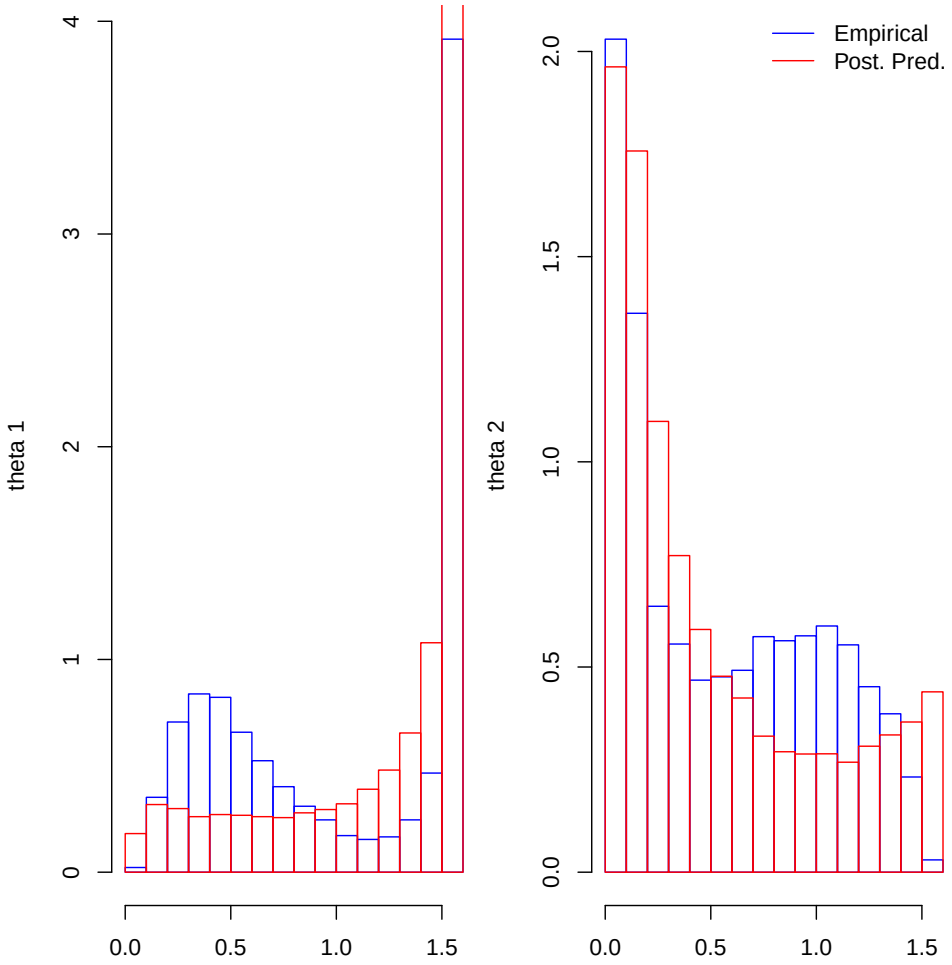

Figure 1: Histograms of Empirical vs Posterior-predictive angular data originating from a simulated 3-dimensional gamma dataset.

3.4.1 Finite Mixture of Generalized Dirichlets

We extend the finite mixture of Dirichlets by, for $l > 1$, allowing σ_{jl} to vary, and placing Gamma hyperpriors for its generating shape and rate parameters. We denote this model

as MGD.

$$
\begin{aligned}
(r_i, x_i) \mid \delta_i = j &\sim \prod l = 1^d \mathrm{Ga}(r x_{il} \mid \zeta_{jl}, \sigma_{jl}) \\
\zeta_{jl} \mid \alpha_l, \beta_l &\sim \mathrm{Ga}(\zeta_{jl} \mid \alpha_l, \beta_l) \sigma_{jl} \mid \xi_l, \tau_l \qquad =\sim \mathrm{Ga}(\sigma_{jl} \mid \xi_l, \tau_l) \text{ for } j = 2, \ldots, d \alpha_l \sim \mathrm{Ga}(\alpha_l \mid 0.5, 0.5) \\
\beta_l &\sim \mathrm{Ga}(\beta_l \mid 2, 2) \\
\xi_l &\sim \mathrm{Ga}(\xi_l \mid 0.5, 0.5) \\
\tau_l &\sim \mathrm{Ga}(\tau_l \mid 2, 2) \\
\lambda &\sim \mathrm{Dir}(0.5)
\end{aligned}
\tag{30}
$$

