Anomaly Detection in Extremes

Peter Trubey and Bruno Sansó
Department of Statistics, University of California Santa Cruz

October 24, 2023

Abstract

We develop anomaly detection scores leveraging the independence between the radial and angular components of vectors in extreme value settings. The angular density is modeled as a Bayesian non-parametric mixture of projected gammas. The resulting posterior predictive density is used for the angular score. For flexible categorical data modeling, we develop an extension of the projected gammas model using a Dirichlet-multinomial kernel. This is coupled with our proposed anomaly detection score in mixed data settings. We evaluate the anomaly detection efficacy of our proposed scores and find that they are generally superior to tested canonical methods.

Keywords: Bayesian non-parametrics, multivariate exteme values, categorical data, directional statistics,

1 Introduction

Anomaly detection, describes a field of methods for identifying observations as *anomalous*; a term that requires defining. For this paper, following the general trend in the literature, we define anomalies as observations that are in some manner different than non-anomalous data. We interpret this to say that anomalies are data that were not produced by the same generating distribution as non-anomalous data, and as such, we would expect observations found in regions of relative data sparsity to be more likely to be anomalous than those observations found in regions of high data abundance. We characterize this assumption as *anomalies stand apart*. In the literature as here, the term *normal data* is used to refer to data which are not anomalous. Normal data tend to cluster into homogenous groups, but anomalous data are heterogenous in their differences.

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

The complete field of anomaly detection is vast. However, most methods can be roughly grouped into three core ideas: statistical model approaches, non-statistical model ap-

proaches, and clustering methods. Common to all approaches is the assumption that anomalous data will tend to stand apart from normal data.

Statistical models for anomaly detection attempt to model the distribution of data, with the goal of estimating the data density around an observation. In specific applications, one might make assumptions about the parametric form of the generating distribution of the data, but for general application, a non-parametric density estimator is frequently used. This might include algorithms such as k-Nearest Neighbors k-NN (Kramer, 2013); kernel density estimation approaches such as the Parzen-Rosenblatt windowing method (Rosenblatt, 1956; Parzen, 1962); or even semi-parametric density estimation methods, such as Gaussian mixture models (McNicholas, 2010). Local Outlier Factor Breunig et al. (2000) is an example of an anomaly score using a non-parametric density estimator.

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

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

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

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

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

The intersection of extreme value theory and anomaly detection is a current topic of research. Some methods employ univariate EVT on estimated densities calculated via other means, such as Clifton et al. (2011) using a Gaussian Mixture model, and Gu et al. (2021) using a Gaussian process. Both then employ EVT on the estimated densities to establish a decision threshold theoretically, avoiding the process of determining said threshold heuristically. Beyond these applications, the applicability of extreme value theory to anomaly

detection is predicated on the assumption that extreme observations are more likely to be anomalous. A discussion on this point is provided by Goix et al. (2017), stating that extreme observations exist at the border between anomalous and non-anomalous regions. Indeed, for most datasets in our testing, the probability an individual observation is anomalous is higher for data in the tails of the distribution. This relative abundance of anomalies among extremes might cause a naive classifier that does not take into account the dependence structure of extremes to classify all extremes as anomalous. If we follow the assumption that anomalies stand apart, then extreme observations that cluster into a homogenous group should not be considered anomalous. For this reason, we desire a classifier that considers the dependence structure of the extremes as well. Goix et al. (2017) offers one such example. Their method is based on transforming the data to a standard Pareto suing the transformation $T(x) = 1/(1 - \hat{F}(x)) \in [1, \infty)$, where $\hat{F}$ corresponds to the empirical distribution function. Then the the space $[1, \infty)^d$ is partitioned into α-cones, defined as subsets where in each dimension the observations are in excess of a α. α-cones with few observations correspond to lower-density regions, so observations falling into these cones are considered more likely to be anomalous.

A central result of multivariate EVT is that, conditional on an observation being extreme, its radial component—or magnitude—is independent of its angular component. In this paper, following Trubey and Sansó (2022), we fit a Bayesian non-parametric mixture of projected gammas to the angular component, and use samples from its posterior predictive distribution to compute an estimate of the density of the angular component. Direct estimation of density via a fitted model is difficult, owing to the bounded nature of the angular distribution. Instead, we employ non-parametric density estimators including k-nearest neighbors and kernel density estimation to produce estimates of angular density. Further, to expand the applicability of this algorithm, we produce an extension of the BNP projected gamma model to include categorical data. Standing alone, this component

represents a highly flexible density model for categorical data, and it efficiently pairs with the projected gamma model for angular data. We develop several anomaly scoring metrics applicable to the angular data, categorical data, and *mixed* data regimes. The major contributions of this paper are thus three-fold: We develop an anomaly detection algorithm for extreme data that accounts for the dependence structure between extremes, approaching density estimation in a continuous space rather than discrete binning in a partition of the space. We obtain a flexible model for multivariate categorical data that efficiently captures the dependence structure between categories in multiple variables, as well as anomaly scores in this setting. Finally, we provide a model that links the scores developed in these two cases, tackling multivariate observations with components of different types.

The paper proceeds as follows: Section 2 provides a brief review of multivariate EVT, explaining the separation of the radial and angular components of the extreme data, as well as an introduction of the angular data model. Section 3 introduces our anomaly scores for angular data, describing the density estimation methods employed, as well as how radial information is incorporated. Section 4 introduces our flexible categorical data model, along with anomaly scores based on it. Section 4.3 provides a link between the two regimes; anomaly scores that include information from both categorical and angular data. Section 4.4 employs the same rank transformation used in Goix et al. (2017) to apply the angular data model to data not already assumed to be in excess of a threshold, widening the applicability of our metrics. Section 5 provides the resulting performance of our anomaly scores as applied to seven reference anomaly detection datasets, as well as comparing to three canonical anomaly scoring methods. Finally, Section 6 provides concluding remarks and discussion.

