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

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

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 $\alpha \subset \{1, \dots, d\}$. 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

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

[3] 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:

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

Thus $\mathbf{V}$ is the projection of the $\mathbf{Z}$ vector onto the unit hypersphere, using the L_∞ norm. 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 hypersphere. To construct a parametric model on that angular distribution, we employ the projected gamma distribution.

5.1 Projected Gamma

The projected gamma distribution, developed in [2], is built upon the product of d independent gamma distributions. That is, for $\mathbf{y} = (y_1, \ldots, y_d)^t$, and $y_i \sim \text{Ga}(\alpha_i, \beta_i)$, we define our starting point:

$$f(\mathbf{y} \mid \alpha, \beta) = \prod_{j=1}^{d} \text{Ga}(y_j \mid \alpha_j, \beta_j), \tag{12}$$

where β is specified as a rate parameter. From that, we transform to d-dimensional spherical coordinates $\mathbf{y} \to (r, \theta)$ as

$$\begin{aligned} y_1 &= r \cos\theta_1, \\ y_2 &= r \sin\theta_1 \cos\theta_2 \\ &\vdots \\ y_{d-1} &= r \sin\theta_1 \ldots \sin\theta_{d-2} \\ y_d &= r \sin\theta_1 \ldots \sin\theta_{d-1} \end{aligned} \tag{13}$$

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

$$\begin{aligned}
\theta_1 &= \cos^{-1}\left[\frac{y_1}{\|y_{1:d}\|_2}\right] \\
\theta_2 &= \cos^{-1}\left[\frac{y_2}{\|y_{2:d}\|_2}\right] \\
&\vdots \\
\theta_{d-1} &= \cos^{-1}\left[\frac{y_{d-1}}{\|y_{(d-1):d}\|_2}\right].
\end{aligned} \tag{14}$$

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*,

$$\mathrm{PG}(\theta \mid \alpha, \beta) = \frac{\Gamma(A)\beta_d^{\alpha_d}}{B^A\Gamma(a_d}\left(\prod_{j=1}^{d-1}\frac{\beta_j^{\alpha_j}}{\Gamma(\alpha_j)}(\cos\theta_j)^{\alpha_j-1}(\sin\theta_j)^{(\sum_{h=j+1}^{d}\alpha_h)-1}\right)\mathcal{I}_{(0,\pi/2)^{d-1}}(\theta) \tag{15}$$

where

$$A = \sum_{j=1}^{d}\alpha_j \qquad \text{and} \qquad B = \beta_1\cos\theta_1 + \sum_{j=2}^{d-1}\left(\beta_j\cos\theta_j\prod_{i=1}^{j-1}\sin\theta_i\right) + \beta_d\prod_{j=1}^{d}-1\sin\theta_j. \tag{16}$$

As is, this model is not identifiable, as taking $\beta^{(2)} = \alpha\beta^{(1)}$ will still yield the same distribution of angles. Following [2], 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 (13), $\mathbf{y}' = \mathbf{y}/r$, 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 [2] 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

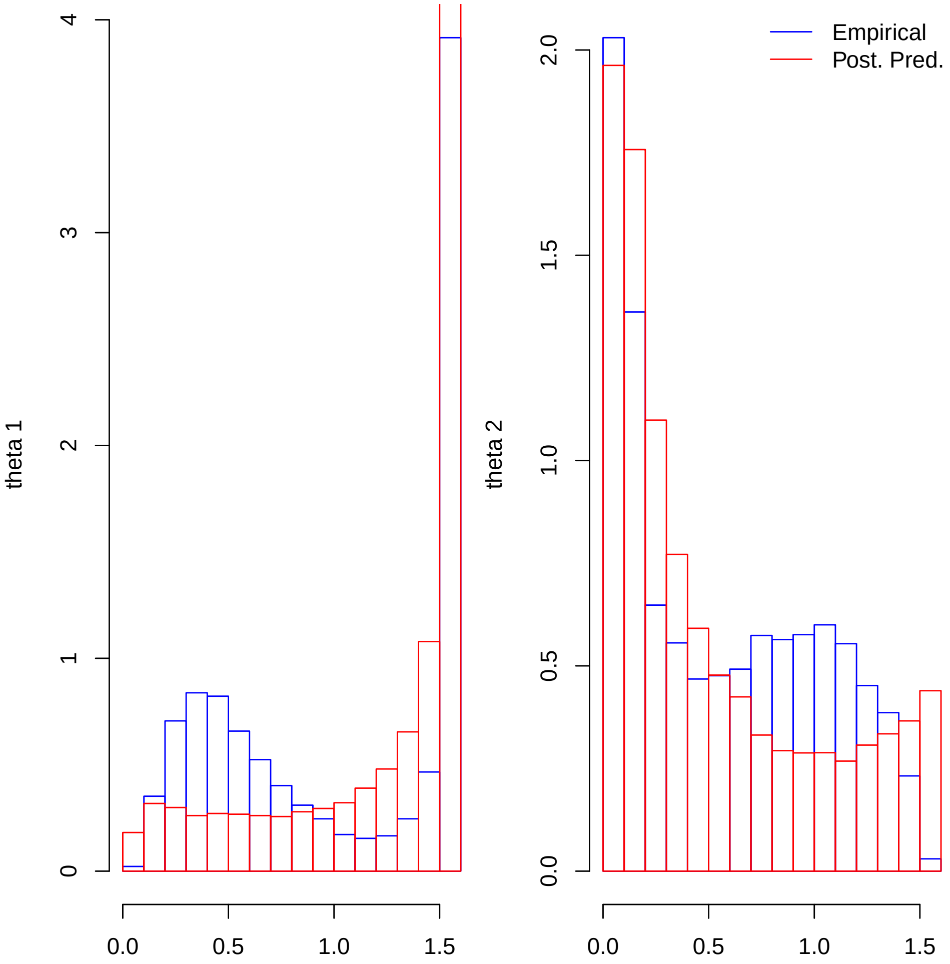


Figure 1: Histograms of Empirical vs Posterior-predictive angular data originating from a simulated 3-dimensional gamma dataset.

