RESEARCH ARTICLE

WILEY

Tracking clusters and anomalies in evolving data streams

Sreelekha Guggilam[1] | **Varun Chandola[1,2]** | **Abani Patra[1,3]**

[1]Computational Data Science & Engineering, University at Buffalo, State University of New York (SUNY), Buffalo, New York, USA

[2]Computer Science & Engineering, University at Buffalo, State University of New York (SUNY), Buffalo, New York, USA

[3]Data Intensive Studies Center, Tufts University, Medford, Massachusetts, USA

Correspondence
Sreelekha Guggilam, Computational Data Science & Engineering, University at Buffalo, State University of New York (SUNY), Buffalo, NY, USA.
Email: sreelekh@buffalo.edu

Abstract

Data-driven anomaly detection methods typically build a model for the normal behavior of the target system, and score each data instance with respect to this model. A threshold is invariably needed to identify data instances with high (or low) scores as anomalies. This presents a practical limitation on the applicability of such methods, since most methods are sensitive to the choice of the threshold, and it is challenging to set optimal thresholds. The issue is exacerbated in a streaming scenario, where the optimal thresholds vary with time. We present a probabilistic framework to explicitly model the normal and anomalous behaviors and probabilistically reason about the data. An extreme value theory based formulation is proposed to model the anomalous behavior as the extremes of the normal behavior. As a specific instantiation, a joint nonparametric clustering and anomaly detection algorithm (INCAD) is proposed that models the normal behavior as a Dirichlet process mixture model. Results on a variety of datasets, including streaming data, show that the proposed method provides effective and simultaneous clustering and anomaly detection without requiring strong initialization and threshold parameters.

KEYWORDS
anomaly detection, Bayesian nonparametric models, clustering-based anomaly detection, evolving stream data, extreme value theory

1 | INTRODUCTION

Anomalies are unusual, unexpected, and surprising phenomena that need to be detected and explained. Identifying, understanding, and prediction of anomalies from data forms one of the key pillars of modern data mining, and has applications in almost every application domain. For instance, effective detection of anomalies can reveal critical information needed to stop malicious attacks, detect and repair faults, and, ultimately, understand the behavior of a complex system. In fact, one of the most practical applications of anomaly detection is for monitoring system behavior and detecting when the system exhibits

anomalous behavior due to external or internal stress factors [22]. In this regard, two types of anomaly detection methods, viz., online anomaly detection [1, 50, 53] and clustering-based anomaly detection [16, 34, 39], are highly relevant. Online methods, that can simultaneously identify clusters and the anomalies from streaming data, are especially beneficial, as complex system behavior typically falls into multiple regimes or clusters.

However, existing anomaly detection methods face two key challenges in this context. First challenge is the reliance of existing anomaly detection methods on an a priori user-defined threshold, which makes them highly sensitive to the choice of the threshold. While a large literature

on anomaly detection exists [9], most of the existing methods follow a general two-phase strategy: (i) learn a model, $\mathcal{N}$, for the normal behavior of the underlying system, and (ii) score a data instance, x, with respect to $\mathcal{N}$ using a scoring function, $s_{\mathcal{N}}()$. Typically, the score is uncalibrated, though some methods produce a calibrated score (probability). However, to identify anomalies, every method requires a notion of a threshold, δ, such that the data instances whose score is above (or below) δ are anomalous. While unthresholded scores are sufficient for evaluation purposes, for example, generating an ROC curve or comparing different methods on a validation dataset, an optimal threshold is necessary in an operational setting. A very high threshold could potentially result in missing many anomalies while a low threshold would have a high false positive rate. The issue is exacerbated in a streaming setting, where both $\mathcal{N}$ and δ can evolve. While current streaming anomaly detection methods allow updates to $\mathcal{N}$, none of them allow for updating the threshold, δ. Second challenge is specific to clustering-based anomaly detection methods. Traditional methods learn the clustering structure from the observed data as a surrogate for the normal behavior, $\mathcal{N}$. Adapting such methods for streaming data requires the ability to allow the clustering to evolve, that is, new clusters can form and old clusters can grow or split. Current clustering-based methods are not equipped to adapt to such evolving stream behavior.

One possible solution would be to explicitly learn a model, $\mathcal{A}$, for the anomalous behavior, and then compare the scores, $s_{\mathcal{N}}(x)$ and $s_{\mathcal{A}}(x)$, to declare if a data instance is normal or anomalous. By allowing both models to "evolve" in a streaming setting, a robust streaming anomaly detector can be developed. However, given the lack of sufficient (or any) anomalous data, learning $\mathcal{A}$ is not possible. We advocate the use of extreme value theory (EVT) [10] to learn a surrogate for $\mathcal{A}$. The core idea is to assume that the anomalous observations are the extreme values of $\mathcal{N}$. Using a key result in EVT, which states that the extreme values can be modeled as a parameterized distribution (referred to as an extreme value distribution or EVD), one can learn $\mathcal{A}$ for a given $\mathcal{N}$.

In principle, this is a fundamental breakthrough in anomaly detection, and some initial work has been recently published in this direction [50]. However, current EVT supports a limited class of base distributions ($\mathcal{N}$); in fact, while dealing with extremes of a univariate and unimodal distribution is well understood in EVT, handling multivariate and/or richer distributions, for example, mixture models, is a challenge. In this paper, we propose an EVT driven strategy that can admit a richer class of n distributions. A generalization of EVT to multivariate and multimodal distributions [12] is employed, which uses

EVT on the likelihood of the observations, thus reducing the problem to a univariate setting.

As an instantiation of the EVT driven strategy, we propose an anomaly detection method in which the normal behavior, $\mathcal{N}$, is modeled as a nonparametric mixture model—Dirichlet process mixture model (DPMM) [21], or DPMM—which allows clustering the data without pre-specifying the number of clusters. This, especially when adapted to the streaming setting, is an invaluable feature for anomaly detection, where the normal clustering pattern can evolve with the stream. This is an invaluable feature for anomaly detection in an online setting, where the normal clustering pattern can evolve with incremental data addition. The anomalous distribution, $\mathcal{A}$, is also a DPMM with a coupling with $\mathcal{N}$ which forces the parameters of $\mathcal{A}$ to be generated from the extremes of the prior distribution that generates the parameters for $\mathcal{N}$. The resulting method can perform joint clustering and anomaly detection and can be adapted to a streaming setting, with robustness to the choice of threshold for identifying anomalies. Experimental results on synthetic and publicly available datasets are provided to demonstrate the effectiveness of the proposed method over state of art methods.