2 Review of Extreme Value Theory

The use of methods based on extreme value theory to perform anomaly detection has the advantage of focusing on the tail of the distribution, which is where anomalous observations are more likely to be found. In the univariate case, asymptotic results provide a unique parametric limiting extreme family. Software to infer the parameters of such family is widely available (see, for example, Coles, 2001). A popular approach to study univariate extremes, is to consider the observations that exceed a threshold, and calculate the excess values, then use them for inference on the parameters of the generalized Pareto distribution that correspond to the theoretical limit. This is known as the peaks over threshold approach (PoT), and will be central to the methods in this paper. In the multivariate case the theory for PoT is well developed (see, for example De Haan and Ferreira, 2006), and it indicates the existence of a limiting distribution that has no parametric representation. This presents a challenge for inference. Furthermore, the difficulty of using a PoT approach is compounded by the fact that there is no unique definition of an exceedance of a multivariate threshold, as there is an obvious dependence on the norm used to measure the vectors sizes.

The multivariate PoT model considered in this paper has been developed in Trubey and Sansó (2022), and it is based on a representation of the limiting distribution proposed in Rootzén et al. (2018). Let $\boldsymbol{W} = (W_1, \ldots, W_d)$ be a d-dimensional random vector with cumulative distribution F. Assume that there exist sequences of vectors $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$, and a d-variate distribution G such that $\lim_{n\to\infty} F^n(\boldsymbol{a}_n\boldsymbol{w} + \boldsymbol{b}_n) = G(\boldsymbol{w})$. G is a d-variate generalized extreme value distribution. Then

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

where H is the multivariate Pareto distribution. Rootzén et al. (2018) provides a number of stochastic representations for H. In particular Remark 1 justifies the representation given in Ferreira and de Haan (2014), consisting of taking $\boldsymbol{W}$, in the limit, as $\boldsymbol{W} = R\boldsymbol{V}$

where R and $\boldsymbol{V}$ are independent. $R = ||\boldsymbol{W}||_\infty$ is distributed as a standard Pareto random variable, and $\boldsymbol{V} = \boldsymbol{W}/||\boldsymbol{W}||_\infty$ is a random vector in $\mathbb{S}_\infty^{d-1}$, the positive orthant of the unit sphere in infinite norm, with distribution Φ. R and $\boldsymbol{V}$ are referred, respectively, as the *radial* and *angular* components of H. The angular measure controls the dependence structure of $\boldsymbol{W}$ in the tails. In view of this, to obtain a PoT model we seek a flexible model for the distribution of $\boldsymbol{V} \in \mathbb{S}_\infty^{d-1}$. Our approach consists of a two-step analysis. We first standardize and subsample using a multivariate threshold. Then we estimate the angular measure.

Starting with a collection of observations $\boldsymbol{w}_i \in \mathbb{R}^d, i = 1, \ldots, n$, we perform thresholding for each marginal. We use as threshold $b_{q,l} = \hat{F}_\ell^{-1}(1 - 1/q)$, where $\hat{F}_\ell$ is the empirical distribucion function for the ℓth component. t is chosen to obtain a large empricial quantile, like 85 or 90%. Thresholded values are assumed to follow a generalized univariate Pareto distribution, and are used to estimate the corresponding scale and shape parameters a_ℓ and ξ_ℓ. We then obtain the standardization

$$z_{i\ell} = \left(1 + \xi_\ell \frac{w_{i\ell} - b_\ell}{a_\ell}\right)_+^{1/\xi_\ell} \tag{1}$$

where $(\cdot)_+$ indicates the positive parts function. Let $r_i = ||\boldsymbol{z}_i||_\infty$ and let $\boldsymbol{v}_i = \boldsymbol{z}_i/r_i$. Due to the thresholding, i ranges from 1 to $m \leq n$, and $r_i > 1$. That is, all vectors have at least one very large component.

Recall that the radial component $R \in \mathbb{R}_+$ follows a standard Pareto distribution, we focus on describing the distribution of the angular component $\boldsymbol{V} \in \mathbb{S}_\infty^{d-1}$. For this purpose we use the samples $\boldsymbol{v}_i,\ i = 1, \ldots, m$. A suitable distribution for $\boldsymbol{V}$ can be approximated by projecting a distribution in $\mathbb{R}_+^d$ onto $\mathbb{S}_p^{d-1}$. Recall that the $\mathcal{L}_p$ norm is $\|\boldsymbol{s}\|_p = (\sum_{\ell=1}^d |s|_\ell^p)^{1/p}$. For $\boldsymbol{x} \in \mathbb{R}_+^d$, we define the transformation

$$T_p(\boldsymbol{x}) = (\|\boldsymbol{x}\|_p, x_1/\|\boldsymbol{x}\|_p, \ldots, x_{d-1}/\|\boldsymbol{x}\|_p) =: (r, \boldsymbol{y}),$$

where $\boldsymbol{y} = (y_1, \ldots, y_{d-1}) \in \mathbb{S}_p^{d-1}$, $r > 0$, and $y_d = \left(1 - \sum_{\ell=1}^{d-1} y_\ell^p\right)^{\frac{1}{p}}$. Thus, a distribution

Algorithm 1 Workflow to fit a distribution to data on $\mathbb{S}_\infty^{d-1}$.

1: $b_{t\ell} := \hat{F}_\ell^{-1}\left(1 - \frac{1}{t}\right)$; For $w_{i\ell} > b_{t\ell}$, fit a_ℓ, ξ_ℓ by likelihood-based method.

2: Transform $\boldsymbol{z}_{i\ell} = \left(1 + \xi_\ell \frac{w_{i\ell} - b_{t\ell}}{a_\ell}\right)_+$ for $\boldsymbol{w}_i \not< \boldsymbol{b}_t$.

3: Transform $(r_i, \boldsymbol{v}_i) = \left(\|\boldsymbol{z}_i\|_\infty, \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|_\infty}\right)$.