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 5.1 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.

5.2 Projected Mixture of Gammas (PMG) Model

One shortcoming of the vanilla method is there could be multiple generating distributions for the extreme events. Indeed, if we attempt to model real data and observe the marginal empirical distributions of θ versus the marginal predicted, we observe a disconnect: the marginal predicted on some axes looks very different from the marginal empirical. We attempt to solve this problem by creating a two-component mixture model for each gamma.

A justification for this model is that, while more complex than the vanilla model, it is significantly less complex than the mixtures of projected gamma, whether finite or non-finite. If we can represent the nuances of real data using a more parsimonious model, then that would be preferred.

As previously stated, $\mathbf{y}$ represents a an array of latent gamma variates. This model considers each marginal gamma as a 2-component mixture model. I.e., $\mathbf{y} = r\mathbf{y}'$, where $\mathbf{y}'$ is a function of the data; considering only the jth column of $\mathbf{y}$, considering only a single observation:

$$f(y_j \mid \alpha_j, \beta_j, \lambda_j) = \lambda_j \text{Ga}(y_j \mid \alpha_{j1}, \beta_{j1}) + (1-\lambda_j)\text{Ga}(y_j \mid \alpha_{j2}, \beta_{j2})$$

To make this Gibbs-able, we add a latent flag, $\gamma_j \in \{0, 1\}$ which indicates which part of the mixture the observation y_j is coming from. That is,

$$f(y_j, \gamma_j \mid \alpha_j, \beta_j, \lambda_j) = \left(\lambda_j \text{Ga}(y_j \mid \alpha_{j1}, \beta_{j1})\right)^{\gamma_j} \left((1-\lambda_j)\text{Ga}(y_j \mid \alpha_{j2}, \beta_{j2})\right)^{1-\gamma_j}$$

Then, considering the likelihood of $\mathbf{y}_j$:

$$\begin{aligned} L(\mathbf{y}_j, \gamma_j \mid \alpha_j, \beta_j, \lambda_j) &= \prod_{i=1}^{n} \left(\lambda_j \text{Ga}(y_{ij} \mid \alpha_{j1}, \beta_{j1})\right)^{\gamma_{ij}} \left((1-\lambda_j)\text{Ga}(y_{ij} \mid \alpha_{j2}, \beta_{j2})\right)^{1-\gamma_{ij}} \\ &= \frac{(\lambda_j)^{n_j} \beta_{j,1}^{\alpha_{j,1} n_j}}{\Gamma(\alpha_{j,1})^{n_j}} \left(\prod_{i:\gamma_{ij}=1} y_{ij}\right)^{\alpha_{j1}-1} \exp\left\{-\beta_{j1} \sum_{i:\gamma_{ij}=1} y_{ij}\right\} \\ &\quad \times \frac{(1-\lambda_j)^{n-n_j} \beta_{j2}^{\alpha_{j2}(n-n_j)}}{\Gamma(\alpha_{j2})^{n-n_j}} \left(\prod_{i:\gamma_{ij}\neq 1} y_{ij}\right)^{\alpha_{j2}-1} \exp\left\{-\beta_{j2} \sum_{i:\gamma_{ij}\neq 1} y_{ij}\right\} \end{aligned}$$

where $n_j = \sum_{i:\gamma_{i,j}=1} 1$. Then extracting the full conditional for $\gamma_{ij} \mid \mathbf{y}_j$, we have:

$$\pi(\gamma_{ij} \mid \ldots) = \text{Ber}\left(\gamma_j \mid \frac{\lambda_j \text{Ga}(y_{ij} \mid \alpha_{j1}, \beta_{j1})}{\lambda_j \text{Ga}(y_{ij} \mid \alpha_{j1}, \beta_{j1}) + (1-\lambda_j)\text{Ga}(y_{ij} \mid \alpha_{j2}, \beta_{j2})}\right).$$

With a prior on λ such that $\pi(\lambda) = \text{Beta}(a_0, b_0)$, we have a Beta posterior:

$$\pi(\lambda_j \mid \ldots) = \text{Beta}\left(a_0 + \sum_i \gamma_{ij}, b_0 + \sum_i (1-\gamma_{ij})\right)$$

Finally, we have the α and β parameters. We constrain the model such that $\beta_{jl} = 1$ for

$j=1$, for $l \in \{1,2\}$. Then the full conditionals for $\alpha_{j,l}$, $\beta_{j,l}$ are updated as follows:

$$\begin{aligned}\pi(\beta_{jl} \mid \alpha_{jl}, \gamma_j, \mathbf{y}_j) &\propto \beta_{jl}^{n_j\alpha_{jl}} \exp\left\{-\beta_{jl}\sum_{i:\gamma_{ijl}=1} y_{ij}\right\} \times \text{Ga}(\beta \mid c, d)\\ &\propto \text{Ga}\left(n_{jl}\alpha_{jl}+c, \sum_{i:\gamma_{ijl}=1} y_{ij} + d\right)\end{aligned}$$

where $\gamma_{ij2} = 1-\gamma_{ij1}$, and $\gamma_{ij1} = \gamma_{ij}$ previously referenced. The full conditional for α_{jl} where the prior for α_{jl} is $\text{Ga}(a,b)$ and β_{jl} has been marginalized out is thus:

$$\pi(\alpha_{j,l} \mid \gamma_j, \mathbf{y}_j) \propto \frac{\alpha^{a-1}(\prod_{i:\gamma_{ijl}=1} y_{ij})^{\alpha_{jl}-1}}{\left(\sum_{i:\gamma_{ijl}=1} y_{ij} + d\right)^{n_{jl}\alpha_{jl}+c}} \frac{\Gamma(n_{jl}\alpha_{jl}+c)}{\Gamma(\alpha_{jl})^{n_{jl}}} \exp\{-b\alpha_{jl}\}$$

Sampling α_{jl} proceeds using a Metropolis Hastings algorithm on the transformed parameter $\log(\alpha_{jl})$.