1.1 | Paper contributions

The paper makes the following key contributions:

1. We propose a method called integrated clustering and anomaly detection (INCAD),[1] that couples Bayesian nonparametric modeling and EVT to simultaneously perform clustering and anomaly detection. INCAD uses a dynamic definition of anomalous and non-anomalous behavior, which makes it well-suited for continuous monitoring applications. At the same time, by using a nonparametric clustering mechanism, that is, DPMMs, the model permits formation of new clusters at subsequent processing steps. This feature helps in address issues in open set classification [4, 25]. Moreover, by explicitly modeling the anomalous behavior, the model can directly produce an anomaly label, instead of relying on a user-defined threshold on a score.
2. We provide a key theoretical result that enables us to extend the EVT formulation to multidimensional data, via the extended generalized Pareto distribution (GPD) modification.
3. We put forward a streaming extension to the INCAD model that captures drift or evolution in streams as illustrated in Figure 1.

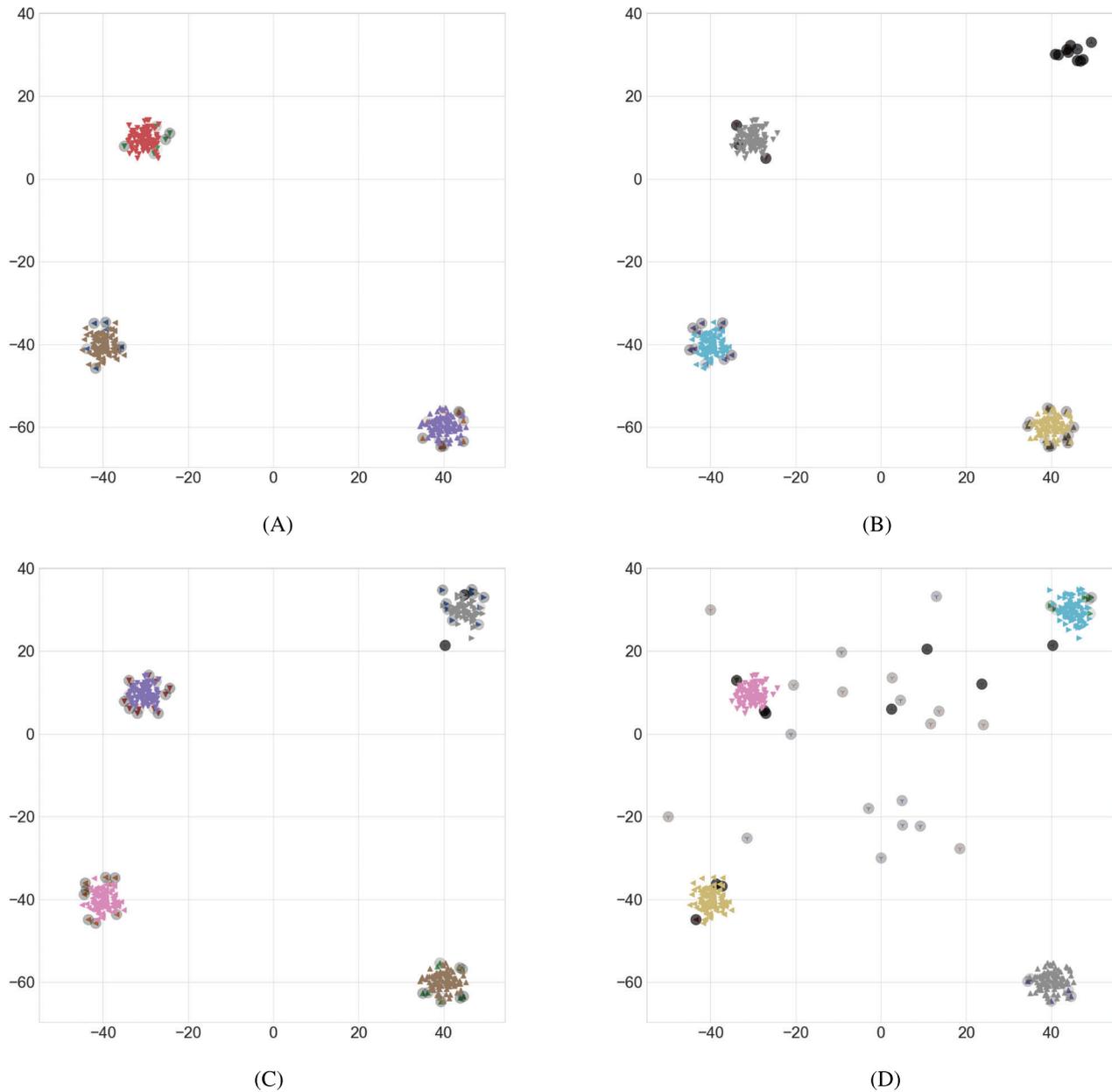[1] A preliminary version of INCAD was published here [33].

FIGURE 1 Illustration of INCAD performance on a synthetic streaming dataset. (A) Before streaming phase: After the initial batch phase, INCAD correctly and automatically identifies three clusters in the data, along with some peripheral data instances as anomalies (denoted by a ∘, where the transparency intensity denotes the probability of observation being anomalous). (B) After initial part of streaming: As new instances arrive in the stream, INCAD first identifies them as anomalies, and then, (C) after introducing all instances for fourth cluster: identifies a new cluster. (D) End of streaming phase: The truly anomalous instances in the stream are labeled as anomalies with higher probability than the false positives (instances on the periphery of the clusters)

4. We provide a comprehensive evaluation of the model on a variety of benchmark datasets to highlight its effectiveness and provide a comparison against existing models.