4: $\boldsymbol{y}_i = \frac{\boldsymbol{z}_i}{\|\boldsymbol{z}_i\|_p}$ is used to facilitate fitting $f(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid \boldsymbol{y})$.

on $\boldsymbol{X}$ induces a distribution on $\mathbb{S}_p^{d-1}$. In particular, assume that $\boldsymbol{X}$ has independent components, each of them distributed as $\mathcal{G}a(\alpha_\ell, \beta_\ell)$, then it can be seen that the density of $\boldsymbol{y}$ is

$$f(\boldsymbol{y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{\ell=1}^{d} \left[\frac{\beta_\ell^{\alpha_\ell}}{\Gamma(\alpha_\ell)} y_\ell^{\alpha_\ell - 1}\right] \left[y_d + \textstyle\sum_{\ell=1}^{d-1} y_\ell^p y_d^{1-p}\right] \frac{\Gamma(\sum_{\ell=1}^{d} \alpha_\ell)}{\left(\sum_{\ell=1}^{d} \beta_\ell y_\ell\right)^{\sum_{\ell=1}^{d} \alpha_\ell}}, \tag{2}$$

(see, Trubey and Sansó, 2022, for details). This distribution is a generalization to $\mathbb{S}_p^{d-1}$ of the projected gamma distribution defined for $p = 2$ in Núñez-Antonio and Geneyro (2019). We will denote its density as $\mathcal{PG}_p(\boldsymbol{y}|\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Note that T is not differentiable at $p = \infty$, so we can not use it to directly model on $\mathbb{S}_\infty^{d-1}$. However, as p increases, the surface $\mathbb{S}_p^{d-1}$ will approach $\mathbb{S}_\infty^{d-1}$. This means that a projected gamma built on $\mathbb{S}_p^{d-1}$ with a sufficiently large p will serve to approximate a distribution on $\mathbb{S}_\infty^{d-1}$. This approximation is leveraged in Trubey and Sansó (2022) to obtain samples of a distribution on $\mathbb{S}_\infty^{d-1}$. To obtain a flexible model for Φ we use the projected gamma density as the kernel of a random measure mixture model, based on the Pitman Yor ($\mathcal{PY}$) process introduced in Perman et al. (1992). Pitman-Yor processes are fully atomic random measures that are specified by two parameters and a centering distribution. They can be formulated, using a stick-breaking representation (Ishwaran and James, 2001a), as

$$\Pr(\alpha \mid \cdots) = \sum_{j=1}^{\infty} p_j \delta_{\boldsymbol{\alpha}_j}; \quad \sum_{j=1}^{\infty} p_j = 1, \quad p_j := \chi_j \prod_{k=1}^{j-1} (1 - \chi_k)$$

where $\delta_{\boldsymbol{\alpha}_j}$ indicates a point mass at $\boldsymbol{\alpha}_j$, and $\boldsymbol{\alpha}_j$ are sampled independently from G_0. $\chi_j \sim \text{Beta}(1 - d, \eta + jd)$. $d \in [0, 1), \eta > -d$ are refereed to as the discount and the

concentration parameters, respectively. Pitman Yor processes have the advantage over the more commonly used Dirichlet processes (Ferguson, 1974) of including a discount parameter along with the concentration parameter, allowing greater control over the formation of new clusters. A hierarchical formulation of the model for observations $\boldsymbol{y}_i \in \mathbb{S}_p^{d-1}$, $i = 1, \ldots, n$, is

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\alpha}_i &\sim \mathcal{PG}_p\left(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_i, \mathbf{1}\right) & G_0 &= \mathcal{LN}_d\left(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Sigma\right) \\ \boldsymbol{\alpha_i} &\sim G & \mu &\sim \mathcal{N}_d\left(\mathbf{0}, \mathbf{1}\right) \\ G &\sim \mathcal{PY}\left(d, \eta, G_0\right) & \Sigma &\sim \mathcal{IW}_d\left(\nu, \Psi\right). \end{aligned} \tag{3}$$

Here $\mathcal{LN}$ denotes a log-normal, $\mathcal{N}$ a normal, and $\mathcal{IW}$ an inverse Wishart. We refer to this model as a *Pitman–Yor mixture of projected gammas* ($\mathcal{PYPG}$). Details of the implementation of this model using and adaptive MCMC approach are provided in the supplementary material. As a kernel density, it was observed in Trubey and Sansó (2022) that the unrestricted form of the $\mathcal{PG}_p$ with both shape and rate parameters offered no improvement in model fidelity on real data compared to the restricted form, where the rate parameters are fixed at 1. For a more parsimonious model, and for compatibility with the categorical model that will be developed in Section 4, in this paper we choose to use the restricted form.

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

$$\chi_j \mid n_j, n_{k>j} \sim \text{Beta}\left(1 + n_j - d, \eta + \sum_{k=j+1}^{J} n_k + jd\right); \qquad p_j := \chi_j \prod_{k=1}^{j-1}(1 - \chi_k) \tag{4}$$

where n_j is the number of observations in cluster j. In this form, the Dirichlet process is a special case of the Pitman-Yor process where the discount parameter $d := 0$. Then the

cluster identifier for observation i, φ_i, is sampled as

$$\Pr\left[\varphi_i = j \mid \boldsymbol{p}, \boldsymbol{\alpha}\right] = \frac{p_j \, \mathcal{PG}_p\left(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_j, \mathbf{1}\right)}{\sum_{k=1}^{J} p_k \, \mathcal{PG}_p\left(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_k, \mathbf{1}\right)}. \tag{5}$$

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

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

3 Novelty Detection Methods

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

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

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

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

That is,

$$\mathcal{S}_i = \mathcal{S}_{i,r} \times \mathcal{S}_{i,\boldsymbol{v}} = f_r(r_i)^{-1} \times \left[\int_\Omega f_v(\boldsymbol{v}_i \mid \boldsymbol{\alpha}) \, \mathrm{d}G(\boldsymbol{\alpha} \mid \mathcal{D}) \right]^{-1}$$

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

To avoid the problems of angular density evaluation in $\mathbb{S}_\infty^{d-1}$ we use a non-parametric angular density estimator based on a sample from the posterior predictive distribution of the model developed in Section 2. Here, we consider two well-established methods: k–nearest neighbors, or *kNN* (Mack and Rosenblatt, 1979), and kernel density estimation, or *KDE* (Parzen, 1962). For both of these methods we make use of pairwise distances between observations from the dataset, and replicates from a posterior predictive sample.