Again, i denotes indexing over observed data, j denotes indexing over clusters, and l denotes indexing over dimensions. Inference conducted on this model is very similar to that of the Dirichlet–we augment the data x_i with r_i, the latent sum of the independent gammas. Doing so allows us to conduct inference on each dimension l independently of the other dimensions. The latent r is generated as

$$
r_i \mid \zeta, \sigma, \delta_i = j \sim \mathrm{Ga}\left(\sum_{l=1}^{d} \zeta_{jl}, \sum_{l=1}^{d} \sigma_{jl} x_l \right), \tag{31}
$$

This results in a more flexible model as compared to the Dirichlet. The choice of somewhat informative hyperparameters for the rate hyperpriors is to ensure that, for numerical stability's sake, rate parameters do not approach 0.

3.4.2 Dirichlet Process Mixture of Generalized Dirichlets

Analogous to the DP extension to the finite mixture of Dirichlets, we place a Dirichlet process prior on the cluster parameters ζ_i, σ_i. As with DPD, and the change made from MD to MGD, this model allows σ_i to vary, and places a Gamma hyperprior on its generating shape and rate parameters. We use the same hyperparameters as the finite mixture of generalized Dirichlets, assuming that shape parameters descend from a Gamma$(0.5, 0.5)$, and rate parameters descend from a Gamma$(2, 2)$.

3.4.3 Generalized Dirichlet with log-normal prior on shape

We have formed analogous finite mixtures and Dirichlet process mixtures using the Dirichlet kernel, along with a log-normal prior on the shape parameters, in the case of the finite mixture, and a log-normal centering distribution in the case of the DP mixture. The rate parameters prior remains a product of independent gammas. That is, for the finite mixture model MGDLN,

$$
\begin{aligned}
(r_i, x_i) \mid \delta_i = j &\sim \prod l = 1^d \mathrm{Ga}(r x_{il} \mid \zeta_{jl}, \sigma_{jl}) \\
\zeta_j \mid \mu_j, \Sigma_j &\sim \log \mathcal{N}(\zeta_{jl} \mid \mu, \Sigma) \sigma_{jl} \mid \xi_l, \tau_l \qquad =\sim \mathrm{Ga}(\sigma_{jl} \mid \xi_l, \tau_l) \text{ for } j = 2, \ldots, d \xi_l \sim \mathrm{Ga}(\xi_l \mid 0.5, 0.5) \\
\tau_l &\sim \mathrm{Ga}(\tau_l \mid 2, 2) \\
\lambda &\sim \mathrm{Dir}(0.5),
\end{aligned}
\tag{32}
$$

and the Dirichlet Process mixture model, DPGDLN

$$
\begin{aligned}
(r_i, \mathbf{x}_i) \mid \zeta_i &\sim \prod_{l=1}^{d} \mathrm{Ga}(r_i x_{il} \mid \zeta_{il}, 1) \\
\zeta_i &\sim \mathrm{DP}(\eta, G) \qquad\qquad G = \log \mathcal{N}(\zeta_i \mid \mu, \Sigma) \prod_{l=2}^{d} \mathrm{Ga}(\sigma_{il} \mid \xi_l, tau_l) \\
\alpha_l &\sim \mathrm{Ga}(\alpha_l \mid 0.5, 0.5) \\
\beta_l &\sim \mathrm{Ga}(\beta_l \mid 2, 2) \\
\eta &\sim \mathrm{Ga}(\eta \mid 2, \kappa) \qquad\qquad \kappa \in \{0.1, 1, 10\}.
\end{aligned} \tag{33}
$$

3.5 Normal model built on Probit representation of Spherical Coordinate Space

The transformation in Equation 25 provides us a mapping from S_2^{d-1} to a $d-1$ dimensional cube, $[0, \pi/2]$. Building on this transformation allows us to represent the data using a true $d-1$ dimensional distribution, rather than generating a latent parameter to induce a d dimensional distribution, as we do with all the gamma based models.