As previously stated, the *appeal* of this model arises from its comparative simplicity relative to the more complex mixtures of projected gamma. However, that advantage is contingent on its ability to represent the nuances of data. As we can see from need to make plot, this method by design ignores information between dimensions, and as a result we see the posterior predictive distribution generated from our fitted model looks startlingly different to the empirical distribution of the data. A more complicated model will be needed.

5.3 Mixture of Projected Gammas (MPG) Model

In 5.2, we introduced a projection of two-component mixture of gammas (pgm) model. Here, we go another way, mixing directly on the projected gamma distribution at the projection level. This allows us to establish essentially clusters of angular vectors. Developing the math on this:

$$\begin{aligned}\text{MPG}(\theta \mid \lambda, \alpha, \beta) &= \sum_{j=1}^{J} \lambda_i \text{PG}(\alpha_i, \beta_i)\\ &= \int \text{MPG}(\theta, \gamma \mid \alpha, \beta) d\gamma\\ &= \int \prod_{j=1}^{J} \left[\lambda_j \text{PG}(\theta \mid \alpha_j, \beta_j)\right]^{\gamma_j} \text{d}\gamma \qquad (17)\\ &= \int_\gamma \int_r \lambda_j^{\gamma_j} \prod_{k=1}^{K} \text{Ga}(r\mathbf{y}' \mid \alpha_\mathbf{j}, \beta_\mathbf{j})^{\gamma_j} |\text{Jac}| \text{d}r\text{d}\gamma \qquad (18)\end{aligned}$$

letting γ_{ij} be an indicator that observation i is in the j'th mixture component. From this, we gather the familiar full conditionals. Let i iterate over observations, j iterate over mixture components, and k iterate over dimensions of the space. Placing a Gamma prior on β as before, we arrive at a Gamma posterior for β_{jk} of the form:

$$\pi(\beta_{jk} \mid \alpha_{jk}, \mathbf{r}, \gamma) = \text{Ga}\left(\sum_{i=1}^{n} \gamma_{ij}\alpha_{jk} + c_0, \sum_{i=1}^{n} \gamma_{ij} r_i y'_{ik} + d_0\right) \tag{19}$$

Integrating β_{jk} from $\pi(\alpha_{jk}, \beta_{jk} \mid \mathbf{r}, \gamma)$, we arrive at the full conditional for α_{jk}, taking the form:

$$\pi(\alpha_{jk} \mid \gamma, \mathbf{r}\gamma) \propto \frac{\alpha_{jk}^{a_0 - 1} \prod_{i=1}^{n} (r_i y'_{ik})^{\gamma_{ij}\alpha_{jk}}}{\Gamma(\alpha_{jk})^{\sum_{i=1}^{n} \gamma_{ij}}} \exp\left[-b_0 \alpha_{jk}\right] \frac{\Gamma(\alpha_{jk} \sum_i \gamma_{ij} + c_0)}{(\sum_i r_i y'_{ik} + d_0)^{\alpha_{jk} \sum_i \gamma_{ij} + c_0}} \tag{20}$$

The radii are generated conditional on the mixture component assignment:

$$\pi(r_i \mid \gamma_i, \alpha, \beta) = \text{Ga}\left(\sum_{j=1}^{J} \gamma_{ij} \sum_{k=1}^{K} \alpha_{jk}, \sum_{k=1}^{K} \left[\sum_{j=1}^{J} \gamma_{ij}\beta_{jk}\right] y'_{ik}\right) \tag{21}$$

Finally, the mixture component indicators γ_i are generated from the Projected Gamma likelihood, rather than introducing the latent r's. The full conditional is thus:

$$\pi(\gamma_i \mid \lambda, \alpha, \beta) = \text{Multinom}\left(1, \left\{\frac{\lambda_j \text{PG}(\mathbf{y}'_i \mid \alpha_j, \beta_j)}{\sum_{l=1}^{J} \lambda_l \text{PG}(\mathbf{y}'_i \mid \alpha_l, \beta_l)}; \qquad j = 1, \ldots, J\right\}\right) \tag{22}$$

And the full conditional for λ is finally:

$$\pi(\lambda \mid \gamma) = \text{Dir}\left(\left\{\sum_{i} \gamma_{ij} + a_0; \qquad j = 1, \ldots, J\right\}\right) \tag{23}$$

The finite mixture of projected gammas model is a rise in complexity versus the projected mixture of gammas, and the vanilla projected gamma model. The question is, is that rise justified? As we can see from need to make plot, for the most part the MPG model is able to capture the nuance of real data. The marginal directional data from the posterior predictive looks very similar to that of the empirical. However, that is true of this dataset. Is it reasonable to stop here? How do we decide what is an appropriate number of mixture components?