As described in Trubey and Sansó (2022), the geodesic distance on $\mathbb{S}_\infty^{d-1}$ is expensive to evaluate as the computational burden grows combinatorically with the number of dimensions. As an alternative they propose an estimate of distance that is computationally cheap to evaluate, bearing a cost equivalent to that of a Euclidean norm. Let

$\mathbb{C}_\ell^{d-1} = \{\boldsymbol{x} : \boldsymbol{x} \in \mathbb{S}_\infty^{d-1}, x_\ell = 1\}$ comprise the ℓth *face* of $\mathbb{S}_\infty^{d-1}$. For a pair of points on the same face, the Euclidean distance corresponds to the geodesic, or length of the shortest possible path between those two points. For a pair of points $\boldsymbol{a} \in \mathbb{C}_\ell^{d-1}$, $\boldsymbol{b} \in \mathbb{C}_\jmath^{d-1}$, we can rotate $\mathbb{C}_\jmath^{d-1}$ into the same hyperplane as $\mathbb{C}_\ell^{d-1}$. Transform $\boldsymbol{b}$ such that

$$\boldsymbol{b}' = P_{\jmath\ell}(\boldsymbol{b}) = \begin{cases} b_i & \text{for } i \neq \jmath, \ell \\ 1 & \text{for } i = \ell \\ 2 - b_\ell & \text{for } i = \jmath \end{cases} \qquad g(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}'\|_2 \tag{6}$$

After transformation, the Euclidean norm between $\boldsymbol{a}$ and $\boldsymbol{b}'$ corresponds to a negative definite kernel that provides an upper bound on geodesic distance on $\mathbb{S}_\infty^{d-1}$ between $\boldsymbol{a}$ and $\boldsymbol{b}$.

3.1 k-Nearest Neighbors Density estimation

We use the kernel g defined in Equation 6 to obtain a local posterior predictive density based on a kNN estimator on $\mathbb{S}_\infty^{d-1}$. To this end we consider a locally uniform density within a $d-1$–dimensional ball $\mathbb{B}$, centered on observation $\boldsymbol{v}_i$. The radius $D_k(\boldsymbol{v}_i)$ is calculated as $g\left(\boldsymbol{v}_i, \boldsymbol{v}^*_{N_k(i)}\right)$, where $\boldsymbol{v}^*_{N_k(i)}$ is the kth nearest neighbor of $\boldsymbol{v}_i$ in a sample from the posterior predictive distribution. The volume of the ball is calculated as

$$\text{Vol}(\mathbb{B}_k^{d-1}) = \frac{\pi^{\frac{d-1}{2}} D_k(\boldsymbol{v}_i)^{d-1}}{\Gamma\left(\frac{d-1}{2} + 1\right)}. \tag{7}$$

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

$$\mathcal{S}_{i,\boldsymbol{v}}^{k\text{nn}} = \frac{N\pi^{\frac{d-1}{2}} D_k(\boldsymbol{v}_i)^{d-1}}{k\Gamma\left(\frac{d-1}{2} - 1\right)} \tag{8}$$

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

analysis.

3.2 Kernel Density Estimation

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

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

where $\boldsymbol{x}_k^*$ are random replicates from F. The choice of kernel function $\mathcal{Q}$, and selection of the bandwidth parameter h are both topics that have been extensively researched. In practice the Gaussian kernel seems to be well regarded for its simplicity, flexibility, and interpretability. The bandwidth parameter in this case corresponds to the standard deviation of the kernel function. The multivariate Gaussian kernel is more flexible, accepting a matrix as the bandwidth parameter. A larger bandwidth serves to smooth the resulting density estimate, where a lower bandwidth is more responsive to individual observations of data. Optimization of h is application and data specific, but there do exist various *rules of thumb* based on summary statistics of the data. For our analysis, we are making use of a distance analogue on $\mathbb{S}_\infty^{d-1}$ described in Equation (6), which precludes the ability to describe bandwidth using a matrix. We therefore consider the univariate case of f in kernel space, where $\|x - x^*\|$ has been replaced with $g(\boldsymbol{v}, \boldsymbol{v}^*)$.

For selection of the bandwidth parameter h, we employ Silverman's rule of thumb (Silverman, 2018), estimating $\hat{h} = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}}\hat{\sigma}$. This then requires the estimation of $\hat{\sigma}$, which in this case we calculate from pairwise distances. Recall that for a random variable X, $\mathrm{E}\left[\|X_j - X_k\|_2\right] = 2\mathrm{Var}(X)$. In that case, $\hat{\sigma} = \sqrt{\frac{1}{2N(N-1)}\sum_{j\neq k} g(\boldsymbol{v}_j^*, \boldsymbol{v}_k^*)}$, where $\boldsymbol{v}_j^*, \boldsymbol{v}_k^*$ are replicates from the posterior predictive distribution. Then $\hat{\sigma}$ is used in the aforementioned rule of thumb for h. Finally, the angular score under KDE is then

calculated as

$$\mathcal{S}_{i,\boldsymbol{v}}^{\text{kde}} = \mathrm{E}_{\boldsymbol{v}^*}\left[\exp\left\{-\left(\frac{g(\boldsymbol{v}_i, \boldsymbol{v}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\exp\left\{-\left(\frac{g(\boldsymbol{v}_i, \boldsymbol{v}_k^*}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{9}$$

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

Algorithm 2 Workflow for anomaly detection on $\mathbb{S}_\infty^{d-1}$.

1: Take r_i, $\boldsymbol{y}_i$ according to Algorithm (1)

2: Fit $\boldsymbol{y} \sim \mathcal{PYPG}$ from Equation (3)

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

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

5: Take $\mathcal{S}_{i,v}$ as per Equations (8,9)

4 Binary and Categorical Data

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

Suppose $\boldsymbol{C}$ is a vector of M random categorical variables. Then C_m is a random categorical variable, with $K_m \geq 2$ categories. Regard $\boldsymbol{W}_m$ as C_m, recoded in one-hot, or multinomial, encoding, and $\boldsymbol{W}$ the concatenation of M one-hot encoded categorical RV's. That is, $\boldsymbol{W}$ is a binary vector of length $K = \sum_{m=1}^{M} K_m$; $\sum_{k=1}^{K} W_k = M$; and

every m subset of $\boldsymbol{W}$ sums to 1. To account for over-dispersion, we consider a Dirichlet-multinomial density for $\boldsymbol{W}_m$. Recall that the Dirichlet distribution is a special case of projected gamma, projected onto $\mathbb{S}_1^{d-1}$, with rate parameters uniformly fixed as $\beta_\ell = \beta = 1$ by convention. We consider a Dirichlet-multinomial distribution, $\mathcal{DM}$ that is obtained by integrating out the latent categorical probability vector from the product of Dirichlet and multinomial distributions; $\mathcal{DM}(\boldsymbol{w} \mid \boldsymbol{\alpha}) = \int_\pi \mathcal{M}(\boldsymbol{w} \mid \pi)\mathcal{D}(\pi \mid \boldsymbol{\alpha})d\pi$. Recalling that a categorical random variable can be considered as a multinomial with size 1, we can further simplify the Dirichlet-multinomial to a Dirichlet-categorical, reducing the computational burden. Thus,