1.2 | Paper organization

An overview of the existing literature on clustering-based anomaly detection and anomaly detection in streaming or online settings is provided in Section 2. Section 3 presents a short overview on EVT along with the extended GPD to high-dimensional settings. The origins of the methodology for a basic one cluster scenario using EVT is introduced in Section 4. The proposed INCAD model (which is an extension to multiple clusters), the Gibbs sampling algorithm for INCAD and its key features are discussed in detail Section 5. The experimental setup for the INCAD model is detailed in Section 6. The results and final evaluations of

the model against state-of-the-art algorithms in the literature are studied in detail in Section 7.

2 | RELATED WORK

This section examines the different aspects of clustering-based anomaly detection. We review existing research on clustering-based and EVT-based approaches and the research extensions that are necessary for studying evolving anomalous behaviors. This section reflects on the need for a synchronized agglomerated clustering and anomaly detection particularly in streaming settings and justifies the extended approach studied in the paper.

2.1 | Clustering-based anomaly detection

Motivated by the natural tendency of complex systems to exhibit clustering behavior, clustering-based methods rely on the assumption that data corresponding to normal behavior would form natural clusters, whereas anomalous data would either form insignificant clusters or get weakly associated with the natural clusters. Thus, clustering-based anomaly detection methods serve a dual purpose: (a) system identification by discovering clusters in the observed data, and (b) identifying critical anomalies in the system behavior. Traditional methods that first perform clustering, followed by an anomaly detection step, risk the negative impact of anomalies on the clustering step [36, 43]. However, recent solutions have been proposed that avoid this risk by jointly identifying the clusters and anomalies [11, 24, 42].

Existing anomaly detection methods, clustering based or otherwise, have a significant shortcoming when applied in practical settings, that is, they cannot adapt to evolving notions of normal and anomalous behavior. Most methods have rigid definitions of such behavior, encoded as parameters (number of clusters, neighborhood size, etc.) or thresholds, which result in poor performance when the underlying behavior changes. For instance, a clustering-based algorithm that assumes the existence of fixed k clusters, will fail if a new cluster evolves over time.

For the most fundamental problem of identifying anomalies within a set of observations, also referred to as unsupervised anomaly detection, existing methods [36, 43] employ different strategies to model the normal and/or anomalous behavior in the data. In particular, clustering-based techniques rely on the assumption that normal observations cluster together into significant clusters, while anomalies either exist as singletons or very small clusters or are far away from the center of the cluster that they are assigned to. While earlier methods operate in two phases, that is, clustering followed by anomaly

detection, methods that simultaneously identify clusters and anomalies have been recently proposed [11, 24, 42]. However, these methods require the user to prespecify the number of clusters, which makes them unsuitable for scenarios where that information is not available or could evolve.

2.2 | Anomaly detection using DPMM

Bayesian approach for nonparametric modeling was first introduced by Ferguson [17] and Antoniak [3]. Modern variants of these models [46] were introduced for unsupervised clustering. Blei et al. [6] and Yerebakan et al. [60] present hierarchical extensions of the DPMM model that enable more flexible clustering for multimodal and skewed clusters. The models are not tailored to incorporate the order of the observations, which makes them unsuitable for studying streaming data. [60] and, Blei and Frazier [5] propose variational inference based variants that address complexity challenges. Additionally, there exists exemplary work that has explored DPMM for the task of anomaly detection [15, 23, 26, 49, 55, 59] that identify anomalies post clustering in a non-streaming setting. But unlike existing work that are based on exchangeable DPMM models, we propose a non-exchangeable evolving model that studies the dependencies in the order of the observations to jointly study clusters and anomalies.

2.3 | Anomaly detection using EVT

A large body of research exists in the area of anomaly detection [9]. There have been limited applications of EVT for detecting anomalies [2, 20, 50–52]. However, these solutions are limited to one-dimensional (1D) data and typically assume that the normal data follows a unimodal distribution (e.g., Gaussian), though limited extensions to multivariate case [12] have been proposed.

Efficient algorithms that can adapt with streaming data still remain a challenge [48].Anomaly detection methods that use EVT have been proposed [12, 29], but are not applicable in a streaming mode. Some streaming algorithms based on EVT [2, 20, 50] have also been recently proposed to adapt to the evolving behavior but differ from our approach. In particular, the approaches in [2, 50] are limited to univariate streams while the method in [20] is tailored to spatiotemporal data. Although EVT's definition of anomalies is more adaptable for streaming datasets, fitting an extreme value distribution on a mixture of distributions or even multivariate distributions is challenging. In [52], the authors proposed framework uses EVT along with sliding windows for detecting outliers in

nonstationary data stream. An approach based on EVT for detecting outliers in streaming univariate and unimodel time-series is proposed in [50]. This approach is shown to be useful for both stationary and nonstationary streaming data since there is no underlying assumption about the distribution of data stream. A combined approach with Gaussian process and EVT for detecting anomalous behavior in streaming data for maritime vessel track analysis is presented in [51]. This approach models the dynamic properties of the distribution governing extreme values through the use of Gaussian processes.

2.4 | Anomaly detection on streaming data

With pervasive use of sensors in different application domains such as healthcare, smart infrastructure and social networking [57], there is an exponential rise in the availability of streaming data [1]. This can be largely attributed to the rise of internet of things (IoT) which has caused the network of real-time data sources to produce infinite, continuous streams of data. Detecting anomalies in streaming data poses challenges as compared to batch data [48]. The significant challenge for an outlier detection technique is to effectively adapt with the changing nature of the distribution of data streams while detecting anomalies.

While most existing solutions operate in a batch or offline mode, requiring the full dataset in advance, it is challenging to adapt them for a streaming setting [31]. Moreover, existing solutions for anomaly detection with streaming data have either focused on 1D data streams [1, 50] or focus on maintaining the density estimates using a tree based data structure [53, 58] in an online fashion. At the same time, several clustering algorithms that can handle streaming data have been proposed [27, 41], which allow the clustering to evolve with the streaming data, that is, new clusters form, old clusters grow or split. However, none of these methods performs joint clustering and anomaly detection.

3 | EXTREME VALUE THEORY

EVT [10] is the study of extremes of data distributions. The foundations were laid by Fisher and Tippett [18] and Gnedenko [28] who demonstrated the closed forms of the distributions of the extreme values of i.i.d. samples. In this paper, we follow the theory by De Haan and Ferreira [13].