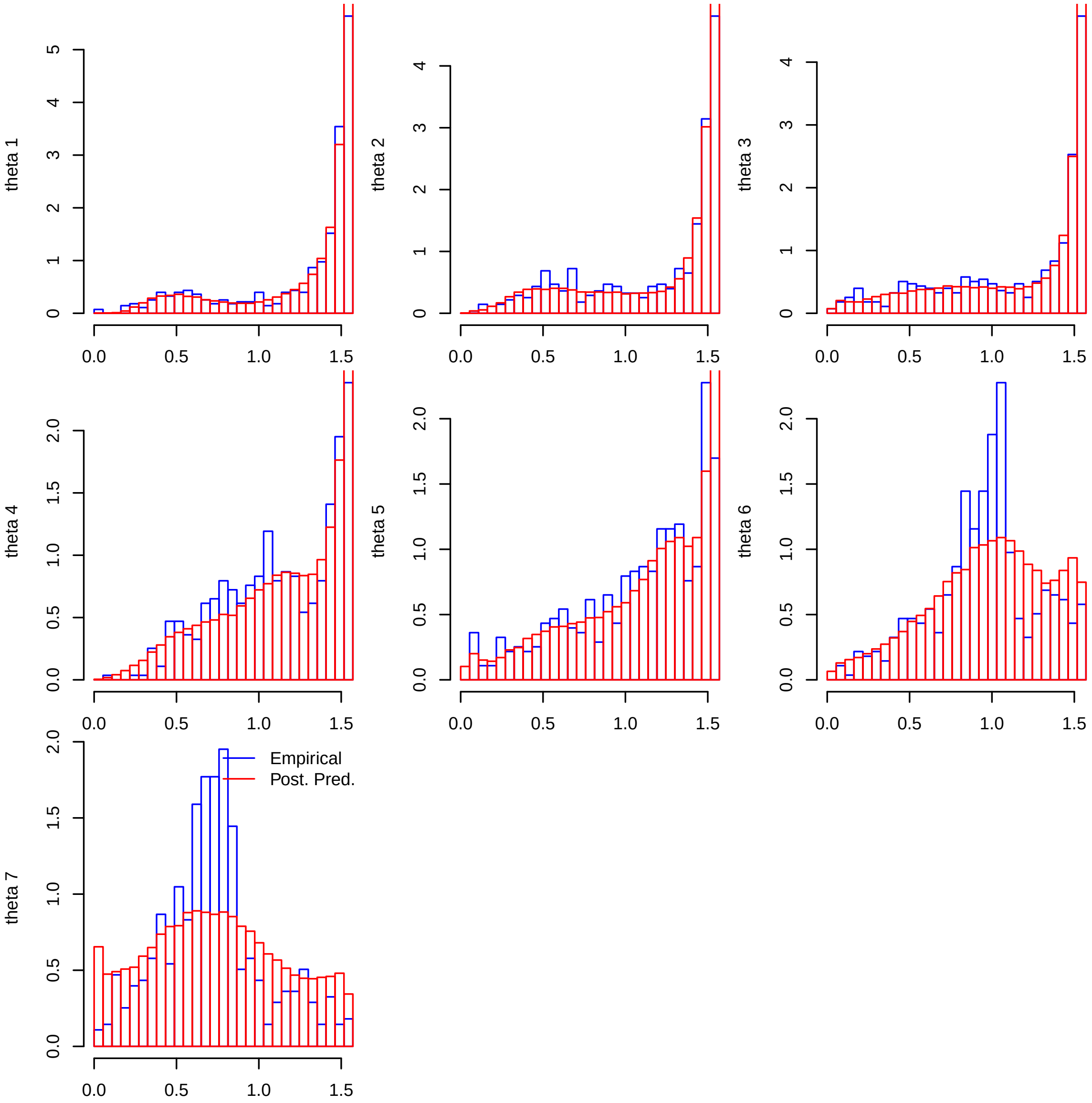
Certainly, on an ad-hoc basis one could continue adding mixture components, and checking some model selection criterion to find some optimal number of components. Alternatively, using Bayesian non-parametric modelling, we can assume a potentially infinite number of components. We explore that model next.

In Figure 5.3, we see the posterior predictive distribution generated by the finite mixture of projected gammas model. This model does well in the first θ's, but we see of θ_6 and θ_7, the model has difficulty in fitting the center spikes. These spikes indicate a high degree of extremal dependence between the last 3 columns, so that's somewhat a problem.

5.4 Non-parametric Mixture of Projected Gammas (NPPG) Model

In 5.3, we fit a finite mixture paradigm on top of the projected gamma model. This raises a legitimate question of how many mixture components will be appropriate. Certainly we

Figure 2: Posterior Predictive distribution from Mixture of Projected Gammas model with 15 mixture components, built on declustered IVT data.



could try many numbers of mixture components, and via some selection criterion choose an appropriate number. Alternatively, if we were sufficiently gluttons for punishment, we could treat the number of components as itself a random variable, and with reversible jump MCMC could assemble a finite mixture model where the number of mixture components is by the model. Alternatively, we can forgo that, and use a Dirichlet process prior on top of the Projected Gamma distribution.

$$\begin{aligned}
\theta_i &\sim \text{PG}\left(\theta_i \mid (\alpha_i, \beta_i)\right) \\
(\alpha_i, \beta_i) &\sim G_i \\
G_i &\sim \text{DP}\left(\eta, G_0\left((\alpha_i, \beta_i) \mid (\mathbf{a}_\alpha, \mathbf{b}_\alpha, \mathbf{a}_\beta, \mathbf{b}_\beta)\right)\right) \\
(\mathbf{a}_\alpha, \mathbf{b}_\alpha, \mathbf{a}_\beta, \mathbf{b}_\beta) &\sim P\left((\mathbf{a}_\alpha, \mathbf{b}_\alpha, \mathbf{a}_\beta, \mathbf{b}_\beta)\right) \\
\eta &\sim \text{Ga}(a_\eta, b_\eta)
\end{aligned} \tag{24}$$