A canonical choice of distribution in this space might be to further transform to $(-\infty, \infty)$ via marginal probit or logit transformation, and represent the data as multivariate normal. We try that here. Let $W_i = \mathrm{Probit}(2\theta_i/\pi)$–that is, scale θ_i to the unit interval, then conduct a probit transformation on it to result in $W_i \in (-\infty, \infty)$. Then we establish a multivariate normal distribution on W. As we are expecting data to descend from a mixture of distributions, we place a DP prior on the multivariate normal kernel distribution. The centering distribution of the DP prior is the product of a multivariate normal and inverse Wishart distribution; and we place multivarite normal and inverse Wishart priors on these parameters. As before, we place a gamma prior on the DP concentration parameter η.

$$
\begin{aligned}
W_i &\sim \mathcal{N}_{d-1}(\mu_i, \Sigma_i) \\
\mu_i, \sigma_i &\sim G_i \\
G_i &\sim \mathrm{DP}(\eta, G_0(\mu_i, \Sigma_i \mid \mu_0, \Sigma_0)) \\
G_0(\mu_i, \Sigma_i \mid \mu_0, \Sigma_0) &= \mathcal{N}_{d-1}(\mu_i \mid \mu_0, \Sigma_0)\mathrm{IW}(\Sigma \mid \nu, \psi) \\
\mu_0 &\sim \mathcal{N}_{d-1}(\mathbf{u}, \mathbf{S}) \\
\Sigma_0 &\sim \mathrm{IW}(\nu_0, \psi_0) \\
\eta &\sim \mathrm{Ga}(\alpha, \beta)
\end{aligned} \tag{34}
$$

There is an advantage in that for inference on μ_i, μ_0, Σ_i, and Σ_0 this model is completely conjugate. However, while the transformation employed in Equation 25 is one to one, small deviations in different dimensions on S_∞^{d-1} have vastly different effects on the resulting transformed variables. This induced distortion may result in an inferior model, when evaluating on S_∞^{d-1}. We will be evaluating this model as representative of models on the $d-1$ dimensional coordinate space, and comparing it against other models, after projecting back onto S_∞^{d-1}.

Another disadvantage of this model is the need to compute $d-1$-dimensional matrix determinants and inversions. If we consider that inversion is a $\mathcal{O}(n^3)$ operation, we

face the real problem of computation time climbing astronomically as the number of dimensions grows. We are currently evaluating this model on 8 and 46 dimensions, we can see that those operations on 46 dimensions will take around 190 times longer than on 8 dimensions. This presents a problem if we want to run this model in any reasonable time scale.

3.6 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 [7] and the energy score criterion of [8]. Both of these metrics require calculating some distance in the target space, and this section will be devoted to that end. As both these metrics operate on the posterior predictive distribution for each observation, we will also be including a metric operating on the overall posterior predictive distribution, in the form of a modified Kullbeck-Liebler divergence.

3.6.1 Posterior Predictive Loss Criterion

The posterior predictive loss criterion, PPL is introduced in [7]. When we assume a squared error loss function, then for the ith observation, the posterior predictive loss criterion is computed as

$$D_k^{(i)} = \mathrm{Var}(X_i) + \frac{k}{k+1}\left(\mathrm{E}[X_i] - x_i\right)^2, \tag{35}$$

where X_i is a random variable from the posterior predictive distribution for x_i. 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. As we are dealing with a multivariate distribution, in keeping with our spatial statistics brethren find a citation where they do this, we will be taking the posterior predictive loss averaged over all dimensions. That is,

$$D_k^i = \frac{1}{d}\sum_{l=1}^{d}\left[\mathrm{Var}(X_{il}) + \frac{k}{k+1}\left(\mathrm{E}[X_{il}] - x_{il}\right)^2\right] \tag{36}$$

Then we report the average D_k, taken over the whole dataset.

3.6.2 Energy Score

The energy score of Gneiting, et al[8], is a generalization of the continuous ranked probability score, or crps, defined for a multi-dimensional random variable.

$$\mathrm{ES}(P, x_i) = \frac{1}{2}\mathrm{E}_p\|\mathbf{X}_i, \mathbf{X}_i'\|_\Omega^\beta - \mathrm{E}_p\|\mathbf{X}_i - \mathbf{x}_i\|_\Omega^\beta \tag{37}$$

I don't understand the first term of this expression. where $\mathbf{X}_i'$ is another replicate from the posterior predictive distribution of $\mathbf{x}_i$. This means, rather than relying on the first

and second moments of the posterior predictive distribution as in the case of posterior predictive loss[7], 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 established distance metrics developed on the positive orthant of the unit hypercube. In the unit simplex, we can assume the use of Euclidean norm as accurately representing the distance between two points. For any $\mathcal{S}_p^{d-1}$ space, $p > 1$, Euclidean norm is going to under-report the actual distance required for travel between two points. That distortion is maximized on $\mathcal{S}_\infty^{d-1}$

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 $d-2$ dimensional cube, and observations in this space have dimensions i, j equal to 1.