Broadly speaking, there are two principal approaches to study extreme values. One of the approaches is to study the block maxima, that is, the largest observations

TABLE 1 Relation between G and ζ

Tail behavior	Tail distribution	Examples
Exponential tail	Gumbel ($\zeta = 0$)	Gaussian, Exponential, Gumbel, Lognormal
Heavy tail	Fréchet ($\zeta > 0$)	Pareto, Fréchet
Bounded tail	Reversed Weibull ($\zeta < 0$)	Uniform, Beta, Reversed Weibull

in multiple large samples (or blocks) of identically distributed observations. For instance, consider a random variable, X, with G as the cumulative distribution function (CDF).[2] Given n realizations of this random variable, $\{X_1, X_2, \dots, X_n\}$, let, $M_n = max\{X_1, X_2, \dots, X_n\}$. If there exists a sequence of constants $a_n > 0, b_n \in \mathbb{R}$, such that $\frac{M_n - b_n}{a_n}$ has a nondegenerate distribution as $n \to \infty$, that is:

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x) \text{ as } n \to \infty \qquad (1)$$

for every continuity point x of the nondegenerate distribution G^{EV}, then G^{EV} is called an extreme value distribution and the class of distributions G satisfying (1) are called the domain of attraction of G^{EV}.

For univariate data, the generalized extreme value (GEV) distribution, $G^{EV}(x)$, takes the following form:

$$G^{EV}(x) = exp\left\{-\left[1 + \zeta\left(\frac{x - \nu}{\beta}\right)\right]^{-1/\zeta}\right\}, \qquad (2)$$

where ν, β and $\zeta \geq 0$ are the location, scale and shape parameters of the distribution. For $\zeta = 0$ the distribution takes the form

$$G^{EV}(x) = exp\left\{-exp\left[-\frac{x - \nu}{\beta}\right]\right\}, \qquad (3)$$

ζ is typically referred to as the extreme value index and depends on the shape of the tail of the data distribution, G. For instance, if G is a univariate Gaussian distribution, then $\zeta = 0$. Table 1 and Figure 2 show the shapes of the tail for different distributions, and the corresponding value for ζ.

Given a distribution, G, and the corresponding EVD, one can calculate the cumulative probability of an observation x to be an extreme value with respect to G. This requires estimation of the shape parameter, ζ, which can be done directly from data. However, the above

---

[2]We will use G_X to denote the CDF of the data X and G_X^{EV} to denote the corresponding tail distribution. Unless needed, the subscript is omitted for ease of notation.
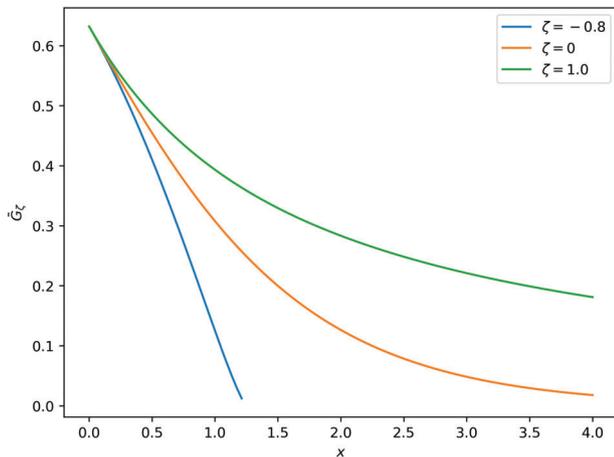
FIGURE 2   Tail distribution for different F for different values of ζ

approach only utilizes maximal value in each block, and is, thus, inefficient. A more economical approach to study extremes, called peaks-over-threshold (POT) [44], studies all large observations which exceed a high threshold. In POT, the excesses over a user-specified threshold, t, that is, $Z = X - t$ can be modeled as a GPD, given by the following CDF:

$$G_Z^{EV}(z) = \begin{cases} 1 - \left(1 + \zeta\left(\frac{z-\mu}{\sigma}\right)\right)^{-\frac{1}{\zeta}} & \text{if } \zeta \neq 0 \\ 1 - \exp\left(-\frac{z-\mu}{\sigma}\right) & \text{if } \zeta = 0 \end{cases} \quad (4)$$

with μ, σ, and ζ as the location, scale, and shape parameters, respectively. The choice of the threshold, t, is often regarded as a bias-variance problem as very large or extreme thresholds lead to fewer observations and over-fitting whereas thresholds resulting in many tail observations result in bias. In this paper, we favor the POT approach due to simplicity in implementation and explanation.

Of course, given a data distribution, G, there is no guarantee that a corresponding EVD exists. A simple theorem from De Haan and Ferreira [13] on domains of attraction for univariate data is used to establish the necessary conditions for the existence of the EVD for G.[3]

Theorem 1. *Let G be a distribution of X with u as the right upper limit on the realizations of X. Assume that second order derivatives G'' exists and the first order derivative G' is positive for all x in the left neighborhood of u. If*

$$\lim_{x \to u}\left(\frac{1-G}{G'}\right)'(x) = \zeta \quad (5)$$

or alternately,

$$\lim_{x \to u}\frac{(1-G(x))(G''(x))}{(G'(x))^2} = -\zeta - 1 \quad (6)$$

then G is in the maximum domain of attraction (MDA)[4] of GEV family of distributions $\mathbf{G}_\zeta^{EV}$ with shape parameter ζ.

3.1 | EVT for multivariate data

In the previous section, we posed the different approaches in EVT in the univariate space. However, most datasets are often multivariate rendering the above approach inapplicable. In this section, we develop the multivariate approach to extreme values.

For the sake of notational simplicity we will discuss a two-dimensional (2D) case, where the random variable, X, is denoted as a tuple (X_1, X_2).