with

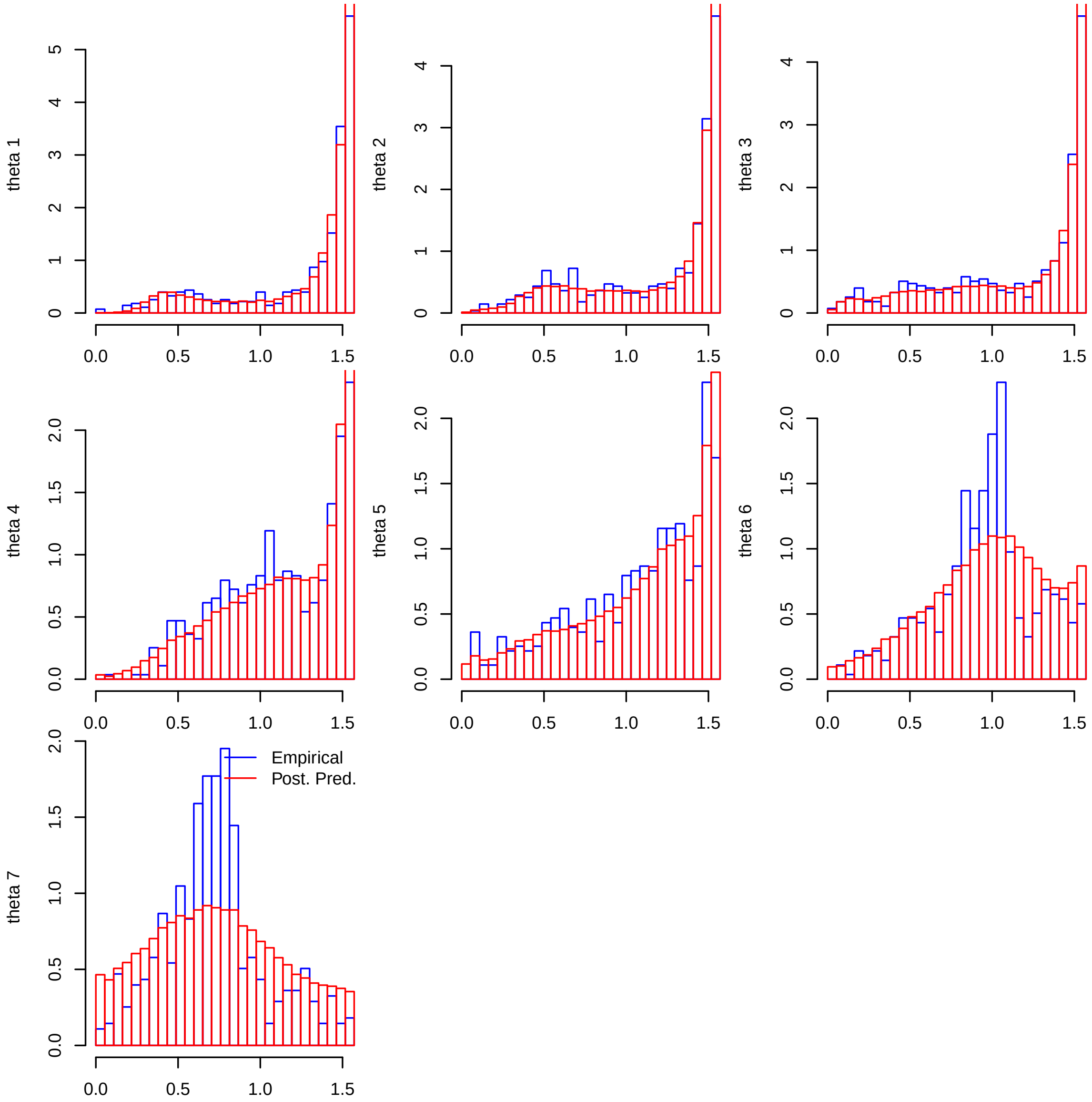
$$\begin{aligned}
G_0\left((\alpha_i, \beta_i)\right) &= \text{Ga}(\alpha_1 \mid a_{\alpha_1}, b_{\alpha_1}) \prod_{j=2}^{d} \text{Ga}(\alpha_j \mid a_{\alpha_j}, b_{\alpha_j}) \text{Ga}(\beta_j \mid a_{\beta_j}, b_{\beta_j}) \\
P((\mathbf{a}_\alpha, \mathbf{b}_\alpha, \mathbf{a}_\beta, \mathbf{b}_\beta)) &= \text{Ga}(a_{\alpha_1}) \text{Ga}(b_{\alpha_1}) \prod_{j=2}^{d} \text{Ga}(a_{\alpha_j}) \text{Ga}(b_{\alpha_j}) \text{Ga}(a_{\beta_j}) \text{Ga}(b_{\beta_j})
\end{aligned}$$

That is, treat the Projected Gamma distribution as the kernel, with a Dirichlet process prior and a product of independent gamma densities as the centering distribution. This has the advantage that the potential number of clusters is infinite, the de-facto number of clusters is random.

An obstacle to fitting this model is calculating the prior predictive density, which is not available in closed form. Another obstacle to fitting this model, is drawing new random samples for α_{ij}. That is, α_j for observation i. Posterior samples for α_j in the other finite mixture model could be sampled via Metropolis Hastings. This process takes time to reach convergence–there is no assurance that the first draw from the sampler for an otherwise empty cluster will be *from the distribution.* Every update to the cluster effectively changes the distribution. As such, these draws will have to be perfomed in some other way. As, after sampling the latent radius, we are dealing with independent gammas, we may treat each dimension independently. One solution we might consider would be to use Gaussian quadrature approximate the CDF of the distribution, and then sample from it using probability integral transform. This is computationally wasteful, so another idea we consider is slice sampling. The slice sampler is a fair amount slower than Metropolis Hastings, but we achieve significantly less correlated draws with no tuning of a proposal density. There is still a starting point that the next draw upon, but by some logic we can find a good starting point, and by successively sampling some iterations, we can arrive at a draw from the posterior that is (effectively) independent of the starting point.

As it turns out, the model as specified above has issues with specifiability, resulting in model instability. The course I took was to fit $b_\alpha = b_\beta = 1$. This at least resulted in a stable model, but as we can see in Figure 5.4, the resulting distribution does not well match the original data. The estimates for a_α (for all columns) are all between 0 and 0.2, while the estimates for a_β (for all columns > 1) are stable around 20. This results, as a prior distribution for possible clusters, in each column having a very unstable distribution independent of other columns. I believe this is why we see such a strong prior effect towards independence in Figure 5.4.

Figure 3: Dirichlet Process Mixture Model with Projected Gamma kernel, independent Gamma priors using declustered IVT data.



5.5 Implications of the distribution

Given the latent radius, the conjugate prior for the rate parameter β is a Gamma distribution, and it is standard practice to assign a Gamma distribution to the shape parameter α as well. We should take note, however, what that entails. In Figure 5.5, we see a $\text{Gamma}(1,1)^d$ distribution cast to the angular space. The point of this exercise is to ask what shape should we expect to see in the marginal θ_i's, if we assume something along the lines of a uniform prior over the unit hypersphere.

Note that this the resulting distribution of d independent Gamma(1,1) random variables, cast to the angular space using Equation 14. The top row represents a two-dimensional gamma distribution, cast to the angular space, and every row thereafter increases the dimensionality of the input space by 1.