$$\boldsymbol{w}_m \mid \boldsymbol{\alpha}_m \sim \mathcal{DC}(\boldsymbol{w}_m \mid \boldsymbol{\alpha}_m) = \frac{\Gamma(\sum_{\ell=1}^d \alpha_{m\ell})}{\Gamma(1 + \sum_{\ell=1}^d \alpha_{m\ell})} \prod_{\ell=1}^d \frac{\Gamma(w_{m\ell} + \alpha_{m\ell})}{\Gamma(\alpha_{m\ell})} \tag{10}$$

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

$$\begin{aligned} \boldsymbol{w}_i \mid \boldsymbol{\alpha}_i &\sim \mathcal{CDC}\left(\boldsymbol{w}_i \mid \boldsymbol{\alpha}_i\right) & G_0 &= \mathcal{LN}\left(\boldsymbol{\alpha} \mid \boldsymbol{\mu}, \Sigma\right) \\ \boldsymbol{\alpha_i} &\sim G & \mu &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{1}\right) \\ G &\sim \mathcal{PY}\left(d, \eta, G_0\right) & \Sigma &\sim \mathcal{IW}\left(\nu, \Psi\right). \end{aligned} \tag{11}$$

Note that there exists a strong negative covariance between categories within a categorical variable. To account for this in our proposed prior, the parameter Ψ is chosen as a block diagonal matrix, with each m block corresponding to the mth categorical variable. Setting the value of the diagonal to ψ_0, the off-diagonals within the m block are set to $-\psi_0 d_m^{-2}$ where d_m is the number of categories in the mth categorical variable. This value corresponds to the covariance of a categorical variable where all category probabilities are equal. In addition to the proposed log-normal model, we investigated using a product of gammas as the centering distribution in Equation (11), but we observed that this choice induces numerical instability. We observed that the log-normal distribution, with its squared exponential

tails and ability to account for negative covariance within the prior, provided stable model fitting.

4.1 Anomaly Detection Methods for Categorical Data

Anomaly scores analogous to the ones proposed in Section 3 can be obtained for categorical variables by transforming the latent variables that define a Dirichlet-Multinomial distribution on $\mathcal{S}_1^{d-1}$ to $\mathcal{S}_\infty^{d-1}$. We start by considering the cluster identifiers. Extrapolating Equation (5) to the categorical model, cluster identifiers φ_i are sampled with probabilities

$$\Pr[\varphi_i = j \mid \boldsymbol{\alpha}, \boldsymbol{p}, \boldsymbol{w}_i] = \frac{p_j \, \mathcal{CDC}\left(\boldsymbol{w}_i \mid \boldsymbol{\alpha}_j\right)}{\sum_{k=1}^{J} p_k \, \mathcal{CDC}\left(\boldsymbol{w}_i \mid \boldsymbol{\alpha}_k\right)} \quad \text{for } j = 1, \ldots, J. \tag{12}$$

For a given sample from the posterior for $\boldsymbol{\alpha}$, first we sample φ_i, then let $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_{\varphi_i}$, and sample

$$\boldsymbol{\varrho}_i \mid \boldsymbol{\alpha}_i \sim \prod_{\ell=1}^{d} \mathcal{G}\left(\varrho_{i\ell} \mid \alpha_{i\ell}, 1\right). \tag{13}$$

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

We develop four methods based on applications of the KNN and KDE metrics previously described. Making an abuse of notation for simplicity of presentation, let $\tilde{\mathrm{E}}\left[\boldsymbol{\nu}_i\right] := T_\infty(\mathrm{E}\left[\boldsymbol{\nu}_i \mid \boldsymbol{w}_i\right])$, the projection of the expectation of $\boldsymbol{\nu}_i$ back onto $\mathbb{S}_\infty^{d-1}$. Evaluating this expectation by Monte Carlo approximation is equivalent calculating the spherical mean

(Mardia et al., 1999), which takes the arithmetic mean of observations in Cartesian coordinates, then projects back onto the sphere.

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

$$\mathcal{S}_{i,\boldsymbol{\nu}}^{\mathrm{hknn}} = \frac{N\,\pi^{\frac{d-1}{2}}}{k\,\Gamma\left(\frac{d-1}{2}+1\right)}\,D_k\left(\tilde{\mathrm{E}}[\boldsymbol{\nu}_i]\right)^{d-1} \tag{14}$$

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

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

$$\mathcal{S}_{i,\boldsymbol{\nu}}^{\mathrm{hkde}} = \mathrm{E}_{\boldsymbol{\nu}^*}\left[\exp\left\{-\frac{1}{2}\left(\frac{g(\tilde{\mathrm{E}}[\boldsymbol{\nu}_i],\boldsymbol{\nu}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K}\sum_{k=1}^{K}\exp\left\{-\frac{1}{2}\left(\frac{g(\tilde{\mathrm{E}}[\boldsymbol{\nu}_i],\boldsymbol{\nu}_k^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{15}$$

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

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

$$\mathcal{S}_{i,\boldsymbol{\nu}}^{\mathrm{lhkde}} = \mathrm{E}_{\boldsymbol{\nu}^*,\boldsymbol{\nu}_i}\left[\exp\left\{-\frac{1}{2}\left(\frac{g(\boldsymbol{\nu}_i,\boldsymbol{\nu}^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \approx \left[\frac{1}{K_{\boldsymbol{\nu}_i}K_{\boldsymbol{\nu}^*}}\sum_{j=1}^{K_{\boldsymbol{\nu}_i}}\sum_{k=1}^{K_{\boldsymbol{\nu}^*}}\exp\left\{-\frac{1}{2}\left(\frac{g(\boldsymbol{\nu}_{i,j},\boldsymbol{\nu}_k^*)}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{16}$$

Computing this for a given sample is more expensive than *hkde* due to the double sum.