Definition 1. Let $\left\{\left(X_{1,i}, X_{2,i}\right)\right\}_{i=1}^n$ be a sequence of independent and identically distributed random tuples with distribution G. Suppose that there exist sequences of constants $a_i, c_i > 0$ and $b_i, d_i \in \mathbb{R}$ and a distribution G^{EV} with nondegenerate marginals for all continuity points of (x_1, x_2). Then any limit function of G^{EV} given below with nondegenerate marginals is called a multivariate extreme value distribution,

$$\lim_{i \to \infty} P\left(\frac{M_{X_1,i} - b_i}{a_i} \leq x, \frac{M_{X_2,i} - d_i}{c_i} \leq y\right) = G^{EV}(x, y), \quad (7)$$

where $M_{X_1,i} = max\left(X_{1,1}, X_{1,2}, \ldots, X_{1,i}\right)$ and $M_{X_2,i} = max\left(X_{2,1}, X_{2,2}, \ldots, X_{2,i}\right)$.

Extending the univariate results to multivariate settings is often arduous and computationally complex. However, as most data is often multivariate, we study using an alternative approach where the probability image space is used to identify anomalies.

3.2 | Using probability image space for handling multimodal and multivariate distributions

Estimation of parameters for extreme value distributions is often infeasible if the distribution is multimodal and/or if the random variable is multivariate [13, 44]. To address this challenge, recent work by Clifton et al. [12] shows

---

[3]The detailed mathematical proofs for the above theorems is given in De Haan et al. [13].

[4]The maximum domain of attraction can be seen as a family of distributions with tail distributions that are unique up to location and scale parameters.
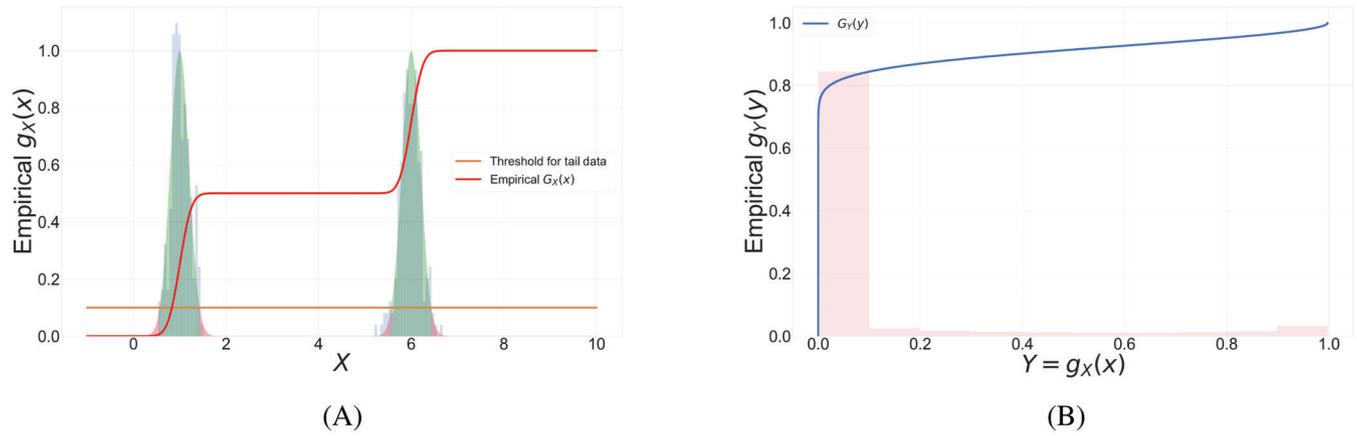
FIGURE 3 Extended GPD distribution using Probability Image Space for Bimodal Univariate Data. Two thousand observations from two random normal distributions with mean and variance (0,2) and (6,2) respectively is shown. (A). Empirical density of the data is shown in green. The observations with probability density less than 0.1 are considered tail observations (shown in red shaded region). The empirical CDF G_X is shown red. (B). The empirical density of the probability image space is shown in red. The cumulative distribution used in the extended GPD approach is shown in blue. (A) Data density. (B) Extended GPD

that it is possible to construct, and examine, an equivalent univariate distribution by considering the probability image space. The result states that for a probability distribution function, $g_X : X \to Y$, where $Y \in \mathbb{R}^+$ is the probability image space, let random variable Y be defined as a distribution G_Y, with following CDF:

$$G_Y(y) = \int_{g_Y^{-1}([0,y])} g_X(x)dx, \qquad (8)$$

where $g_Y^{-1}([0, y])$ denotes all the values of the random variable X, whose probability density is between 0 and y. Using the POT result [44], as discussed earlier, it can be shown that for a small positive value, u, the tail of G_Y can be modeled as a GPD for $y \in [0, u]$, as $u \to 0$, such that if an observation x is extreme with respect to the original distribution, G_X, if $g_X(x) < u$, then $y = g_X(x)$ will be extreme with respect to G_Y. The corresponding GPD for $(u - y)$, denoted as G_Y^{EV}, can be used to calculate the probability of x to be extreme, with respect to G_X.

A simulated example is shown in Figure 3, where 2000 observations from two univariate Gaussian distributions are studied. Unlike the traditional EVT approach that can only study tail distributions for unimodal data, the Ext-GPD approach is able to include rare observations between the two modes as seen in the shaded red zone in Figure 3A. The probability image space of the mixture distribution is used to study the observations with low probabilities, that is, the rare tail observations. The resulting image space is considered as the 1D projections of the original data and the anomalies are identified by studying the left tail in Figure 3B. The Ext-GPD approach is discussed in detail in Section 3.3. The theory behind

the extended GPD approach has not been presented earlier [12]. Hence, we present the necessary conditions 1D data in Section 3.3. The proof for multidimensional case is similar and has been included in the Appendix.

3.3 | Ext-GPD approach

In this section, we derive the necessary conditions required for the Extended GPD approach. For this, consider the following setting in the univariate space.[5]

Let $X \in \mathbb{R}$ be the data space with pdf[6] $g_X : \mathbb{R} \to \mathbb{R}^+$. Let $Y \in \mathbb{R}^+$ be the corresponding image space, that is, $Y = g_X(X)$ and $Y_m = sup(g_X(X))$. As the limit distribution of the minima of Y is of interest, we wish to study the limit distribution of maxima of $Z = Y_m - Y$. Let the CDF of Z is given by G_Z. Then, we show that the Theorem 2 holds.