We see that in the two-dimensional case, the Gamma$(1, 1)$ distribution creates a density that is biased towards either end of the spectrum, though not excessively. There is no combination of gamma distributions that will create a uniform density over this space. I probably should show this. If we extend this space to three and more dimensions, we see the distribution of θ_1 through θ_{d-2} shift towards the right, indicating that the mass of the norm of a particular observation would tend to fall into the latter columns. We show this here to demonstrate that this is right and expected behavior, so we should expect to some degree progressively less right-skewed marginal θ_i distributions when looking at real data.

5.6 Non-parametric Mixture of Multivariate Normals

I do not regard this as a viable model.

Note that $\mathbf{V}$ is the projection of the Generalized Pareto after marginal standardization mentioned previously onto the unit hypersphere on L_∞. From this, we see $V_i \in [0, 1]$, $\max_i V_i = 1$. I had a thought regarding this wondering how well a simple Normal-Normal model would recover the original distribution, when each observation is cast into probit space. That is, $W_i = \text{Probit}(V_i)$, where $\text{Probit}(\cdot) = \Phi^{-1}(\cdot)$, with $\Phi(\cdot)$ the CDF of a standard normal distribution. This transformation pushes $\mathbf{V} \in [0, 1]$ to $\mathbf{W} \in \mathcal{R}^d$. As far as why this is not a viable model, W has d degrees of freedom, where it should only have $d - 1$.

As every observation V has potentially some $V_i = 0$ and one $V_i = 1$, under this transformation these values would become $-\infty$ and ∞. This would not be possible to fit under the normal model, so I jitter these observations by some $\epsilon > 0$ such that $\min V_i' = \epsilon$, and $\max V_i' = 1 - \epsilon$, and $W_i = \text{Probit}(V_i')$. Thus, the model becomes

$$\begin{aligned}
W_i &\sim \mathcal{N}_d\left(\mu_i, \Sigma_i\right) \\
\mu_i, \sigma_i &\sim G_i \\
G_i &\sim \text{DP}(\eta, G_0(\mu_i, \Sigma_i \mid \mu_0, \Sigma_0)) \\
G_0(\mu_i, \Sigma_i \mid \mu_0, \Sigma_0) &= \mathcal{N}_d(\mu_i \mid \mu_0, \Sigma_0)\text{IW}(\Sigma \mid \nu, \psi) \qquad (25) \\
\mu_0 &\sim \mathcal{N}_d\left(\mathbf{u}, \mathbf{S}\right) \\
\Sigma_0 &\sim \text{IW}(\nu_0, \psi_0) \\
\eta &\sim \text{Ga}(\alpha, \beta)
\end{aligned}$$

The updates to μ_i, μ_0, Σ_i, and Σ_0 are thus known forms; $\mu_i \mid \mu_0, \mathbf{W}$ and $\mu_0 \mid \mu_i$ follow multivariate normal distributions. $\Sigma_i \mid \mathbf{W}$ and $\Sigma_0 \mid \mu_i, \mu_0$ follow inverse Wishart distributions.

As stated, this model results in a distribution with d degrees of freedom, whereas it should properly have $d - 1$. To generate the posterior predictive distribution and induce this $d - 1$ degrees of freedom, I cast the generated observations back to the hypercube. That is,

$$\mathbf{V}^{\text{new}} = \frac{\Phi(\mathbf{W}^{\text{new}})}{\max_i \mathbf{W}_{\mathbf{i}}^{\text{new}}}.$$

In Figure 5.6, we see the resulting posterior predictive distribution, after casting back to the angular space. In the Marginal θ's, this model captures pretty well the spikes in θ_6, θ_7 that have thus far confounded the mixture of projected gammas model. However, it adds an unwarranted spike in θ_4 and less so in θ_3. In θ_6, we also see a strong tendency towards the upper end of the range that is not evident in the original data. In θ_7, we see the reverse. Odd.

5.6.1 Casting θ to Probit space directly

Given that the previous model had d degrees of freedom, whereas by construction we need a model with $d-1$ degrees of freedom. When we look at the angular representation used in the projected Gamma model, that gives a $(d-1)$-dimensional representation of the data, with each $\theta_i \in [0, \pi/2]$. It stands to reason, that we can represent this in $[0,1]$ by dividing the vector by $\pi/2$, and, as before, cast this into probit space. Then we build the same normal-normal model used above, directly on this probit space. This results in one of the more compelling models thus far.

In Figure 5.6.1, we see the resulting posterior predictive distribution after fitting on the declustered IVT data. The resulting model is able to pick up the spikes in the later θ's that the mixture of projected gammas model was unable to account for. It also does not exhibit the strange edge behavior in the later θ's of the previous normal model. We might ask for a more granular model, as it appears that this model has been unable to fit some of the smaller nuances of the data, but that can likely be solved via prior specifications on η, Σ, and μ.

5.7 Non-parametric Mixture of Projected Gammas with Lognormal Prior

$$\begin{aligned}
\theta_i &\sim \text{PG}(\theta_i \mid \alpha_i, \beta_i)\\
r_i &\sim \text{Ga}(r_i \mid \alpha_i, \beta_i)\\
(\alpha_i, \beta_i) &\sim \text{DP}\left((\alpha_i, \beta_i) \mid \eta, G_0\right)\\
G_0 &= \text{Log}\mathcal{N}(\alpha_i \mid \mu, \Sigma)\prod_{j=2}^{d}\text{Ga}(\beta_{ij} \mid a, b)\\
\mu &\sim \mathcal{N}(\mu \mid \mu_0, \Sigma_0)\\
\Sigma &\sim \text{IG}(\Sigma \mid \nu, \psi)
\end{aligned} \tag{26}$$

5.8 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

$$Z(\mathbf{s}) = \left(1 + \gamma(\mathbf{s})\frac{X(\mathbf{s}) - b_t(\mathbf{s})}{a_t(\mathbf{s})}\right)_+^{1/\gamma(\mathbf{s})} \tag{27}$$

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

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] 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.

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

Figure 4: The Marginal density of a Gamma$(1,1)^d$ after casting to angular coordinates. The top row is from a two-dimensional the second a three-dimensional, the third a four-dimensional, and the bottom a five-dimensional distribution.