A distance in this space is 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:

$$a_2' = \begin{bmatrix}1\\2\\0\end{bmatrix} + \begin{bmatrix}0 & 0 & 0\\-1 & 0 & 0\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}x_2\\1\\z_2\end{bmatrix} = \begin{bmatrix}1\\2-x_2\\z_2\end{bmatrix} \tag{38}$$

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:

$$a_2' = \begin{bmatrix}1\\2\\2\end{bmatrix} + \begin{bmatrix}0 & 0 & 0\\0 & 0 & -1\\-1 & 0 & 0\end{bmatrix}\begin{bmatrix}x_2\\1\\z_2\end{bmatrix} = \begin{bmatrix}1\\2-z_2\\2-x_2\end{bmatrix} \tag{39}$$

Every successive rotation is relative to the last face. So, as the number of dimensions grows, the number of possible rotations grows as well. 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 that first rotation–we held that as a numerically easier analogue to our actual goal–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,

$$\|\mathbf{a}, \mathbf{b}\|_H = \|\mathbf{a}, \mathbf{c}\|_2 + \|\mathbf{c}, \mathbf{b}\|_2 \tag{40}$$

If we can be assured of symmetry in the functional arguments, then the requirements for a negative definite kernel are trivially proved. And from geometryneed citation on this, we can assert that on a one-dimension unfolding, the line segment connecting the starting and ending points will travel through that optimal net.

This whole paragraph can be made much shorter and crispier. Introduce the problem, then mention the difficulty of defining an actual distance in S_∞^{d-1} and jump into the negative kernel idea, then write the lemma from the paper by Gneiting and Raftery, define a suitable kernel and write a proposition proving that it is negative definite, then comment on how you calculate it. BTW, I still don't understand the notation ——a,b——.

3.6.3 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. 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. Comparing our energy score results against this value offers a metric of how well we are partitioning the data into separate distributions to be evaluated.

3.7 Kullbeck Liebler Divergence

The Kullbeck Liebler divergence D_{KL} offers a means of comparing two densities. It is, at its core, a logratio of the two densities, integrated over the first density. That is,

$$D_{\mathrm{KL}}(A, B) = \int_{x \in \Omega(A)} A(x) \log\left(\frac{A(x)}{B(x)}\right) dx \tag{41}$$

where $\Omega(\cdot)$ indicates the support, and A, B are densities. In our case, as the densities are intractible, we will be estimating the densities using a numerical method. There is a wealth of literature on this topic, but for reasons to be made clear, I will discuss two and use one. You need to sharpen this explanation and possibly include a formula.

Parzen [13] offers a method of density estimation involving setting a window with radius σ around a point, and counting observations within that window. The number of observations within the window, divided by the total number of observations within the dataset, and the area of the window provides an estimate of the density at that point. This is highly dependent upon the choice of window radius σ, but importantly for our purpose, the area of the window on S_∞^{d-1} is not well defined. Consider S_∞^2, the surface of the positive orthant of the cube. Now imagine a circle. Imagine plastering that circle on different points of the cube. If the circle is entirely contained on one face of the cube, it can spread out and exhibit its full area. If it crosses between 2 faces, it can still exhibit its full area. But if we place that circle with its center at $(1, 1, 1)$, then plaster down the sides, we are left with $1/4$ of the circle unable to be pasted down. That is, The circle at that point has an area of only $\frac{3}{4}\pi r^2$ whereas a circle of radius r not intersecting that point will have an area of πr^2. In higher dimensions, this relationship becomes more murky.

Even outside of our specific geometry, the Parzen method suffers another issue–the curse of dimensionality. [3] shows, for finite sample size, that as the number of dimensions increases, the probability of observations falling into the window on the other dataset decreases. Effectively, the quality of density estimation goes down as dimensionality goes up.