However, plugging in an estimate of $\mathrm{E}[\boldsymbol{v}_{\boldsymbol{w}_i}]$ removes a significant degree of uncertainty around the distribution of $\boldsymbol{v}_{\boldsymbol{w}_i}$, which may be relevant.

If, instead of projecting the unnormalized probability vectors onto a unified hypersphere $\mathbb{S}_\infty^{d-1}$, we normalize each m-component onto its associated simplex, $\mathbb{S}_1^{d_m-1}$. Using Manhattan distance on the simplex, we obtain the latent simplex KDE *(lskde)*.

$$\mathcal{S}_{i,\boldsymbol{\pi}}^{\text{lskde}} = \mathrm{E}_{\boldsymbol{\pi}_i,\boldsymbol{\pi}^*}\left[\exp\left\{-\frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_i - \boldsymbol{\pi}^*\|_1}{\hat{h}}\right)^2\right\}\right] \approx \left[\frac{1}{K_{\boldsymbol{\pi}^*}K_{\boldsymbol{\pi}_i}}\sum_{j=1}^{K_{\boldsymbol{\pi}_i}}\sum_{k=1}^{K_{\boldsymbol{\pi}^*}}\exp\left\{-\frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_{ij} - \boldsymbol{\pi}_k^*\|}{\hat{h}}\right)^2\right\}\right]^{-1} \tag{17}$$

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

Algorithm 3 Workflow for anomaly detection for categorical data

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

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

3: Fit $\boldsymbol{w} \sim \mathcal{PYCDC}$ from Equation (11)

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

5: From $\boldsymbol{\alpha}_i \mid \boldsymbol{w}_i$ sampled as per Equations (12-13)

sample $\boldsymbol{\varrho}_{ik} \mid \boldsymbol{\alpha}_i \sim \prod_\ell \mathcal{G}(\alpha_\ell)$ for $k = 1, \ldots, K_{\nu_i}$

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

7: Take $\mathcal{S}_{i\boldsymbol{v}}$ as per Equations (14–17)

4.2 Mixed Models

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

$$(\boldsymbol{y}, \boldsymbol{w}) \sim \int_{\boldsymbol{\alpha}} \mathcal{PG}_p(\boldsymbol{y} \mid \boldsymbol{\alpha}_{\boldsymbol{y}}, \mathbf{1})\, \mathcal{CDM}(\boldsymbol{w} \mid \boldsymbol{\alpha}_{\boldsymbol{w}})\, dG(\boldsymbol{\alpha}) \tag{18}$$

with the distribution of $\boldsymbol{\alpha} = (\boldsymbol{\alpha_y}, \boldsymbol{\alpha_w})$ as defined in Equation 11. The dimensions are, respectively, $d_{\boldsymbol{y}}$ and $d_{\boldsymbol{w}}$. Note that for the projected gamma distribution, we restrict the rate parameters to $\beta_\ell := 1$. Also note that for the mixed model, the hyperparameter for the covariance matrix $\Sigma_{\boldsymbol{\alpha}}$ is taken as a blocked diagonal matrix, with the block corresponding to the angular component being a diagonal matrix.

4.3 Mixed Model Anomaly Scores

Let $d = d_{\boldsymbol{y}} + d_{\boldsymbol{w}}$ be the total number of dimensions. Then, for the mixed model, let $\boldsymbol{\nu}_i = T_\infty(R_i \boldsymbol{y}_i, \boldsymbol{\varrho}_{i\boldsymbol{w}})$, and $\boldsymbol{\nu} = T_\infty(\boldsymbol{\varrho})$. The *hknn* score can be adapted to the mixed model by re-projecting the angular data and the latent categorical component into the same sphere. This requires moving $\boldsymbol{y}_i$ back to $\mathbb{R}_+^{d_{\boldsymbol{y}}}$, by multiplying by the radial component generated according to

$$R_i \mid \boldsymbol{\alpha} \sim \mathcal{G}\left(R_i \,\middle|\, \sum_{\ell=1}^{d_{\boldsymbol{y}}} \alpha_\ell, 1\right). \tag{19}$$

Then $\boldsymbol{\nu}_i = T_\infty(R_i \boldsymbol{y}_i, \boldsymbol{\varrho}_{i\boldsymbol{w}})$ is the latent projection of both the real component and categorical component into the same sphere. Also, let $\boldsymbol{\nu} = T_\infty(\boldsymbol{\varrho})$ be the generic $\boldsymbol{\nu}$ not specifically dependent on observation i. To obtain the corresponding anomaly scores we can proceed by using the expression in equations (14), (15) and (16) All three scores seek a unifying approach for all data, projecting onto a the same sphere, and calculating a consistent distance metric. An alternative is to, instead, evaluate distances between angular data their own space, and, separately, latent posterior class probabilities in their own space, with the appropriate distance metric for each. In effect, this approach combines *hkde* from the angular component and *lskde* from the categorical component yielding:

$$\begin{aligned} \mathcal{S}_{i,\boldsymbol{v}}^{lmkde} &= \mathrm{E}_{\boldsymbol{v}^*,\boldsymbol{\pi}^*,\boldsymbol{\pi}_i}\left[\exp\left\{-\frac{1}{2}\left(\frac{d\left(\boldsymbol{v}_i, \boldsymbol{v}^*\right)}{\hat{h}_{\boldsymbol{v}^*}}\right)^2 - \frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_i - \boldsymbol{\pi}^*\|_1}{\hat{h}_{\boldsymbol{\pi}^*}}\right)^2\right\}\right]^{-1} \\ &\approx \left[\frac{1}{K_{\boldsymbol{\pi}^*} K_{\boldsymbol{\pi}_i}} \sum_{j=1}^{K_{\boldsymbol{\pi}_i}} \sum_{k=1}^{K_{\boldsymbol{\pi}^*}} \exp\left\{-\frac{1}{2}\left(\frac{d(\boldsymbol{v}_i, \boldsymbol{v}_k^*)}{\hat{h}_{\boldsymbol{v}^*}}\right)^2 - \frac{1}{2}\left(\frac{\|\boldsymbol{\pi}_{ij} - \boldsymbol{\pi}_k^*\|_1}{\hat{h}_{\boldsymbol{\pi}^*}}\right)^2\right\}\right]^{-1} \end{aligned} \tag{20}$$

This choice to evaluate each component within its own space presents some loss of information as to the dependence structure between $\boldsymbol{y}$ and $\boldsymbol{w}$ within the score. We will explore to what extent that loss of information is relevant.