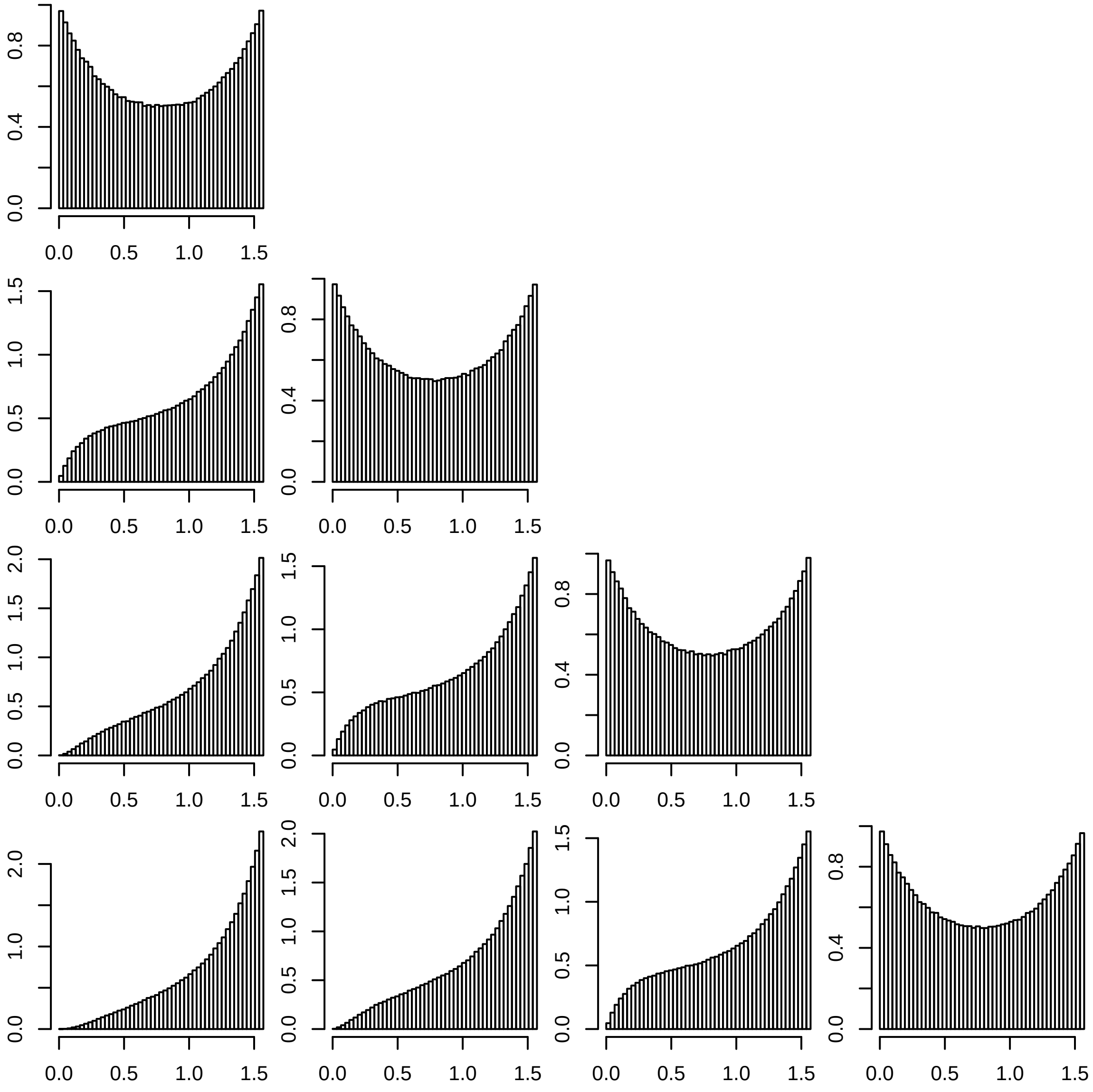


Figure 5: Dirichlet Process Mixture Model with multivariate normal kernel over probit space cast on unit hypercube using declustered IVT dataset