An alternative to this method that we will use is based on distance between observations from the empirical and posterior predictive distributions. Similar to the Parzen method, we're working with samples from the distribution, but instead of establishing a window of some width, we're using distance from the target observation to its k nearest neighbors in the dataset. From hence, we get the name KNN Based KL Divergence. The FNN package in R implements a Euclidean distance based version of this approach. As we operate on the hypersphere, we again recognize that the Euclidean distance metric is not appropriate, delivering results strongly biased downwards. Our one-rotation norm method described previously is employed here as a distance. We recognize that the one-rotation norm is on average biased slightly longer as compared to a proper geodesic, but it is closer to truth than any other distance metric we are aware of. To that end, we employ it as the pairwise distance measure required by KNN for this KNN based KL divergence. We employ the formula of Boltz et al [3], as

$$D_{\mathrm{KL}}^{(k)}(A, B) = \log \frac{n(B)}{n(A)} + c(A) \left[\rho_A^{(k)}(B) - \rho_A^{(k)}(A)\right] \tag{42}$$

where $\rho_A(\cdot)$ denotes the average log-distance for observations in A to their kth nearest observation in $\cdot$. Boltz et al acknowledge that their method is biased, does not integrate to 1, but its disadvantages in bias as compared to the Parzen windowing method are less apparent at higher dimensions, where the windowing method's weakness to dimensionality tends to rise. As we regard this as a numerical approximation, we offer this metric in the spirit of complementary evidence, rather than an authority, or oracle, to show us the way.

3.8 Spatial Threshold Modeling

Following the work of [4], 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

$$Z(\mathbf{s}) = \left(1 + \gamma(\mathbf{s})\frac{X(\mathbf{s}) - b_t(\mathbf{s})}{a_t(\mathbf{s})}\right)_+^{1/\gamma(\mathbf{s})} \tag{43}$$

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.

4 Results

4.1 Integrated Vapor Transport

We use data from the integrated water vapor transport (IVT) modelneeds citation of atmospheric rivers as a means of testing these models. An atmospheric river is a meteorological event of local concentration of water vapor in the atmosphere that moves with wind patterns. Understanding dependence among extreme events in atmospheric rivers finish thought.

Fitting our models to this data requires some pre-processing. The marginal distributions of the IVT data appear naturally log-normal, which falls into the domain of attraction of a Gumbel distribution. Given that, we can apply thresholding and exceedances over that threshold will follow a generalized Pareto distribution. Our model begins with that assessment. As estimating the Pareto parameters is not yet our focus in this analysis, we choose to apply the threshold using the empirical CDF. That is, for a given t, let $b_t = \hat{F}^{-1}(1-t^{-1})$. For this analysis, we set $t = 20$, indicating the 95 percentile. The other parameters of the generalized Pareto–the scale parameter α_t and the extremal index χ–are set via maximum likelihood. A fully Bayesian model formulation will allow their varying within a distribution, and we hope to eventually allow that. However, such a model will not lend itself to being fitted by MCMC.

After the thresholding and maximum likelihood estimation of the parameters of the Pareto, we scale the data to the standard multivariate Pareto. Dividing each observation by its $\mathcal{L}_\infty$ norm, we arrive at data on the hypercube. Data in sequence represents observations in time, and these are heavily correlated. As such, we choose to decluster the observations by, observing a sequence of observations $\mathbf{z}$ for which $\|z_i\|_\infty > 1$ for each observation in sequence, we keep only the observation with the maximum observed $\mathcal{L}_\infty$ norm.

We fit the proscribed models to the data via whatever projection is necessary, run the models, and use the fitted model parameters to generate posterior predictive distributions–both overall, for use with the KL divergence metric, and conditioned on individual observations, for use with the posterior predictive loss and energy score metrics.

We have, at our disposal, two datasets from the IVT model. One records data from 8 grid cells, covering generally the coast of California. The other, with a higher resolution, records data from 46 grid cells covering the same area. We fit our models to both datasets.