Theorem 2. *G_Z is in the maximum domain of attraction of a GEV distribution $\mathbf{G}_\zeta^{EV}$, iff $\frac{dg_X(x)}{dz}$ and $\frac{d^2 g_X(x)}{dz^2}$ exists $\forall x \in g_X^{-1}(Y_m - z)$ in some neighborhood of Y_m.*

Proof To derive the necessary conditions for the Ext-GDP approach, we make the following claims.

Claim 1. G_Z is a CDF.

Proof As the limit distribution of the minima of Y is of interest, we wish to study the limit distribution of maxima

---

[5]The proof for the higher dimensional space is presented in the Supporting Information.
[6]Note: g_X^{-1} represents an image set as the function g_X is a many-to-one (noninjective) function.

of $Z = Y_m - Y$. Then the CDF of Z is given by G_Z is

$$
\begin{aligned}
G_Z(z) &= P(Z \leq z) \\
&= P(Y_m - Y \leq z) \\
&= P(Y \geq Y_m - z) \\
&= 1 - G_Y(Y_m - z) \\
&= \int_{g_X^{-1}([Y_m - z, Y_m])} g_X(x) dx, \quad (9)
\end{aligned}
$$

$\forall z \in [0, Y_m]$.

For, G_Z, the corresponding maximum value, $x^* = Y_m$.

Claim 2. G_z' exists and is positive in some neighborhood of Y_m.

Proof   If F be a distribution in 1D, $\exists\ \{x_1 = \infty, x_1, x_2 \ldots, x_{2N} = \infty\}$ and intervals $I_1, I_2, \ldots, I_{N-1}$ such that $I_n = \left[x_{\frac{n}{2}}, x_{\frac{n+1}{2}}\right]\ \forall\ n = 1, 2, \ldots, N$ and $g_X^{-1}([0, Y_m - z]) = \cup_{n=1}^N I_n$

$$
\begin{aligned}
\int_{g_X^{-1}([0, Y_m - z])} g_X(x) dx &= \int_{\cup_{n=1}^N\ I_n} g_X(x) dx \\
&= \sum_{n=1}^N \int_{I_n} g_X(x) dx \\
&= \sum_{n=1}^N G_n(z), \quad (10)
\end{aligned}
$$

where $G_n(z) = \int_{I_n} g_X(x) dx$ and $\{x_2, x_2 \ldots, x_{2N-1}\}$ are the solutions to $g_X^{-1}(Y_m - z)$.

Then,

$$
\begin{aligned}
G_Z'(z) &= \frac{d}{dz} \int_{g_X^{-1}([Y_m - z, Y_m])} g_X(x) dx \\
&= \frac{d}{dz} \left(1 - \int_{g_X^{-1}([0, Y_m - z])} g_X(x) dx\right) \\
&= -\frac{d}{dz} \sum_{n=1}^N G_n(z). \quad (11)
\end{aligned}
$$

Since $G_n(z) = \int_{I_n} g_X(x) dx = \int_{x_{n-1}}^{x_n} g_X(x) dx$, by Leibniz integral rule, we get,

$$
\begin{aligned}
\frac{d}{dz} G_n(z) &= \frac{d}{dz} \int_{x_{n-1}}^{x_n} g_X(x) dx \\
&= g_X(x_n) \frac{dx_n}{dz} - g_X(x_{n-1}) \frac{dx_{n-1}}{dz} + \int_{x_{n-1}}^{x_n} \frac{d}{dz} g_X(x) dx \\
&= (Y_m - z) \left(\frac{dx_n}{dz} - \frac{dx_{n-1}}{dz}\right) \\
&= -(Y_m - z) \left(\left|\frac{dx_n}{dz}\right| + \left|\frac{dx_{n-1}}{dz}\right|\right). \quad (12)
\end{aligned}
$$

Then,

$$
\begin{aligned}
G_Z'(z) &= \sum_{n=0}^{2N} (Y_m - z) \left|\frac{dx_n}{dz}\right| \\
&= (Y_m - z) \sum_{x \in g_X^{-1}(Y_m - z)} \left|\frac{dx}{dz}\right|. \quad (13)
\end{aligned}
$$

Claim 3. G_Z'' exists iff $\frac{dg_X(x)}{dz}$ and $\frac{d^2 g_X(x)}{dz^2}$ exists $\forall x \in g_X^{-1}(Y_m - z)$.

Proof

$$
\begin{aligned}
G_Z''(z) &= \frac{d}{dz} G_Z'(z) \\
&= \frac{d}{dz} \left[(Y_m - z) \sum_{x \in g_X^{-1}(Y_m - z)} \left|\frac{dx}{dz}\right|\right]. \quad (14)
\end{aligned}
$$

It can be seen that G_Z'' exists iff $\frac{dg_X(x)}{dz}$ and $\frac{d^2 g_X(x)}{dz^2}$ exists $\forall x \in g_X^{-1}(Y_m - z)$. This is true for all distributions in the exponential family.

Claim 4. G_Z is in the maximum domain of attraction of a GEV distribution $\mathbf{G}_\zeta^{EV}$, where $\zeta \in \mathbb{R}$ is the rate parameter of the GEV distribution.

Proof   By von Mises' condition,[7] and Claims 2 and 3, we can see that the G_Z' is positive and G_Z'' exists in some neighborhood of Y_m. Hence, G_Z is in domain of attraction of $\mathbf{G}_\zeta^{EV}$.

Using Claims 1–4, we get the necessary conditions for the above claim.

The extension to the multivariate case is shown in Theorem 3. The proof is included in the Appendix to keep the presentation here focused.