Algorithm 4 Workflow for anomaly detection for *mixed* data

1: Take r_i, $\boldsymbol{y}_i$ according to Algorithm (1); $\boldsymbol{w}_i$ as in Algorithm 3.

2: Fit $(\boldsymbol{y}, \boldsymbol{w})$ using mixed model from Equation (18)

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

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

5: From $\boldsymbol{\alpha} \mid \boldsymbol{y}_i, \boldsymbol{w}_i$: sample R_i according to Equation (19), $\boldsymbol{\varrho}_{iw}$ similar to Algorithm 3.

6: Take $\boldsymbol{\nu}_i = T_{infty}(R_i \boldsymbol{y}_i, \boldsymbol{\varrho}_{iw})$; $\boldsymbol{\nu}^* = T_\infty(\boldsymbol{\varrho}^*)$.

7: Apply Score function.

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

9: From $\boldsymbol{\alpha} \mid \boldsymbol{y}_i, \boldsymbol{w}_i$, sample $\boldsymbol{\varrho}_{iw}$ similar to Algorithm 3.

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

11: Apply Score function.

12: **end if**

4.4 Relaxing the assumption of independence

A valid critique of the model presented thus far is that in order to justify modelling the radial component of $\boldsymbol{Z}$ as independent to its angular component—the fundamental result of the multivariate extreme value theory presented—it is necessary to subset data to those observations $\boldsymbol{X}$ which exceeded a large threshold in at least one dimension. For some applications, this represents a very powerful data reduction with little loss of information pertaining to anomalies, as anomalies tend to be in the tails (see, for example, Table 1). For other applications, this data reduction represents a significant loss of information about possible anomalies not corresponding to the tails. For this second group, one available avenue

is to relax the assumption of independence between the angular and radial components.

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

$$(\boldsymbol{y}_i, \boldsymbol{w}_i, r_i) \sim \int_{\boldsymbol{\alpha}} \mathcal{PG}_p(\boldsymbol{y}_i \mid \boldsymbol{\alpha}_{\boldsymbol{y}}, \mathbf{1})\, \mathcal{CDM}(\boldsymbol{w}_i \mid \boldsymbol{\alpha}_{\boldsymbol{w}})\, \mathcal{P}(r_i \mid \alpha_r)\, \mathrm{d}G(\boldsymbol{\alpha}) \tag{21}$$

As $\alpha_r > 0$, we augment the kernel parameters to $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{\boldsymbol{y}}, \boldsymbol{\alpha}_{\boldsymbol{w}}, \alpha_r)$, and use a joint log-normal as the center of the random measure prior for G. The scores developed previously in Section 4.3 remain applicable.

5 Results

As mentioned in 3, our goal is to produce novelty scores to rank observations according to how likely they are of being anomalous. This creates another problem: threshold selection—anomaly scores beyond what level are determined anomalous? We mentioned Clifton et al. (2011) and Gu et al. (2021) as examples of computing thresholds theoretically, but in general, thresholds are determined heuristically, using performance criteria. In some applications, heuristic determination can be extremely costly.

One such criteria is the receiver operating characteristics, or *ROC*, curve. For a given score threshold, one can compute the true positive rate, or TPR, as the number of anomalous observations with scores above the threshold, divided by the total number of anomalous observations. The false positive rate, or FPR, is similarly the number of non-anomalous observations above the threshold, divided by the total number of non-anomalous observa-

tions. The ROC curve is formed as the TPR is plotted on the vertical axis against the FPR on the horizontal axis for a range of possible thresholds. The curve is non-decreasing, starting at the origin $(0, 0)$, and ending at unity $(1, 1)$. Threshold selection might include specifying an acceptable FPR, and determining the threshold that produces that FPR.

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

In developing our model, we employ the blocked Gibbs sampler for stick-breaking priors detailed in Ishwaran and James (2001b). We set a discount factor of 0.1, and a concentration parameter of 1.0. In our testing, in the neighborhood around these values we found the resultant number of extant clusters to be relatively stable. We use $(\mu_0 = \mathbf{0}_d, \Sigma_\mu = \boldsymbol{I}_d)$ as prior parameters for μ, and $(\nu = d + 50, \Psi = \nu I_d)$ as prior parameters for Σ, except for the categorical components of the shape vector as described in Section 4. Deviations in μ_0 towards the negative direction bias the model towards asymptotic independence, which in our testing resulted in lower model fidelity. To update the cluster shape vectors, we employ a joint proposal step in log-space using a multivariate normal proposal, where the proposal covariance is informed with an adaptive Metropolis algorithm.(Haario et al., 2001). To hasten updates to the shape parameters, and speed convergence of the model, we employ a parallel tempering algorithm where parallel MCMC chains are sampled at an ascending temperature ladder, where density is exponentiated to the reciprocal of the chain temperature t: $f_t(\theta) = f(\theta)^{1/t}$. Chains with higher temperatures have flatter posteriors, and thus more readily move around the parameter space. Chain states are randomly exchanged via a Metropolis step with probability $p_{1,2} = \exp\left\{(t_2^{-1} - t_1^{-1})(L_2 - L_1)\right\}$, where L refers to the *energy*, or log-density of the chain at its current state. The sample history of the cold chain, where $t := 1$, is preserved as draws from the posterior distribution.

Table 1: Characteristics of datasets used in the analysis. For a given model, N and A refer to the number of observations and anomalies in the fitting set, respectively. M identifies the number of categorical variables, with $d_{\boldsymbol{v}}$ and $d_{\boldsymbol{w}}$ identifying the total number of real and categorical columns respectively. t is the time (in hours) to fit the model. Discrepancy in d between peaks–over–threshold and rank–transformation reflects differences in data transformation, as well as the additional column for the radial component in the rank–transformed model.

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

For each example dataset, the sampler was ran for $50,000$ iterations, discarding the first 40,000 as burn-in. The resulting chain was thinned, keeping only every 10th iteration. For evaluating density under the posterior predictive distribution, we generate 10 replicates from each iteration kept.