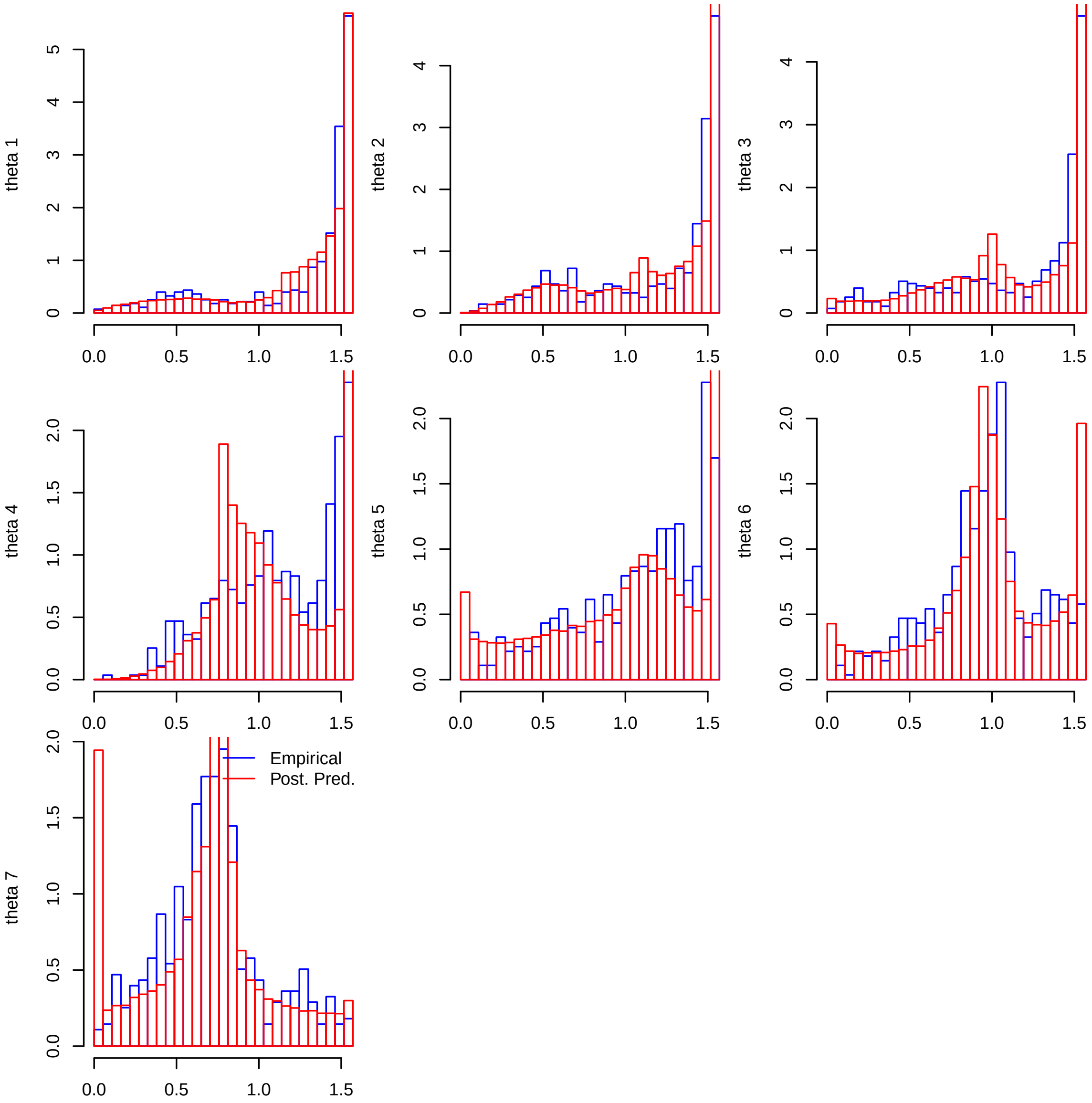
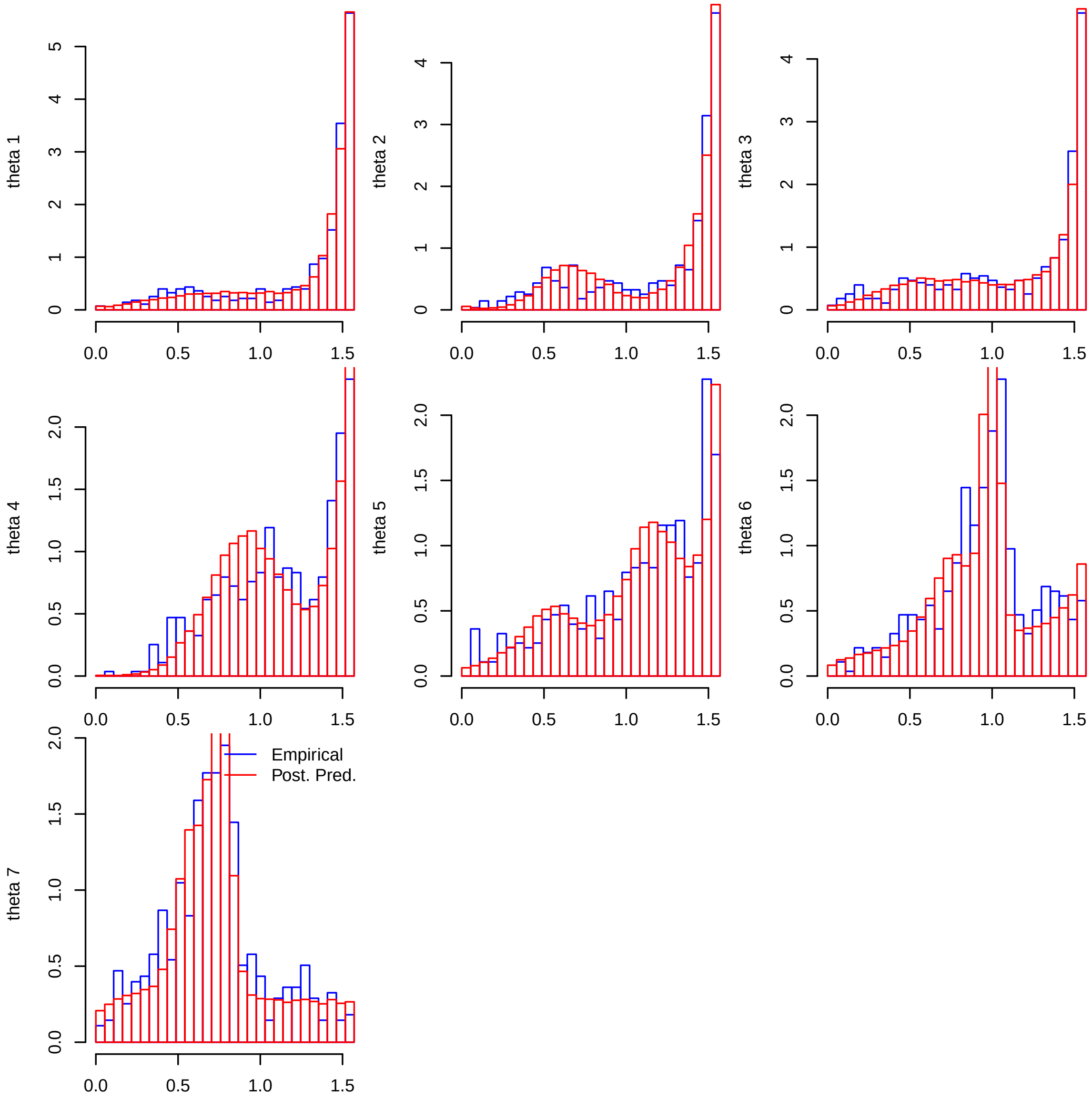


Figure 6: Dirichlet Process mixtue model with multivariate normal kernel over probit space cast on angular representation using declustered IVT dataset


theta 1
theta 2
theta 3
theta 4
theta 5
theta 6
theta 7
Empirical
Post. Pred.