Theorem 3.   *Let $\vec{\mathbf{X}} \in \mathbb{R}^n$ be the data space with pdf $g_X : \mathbb{R}^n \to \mathbb{R}^+$. Let $Y \in \mathbb{R}^+$ be the corresponding image space. Let $\vec{\mathbf{X}} \in \mathbf{R}^n$ and $g_{\vec{\mathbf{X}}}^{-1}([0, Y_m - z]) = D(Y_m - z)$ be a n-manifold with a boundary $\partial D(Y_m - z)$. G_Z is in the maximum domain of attraction of a GEV distribution iff:*

1. *$D(Y_m - z)$ is an n-manifold with a boundary $\partial D(Y_m - z)$,*
2. *The Eulerian velocity of the boundary $\vec{\mathbf{v}}_b = \frac{dD(Y_m - z)}{dz}$ exists,*
3. *$d_x\left[g_X(\vec{\mathbf{x}})\vec{\mathbf{v}}_b \cdot d\mathbf{\Sigma}\right]$ exists, and*
4. *$i_{\vec{\mathbf{v}}}\left(d_x\left[g_X(\vec{\mathbf{x}})\vec{\mathbf{v}}_b \cdot d\mathbf{\Sigma}\right]\right)$ exists.*

---

[7]von Mises' condition: Let F be a distribution function and x^* is its right end point. Suppose F'' exists and F' is positive for all x in some neighborhood of x^*. If $\lim_{t \to x^*} \left(\frac{1-F}{F'}\right)(t) = \zeta$ then, F is in the MDA of $\mathbf{G}_\zeta^{EV}$.
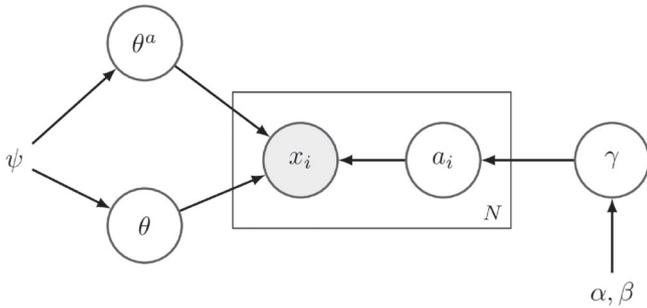
FIGURE 4 Graphical representation of the proposed probabilistic model

4 | ANOMALY DETECTION USING EVT FOR UNIMODAL DATA

EVT plays a significant role in studying rare events and so, several methods have been proposed that incorporate these features in anomaly detection. Here, we present a novel methodology which involves both EVT and non-parametric modeling for anomaly detection. The core principles that lead to the development of the integrated algorithm are discussed here. We start with a basic case of one cluster data with anomalies.

Based on the EVT concepts discussed above, we first propose a simple anomaly detection model (Figure 4), which is equivalent to the following generative distributions:

$$\theta|\psi \sim \mathrm{G}_0(\psi), \tag{15}$$

$$\theta^a|\psi \sim \mathrm{G}_0^{\mathrm{EV}}(\psi), \tag{16}$$

$$\gamma|\alpha, \beta \sim \mathrm{Beta}(\alpha, \beta), \tag{17}$$

$$a_i|\gamma \sim \mathrm{Bernoulli}(\gamma), \tag{18}$$

$$x_i|a_i, \theta, \theta^a \sim \begin{cases} G(\theta) & \text{if } a_i = 1 \\ G(\theta^a) & \text{if } a_i = -1 \end{cases}, \tag{19}$$

The model is a mixture of two components, $\mathcal{N}$ and $\mathcal{A}$, parameterized by θ and θ^a, respectively. a_i is an indicator latent variable denoting if x_i is normal or anomalous, and γ is the mixture weight with a `Beta` distribution prior.

The mixture of models representation allows us to sketch a Gibbs sampling-based inference scheme, similar to a mixture model [19], using the following conditional posteriors:

$$p\left(\gamma|\mathbf{a}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) = \mathrm{Beta}\left(\alpha + n^a, \beta + n - n^a\right), \tag{20}$$

where $\mathbf{x}$ denotes the vector of n observed data instances, $\mathbf{a}$ is a binary indicator vector, that is, $a_i = -1 \Rightarrow x_i$ is anomalous, and n^a is the number of anomalous instances. The

posteriors for the indicators can be computed as:

$$p\left(a_i = -1|\mathbf{a}_{-i}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto \gamma p_G\left(x_i|\theta^a\right), \tag{21}$$

$$p\left(a_i = 1|\mathbf{a}_{-i}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto (1-\gamma)p_G\left(x_i|\theta\right). \tag{22}$$

Finally, the posteriors for the mixture parameters, θ and θ^a, can be computed as:

$$p\left(\theta|\mathbf{a}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto p_{\mathrm{G}_0}(\theta|\psi) \prod_{i:a_i=1} p_G\left(x_i|\theta\right), \tag{23}$$

$$p\left(\theta^a|\mathbf{a}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto p_{\mathrm{G}_0^{\mathrm{EV}}}\left(\theta^a|\psi\right) \prod_{i:a_i=-1} p_G\left(x_i|\theta^a\right). \tag{24}$$

Starting from an initial estimate of the latent variables, γ, $\mathbf{a}$, θ, and θ^a, the inference can be done via Gibbs update, in which new estimates for the latent variables are sampled from the conditional posteriors given in (20), (22), and (24), respectively.

4.1 | Modified posterior expressions

Let y_i denote the pdf of an observation x_i according the to the normal distribution, that is, $y_i = p_G(x_i|\theta)$. Using a threshold u,[8] we define the "tail" of the distribution G_Y using samples $\{y_i\}_{i:y_i \le u}$. A GPD, G_Y^{EV}, is fitted on the samples $\{u - y_i\}_{i:y_i \le u}$. The conditional posteriors for a_i for tail instances can be written as:

$$p\left(a_i = -1|\mathbf{a}_{-i}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto \gamma\left(1 - P_Y^{EV}\left(u - y_i\right)\right), \tag{25}$$

$$p\left(a_i = 1|\mathbf{a}_{-i}, \mathbf{x}, \theta, \theta^a, \alpha, \beta, \psi\right) \propto (1-\gamma)P_Y^{EV}\left(u - y_i\right), \tag{26}$$

where $P_Y^{EV}\left(u - y_i\right)$ is the probability of observing y_i in the tail of G_Y. Since *GPD* is a unimodal distribution, we use the survival function value, $1 - G_Y^{EV}\left(y - u_i\right)$, instead of the exact probability. For non-tail instances, that is, $y_i > u$, the conditional probability $p\left(a_i = -1|\ \ldots\right)$ is set to 0. Under this modified model, computing the posterior for θ^a in (24) is not needed anymore. If the form of the normal model is known, for example, a unimodal Gaussian or a mixture of Gaussians[9] (Figure 5), the anomalies and the model

---

[8]Note that u is not a threshold for determining if an observation is anomalous or not; instead, it defines the "tail" of the original distribution, which are then used to determine the parameters of the corresponding GPD.