We compared our four proposed scores against three canonical novelty detection algorithms, including isolation forest *(iso)* Liu et al. (2008), local outlier factor (lof) Breunig et al. (2000), and one-class SVM (svm) Chang and Lin (2011). Each dataset was subject to 5-fold cross-validation, and out-of-sample performance scores were averaged to produce the resulting performance tables seen in this section. This additional step of cross–validation turned out to be unnecessary for our model, as out–of–sample performance did not markedly differ from in–sample or full–sample performance for the tested datasets. Table 1 provides a summary description of the datasets used in the analysis. For larger datasets, we subset-

ted the raw data to reduce computation time for the rank–transformation and categorical applications. Note that the categorical versions of *cover*, *pima*, and *yeast* are created from discretizing the rank–transformation subsets. First, we present score efficacy on our purely categorical data model, then mixed scoring with thresholding on continuous variables. Finally we present mixed scoring on rank-transformation data.

5.1 Categorical anomalies

The categorical transformation of *cover*, *pima*, and *yeast* discretized the real-valued and ordinal variables in those datasets. For *cover* in particular, it seems this transformation lost a significant amount of data. From Table 1, it seems a large portion of data regarding anomalies is contained within the radial component, so a categorical transformation loses that information. Likely for this reason, none of the methods offer exceptional performance on this dataset. The dataset *solarflare* was also unique in our analysis, being the only truly

Table 2: Area under the *ROC* curve for various anomaly detection schemes, on *strictly categorical* datasets. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

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

categorical dataset used. Our algorithm *lskde* very slightly trailed the performance of *one-class SVM*, the best performing algorithm on this dataset. On both *pima* and *yeast*, latent-simplex KDE performed significantly better than any of the canonical methods. On this

analysis, *hkde* and *hknn* both performed poorly. It seems the projection of the categorical probability vectors into a unified sphere induces some loss of information.

5.2 Peaks-over-Threshold anomalies

We subjected six datasets to multivariate thresholding, only keeping observations that exceeded the threshold in at least one dimension. Table 1 indicates what quantile was used for the threshold, as well as the number of anomalies in excess of the threshold. For *cover*, we further sub-sampled the excesses to produce a more manageable sized dataset. For variables that did not exhibit properties that would allow for a peak-over-threshold model to apply, these variables were instead converted to discrete values with two or three categories. We built the mixed data model, and evaluated performance of the mixed scores, compared against the canonical methods. Of particular interest here is the *annthyroid* dataset, for which all of our scores performed comparably, and significantly better than the canonical scores. Of the other tested datasets, on *cardio*, *lmkde* approached the performance of *isolation forest* and *one-class SVM*, but all other methods performed worse. For the datasets *cover* and *mammography*, *hknn*, *lhkde*, and *lmkde* performed comparably, and each significantly better than any of the canonical methods. We see that *lmkde*, being the inheritor of the latent simplex KDE score, performs reasonably well reliably among datasets thus far in the peaks-over-threshold setting, but is outperformed by other metrics on each dataset. We may see some effect of the loss of information relating to the dependence structure between $\boldsymbol{w}$ and $\boldsymbol{v}$ on the derived performance. On that note, *lhkde* performed comparably to *lmkde* on *annthyroid*, *cover*, *pima*, and *yeast*, but slightly exceeded its performance on *mammography*. We saw in the categorical datasets, *lskde* performed generally well, so the projection onto a unified sphere may induce loss of information. In that regard, it may be the case that preserving information about the dependence structure between $\boldsymbol{v}$ and $\boldsymbol{w}$ had a greater effect than a greater effect than preserving information within $\boldsymbol{w}$ specifically.

Table 3: Area under the *ROC* curve for various anomaly detection schemes, on *mixed* data where the real component has undergone the *threshold* standard Pareto transformation. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

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

As to the poor performance of every method on *pima* and *yeast*, these reported *AuROC* values are conditional on the data exceeding the multivariate threshold used in building the model. As we see in Table 1, these datasets do not meet the assumption that anomalies are concentrated in the tails. Scores depending on r_i, the radius component of $\boldsymbol{z}_i$, or *magnitude* of the extremal observation, are going to perform poorly relative to metrics that do not make that assumption.

5.3 Rank Transformation anomalies

We subjected the same six datasets used in the peak-over-threshold model to rank transformation on the real and ordinal variables. We then built the mixed model including radius described in Section 4.4 on the transformed datasets. Large datasets used in rank-transformation and categorical models were sub-sampled to reduce computation time. Note that rank transformation preserves the entire dataset, so we should not consider the values

in Table 4 to be comparable to the values in Table 3.

Table 4: Area under the *ROC* curve for various anomaly detection schemes, on *mixed* data where the real component has undergone the *rank* standard Pareto transformation. Reported here is arithmetic mean of out-of-sample performance for 5-fold cross-validation. Values closer to 1 are preferred.

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

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

6 Conclusion

In this paper, we have proposed a method of *scoring* observations as anomalous based on their posterior-predictive angular density, using the result from multivariate extreme value theory that—assuming the existence of a limiting behavior—given observations are in excess of a high threshold, after transformation their angular distribution on $\mathbb{S}_\infty^{d-1}$ is independent of the radial distribution on $\mathbb{R}_+$. In the anomaly detection setting, this independence

allows us to separate anomaly scores into an angular and radial component, and treat them separately. To define an angular anomaly score, a Bayesian non-parametric model is developed on the angular data projected onto $\mathbb{S}_p^{d-1}$, and as a true density on $\mathbb{S}_\infty^{d-1}$ is not available, anomaly scores are obtained using a non-parametric estimator to that angular density built on a sample from the posterior predictive distribution of the fitted model. The non-parametric estimators we used were k-nearest neighbors, and kernel density estimation.

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

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

In this paper, we have presented a highly flexible model–based method for anomaly detection that scales to moderately large dimensions and sample sizes. However, as seen in

Table 1, even for the dimensions and sample sizes presented, model fitting can take several hours. Scaling this model beyond some thousands of observations or tens of columns will require a paradigm shift in *how* the model is fit. For this reason, we are investigating faster means of model fitting, including a variational approach.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Trubey, P. and B. Sansó (2022). Inference for multivariate peaks over threshold models. Technical Report UCSC-SOE-22-02, University of California Santa Cruz.