			dim = 8		dim = 46	
Model	Mix	Prior	PPL	ES	PPL	ES
Dirichlet			1.819	0.819	6.776	1.547
Dirichlet	DP	Gamma	0.200	0.173	2.429	0.833
Dirichlet	DP	LogNormal	0.230	0.193	1.643	0.656
Dirichlet	M	Gamma	0.214	0.190	1.412	0.620
Dirichlet	M	LogNormal	0.222	0.190	1.445	0.616
Gen. Dirichlet			1.818	0.841	6.616	1.633
Gen. Dirichlet	DP	Gamma	0.795	0.422	4.512	1.257
Gen. Dirichlet	DP	LogNormal	0.386	0.278	2.400	0.822
Gen. Dirichlet	M	Gamma	0.525	0.349	3.029	0.976
Gen. Dirichlet	M	LogNormal	0.637	0.372	3.774	1.105
Probit Normal	DP	Normal	0.843	0.446	8.018	1.609
Proj. Gamma			1.815	0.841	6.614	1.633
Proj. Gamma	DP	Gamma	0.778	0.408	5.572	1.480
Proj. Gamma	DP	LogNormal	0.370	0.265	2.424	0.835
Proj. Gamma	M	Gamma	0.578	0.365	2.824	0.935
Proj. Gamma	M	LogNormal	0.576	0.349	5.165	1.396
Proj. Res. Gamma			1.818	0.818	6.776	1.546
Proj. Res. Gamma	DP	Gamma	0.202	0.176	2.039	0.772
Proj. Res. Gamma	DP	LogNormal	0.228	0.191	1.470	0.584
Proj. Res. Gamma	M	Gamma	0.217	0.189	1.796	0.704
Proj. Res. Gamma	M	LogNormal	0.249	0.202	1.269	0.528

As is visible from Table **??**, we see a strong preference for the mixture models as compared to the vanilla models both in posterior predictive loss and energy score. As these measure model performance conditional on the observed data's posterior mixture component flags, it would likely be the case that a mixture model is preferred as compared to a single model. We also employ the KL divergence metric to look for differences in kernel density.

That said, there is much we can glean from this table. The normal model, which we developed after having cast the data from S_∞^{d-1} to the $d-1$ cube $[0, \pi/2]^{d-1}$ via spherical coordinate transformation, then scaled via probit transformation to $(-\infty, \infty)$, was not competitive when we transform samples from its posterior predictive distribution back onto S_∞^{d-1}. The spherical coordinate transformation induces a high degree of dependence between dimensions, so it is difficult to think of the geometry of that space as orthogonal. Never the less, fitting a normal model in that space assumes something like orthogonality, which obviously doesn't hold. The resulting low model performance in the normal model is likely a result of this distortion.

Another inference to learn from this table is the generally better results of the restricted Gamma models–Dirichlet, projected restricted Gamma–as compared to the Gamma models that allow for a varying rate parameter–generalized Dirichlet and projected Gamma. Note that regardless of what prior was used for the shape parameter, we always assumed

a Gamma prior for the rate parameter if we allowed it to vary. In one model formulation, we attempted a multivariate log-normal prior that covered both the rate and shape parameters, but this overall resulted in worse model performance even comparing to our current results.

It is no surprise that Dirichlet process models are generally preferred as compared to the finite mixture models. Properly tuned, finite mixture models will likely only do as well as the Dirichlet process mixture model when their number of extant model components are in rough agreement. The Dirichlet process just allows us to bypass to some extent that tuning process. That said, for a given performance, the finite mixture model would be preferred, as much of the model sampling can be accomplished in parallel.

The final, and most important, revelation this table contains concerns relative model performance between the incarnations of the restricted Gamma model. Given the mixture method, the Dirichlet and projected restricted Gamma model are effectively the same model, just built on a different projection of the original data. And indeed, on the 8 dimensional data, the model performance is comparable. But on the higher dimensional data, we see model performance favor the model built on S_2^{d-1} rather than S_1^{d-1}, the simplex. That difference in model performance is of interest.

Interpreting the KL divergence curves in Figure 4.1, KL divergence at its core is a log-ratio of densities. We would prefer, all else being equal, a KL divergence near 0. Interpreting this particular divergence metric, a KL divergence less than 0 means on average, that the kth closest sample from the posterior predictive dataset is closer to the observation than the kth closest sample from the empirical dataset. There is a logged ratio of sample sizes to account for differences in cardinality between the empirical dataset and posterior predictive.

From this, we would interpret the best model as that model which remains closest to the horizontal line at 0. For the 8-dimension model, this means the generalized Dirichlet and projected Gamma models, the unrestricted Gamma based models, seem to perform the best. This is in stark contrast to our inference from the earlier posterior predictive loss and energy score criterions, which sharply favored the restricted Gamma models. However, as dimensionality increases, we see that the restricted Gamma models again become favored, indicating that the greater flexibility of the unrestricted Gamma models becomes burdensome to fit in a high-dimensional case.

5 Conclusion

References

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

[2] David E. Allen, Abhay K. Singh, and Robert J. Powell. Evt and tail-risk modelling: Evidence from market indices and volatility series. The North American Journal of Economics and Finance, 26:355–369, 2013.

[3] Sylvain Boltz, Eric Debreuve, and Michel Barlaud. High-dimensional statistical measure for region-of-interest tracking. IEEE Transactions on Image Processing, 18(6):1266–1283, 2009.

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

[5] Ronald Aylmer Fisher and Leonard Henry Caleb Tippett. 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, 1928.

[6] Maurice Fréchet. Sur la loi de probabilité de l'écart maximum. Ann. Soc. Math. Polon., 6:93–116, 1927.

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

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

[9] Emil Julius Gumbel. Les valeurs extrêmes des distributions statistiques. Annales de l'institut Henri Poincaré, 5(2):115–158, 1935.

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

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

[12] 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.

[13] Emanuel Parzen. On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3):1065–1076, 1962.

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

[15] Jill C Trepanier and Clay S Tucker. Event-based climatology of tropical cyclone rainfall in houston, texas and miami, florida. Atmosphere, 9(5):170, 2018.

[16] Waloddi Weibull et al. A statistical distribution function of wide applicability. Journal of applied mechanics, 18(3):293–297, 1951.

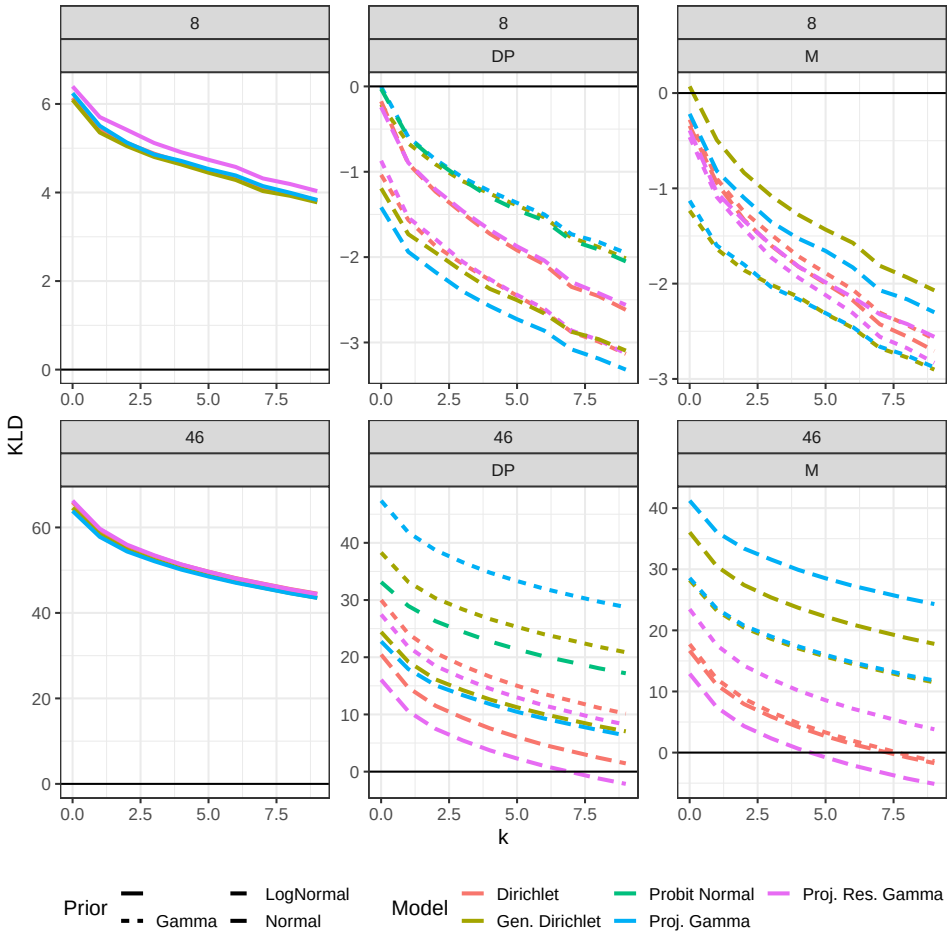

Figure 2: KL Divergence Curves calculated through the KNN-KL Metric, evaluated between the empirical dataset and posterior predictive datasets of various models. The top row corresponds to the 8-dimensional data, while the bottom row the 46. The left column corresponds to the vanilla models with no mixture method; the middle column a Dirichlet process prior, and the right column a finite mixture model.