[9]In presence of multiple clusters, the prior G_0 can be chosen as a mixture of individual priors generating the non-anomalous components ensuring that low probability or tail region of the distribution is associated with generating parameters associated with anomalous components.
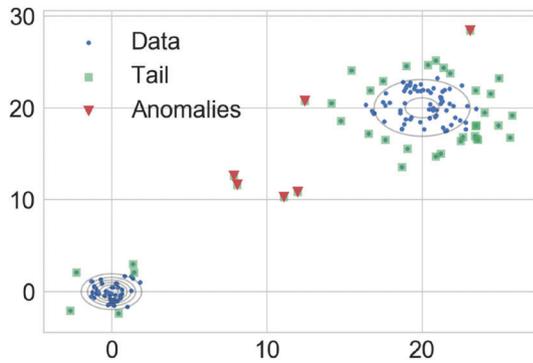
FIGURE 5 Results for a synthetic 2D case, with a fixed Gaussian mixture model as G_0. The model identifies the anomalies (red) with respect to the tail of G_0 (green) as well as the parameters for G_0 (shown as contour lines)

parameters can be inferred via Gibbs sampling, using the above mentioned conditional distributions. However, in the next section, we show how the Bayesian formulation can be extended to a richer class of the base distribution, G_0, that is, nonparametric mixture models.

Challenges: If G_0 is the conjugate prior of G, one can get an analytical form for the posterior in (24). The posterior for θ^a is the main challenge here, for two reasons: (a) G_0^{EV} exists only for a limited base distributions, G_0, and, (b) even for known G_0^{EV}, it is unlikely that the posterior in (24) will have an analytical form.

We first note that the quantity $p_G(x_i|\theta^a)$ is the probability of the observation x_i to be generated by the distribution G, parameterized by θ^a, which, in turn, is sampled from the EVD for G_0, that is, G_0^{EV}.

For distributions belonging to the exponential family, one can show that if G_0 is the conjugate of G, then sampling x_i from $G(.|\theta^a)$, where $\theta^a \sim G_0^{EV}$, is equivalent to (under expectation): first sampling θ from G_0, and then sampling x_i from the EVD of G (or G^{EV}), parameterized by θ, that is, $\mathbb{E}_{\theta^a \sim G_0^{EV}}\left[p_G(x_i|\theta^a)\right] = \mathbb{E}_{\theta \sim G_0}\left[p_{G^{EV}}(x_i|\theta)\right]$.

We show that this claim will hold for the following simple setting, and omit the general proof in the interest of space. Let $G \sim \mathcal{N}(\mu, 1)$, that is, G is a univariate Gaussian distribution with fixed variance and the mean is generated from a Gaussian prior, that is, $G_0 \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Note that the EVD for a Gaussian distribution is a Gumbel distribution, that is, $G_0^{EV} \sim Gumbel(\mu_0, \sigma_0)$.

Assuming that x_i is an anomaly, that is, x_i is sampled from a Gaussian, $\mathcal{N}(\mu^a, 1)$, where $\mu^a \sim Gumbel(\mu_0, \sigma_0)$, then we can show that for any $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$, the probability that x_i is not in the tail of $\mathcal{N}(\mu, 1)$ will be very small, since:

$$\mathbb{E}_{X \sim \mathcal{N}(\mu^a, 1)}\left[G_{\mathcal{N}(\mu, 1)}(X)\right]$$

$$= \int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu^a)^2}{2}\right) dx$$

$$\propto \exp\left(-\frac{(\mu-\mu^a)^2}{4}\right).$$

Thus, the claim will hold in this case because the prior distribution is Gaussian, for which $|\mu - \mu^a| \gg 0$.

5 | EXTENSION TO DATA WITH MULTIPLE CLUSTERS

While the previous result is an interesting step toward explicitly modeling the anomaly distribution, it is still limited to the case where the normal data is being generated from a single cluster. A natural extension to the presented preliminary model is the scenario where the normal data could be generated from multiple clusters. The key challenge in performing anomaly detection on such data is the method to identify the generative model that is robust to anomaly presence.

Why integrate EVT and DPMM?: Anomalies with significantly large deviations are inherently caught by most anomaly detection algorithms including traditional DPMM. The distinction between the algorithms is observed when identifying anomalies with relatively similar behavior to normal data. Such anomalies are found in the vicinity of clusters and are often clustered into being normal. Traditional DPMM algorithm can identify such anomalies by increasing the concentration parameter but the choice of the new value has the same challenges as the choice of a threshold thus arising a need for an external algorithm like EVT that studies these tail points separately and an integrated approach would ensure enhanced and robust clustering.

5.1 | Background on mixture models

Finite mixture models (FMM) are a useful clustering tool to identify and study subpopulations within data. However, they require prespecifying the number of clusters, which is not always known. This is especially important for anomalous data for which accurate knowledge is not available, and can lead to some significantly inaccurate (and in some cases unreliable) interpretations of the data. Nonparametric mixture models, for example, DPMMs [21], can be used in such settings.

DPMMs: A DPMM can be thought of as an infinite extension of an FMM, which is equivalent to the following distributions:

$$\boldsymbol{\pi}|\alpha \sim \text{Dir}(\alpha/K, \ldots, \alpha/K), \tag{27}$$

$$z_i | \pi \sim \text{Multi}(\pi), \tag{28}$$

$$\theta_k | \psi \sim \text{G}_0(\psi), \tag{29}$$

$$x_i | z_i, \{\theta_k\}_{k=1}^K \sim G\left(\theta_{z_i}\right). \tag{30}$$

Each observation x_i is generated by first sampling a cluster index, z_i from a multinomial distribution, parameterized by a K length vector, π. A symmetric Dirichlet prior is used to generate π. The observations are sampled from a cluster specific distribution, G, parameterized by θ_k. The cluster specific distribution parameters are also generated from a prior (or base) distribution, G_0, parameterized by ψ.

A DPMM is an extension of FMM to the case where $K \rightarrow \infty$. While several equivalent representations of DPMM exist, we will use the stick breaking representation, which shows DPMM as a natural extension of FMM. The stick breaking representation allows sampling the mixture weights, with possibly infinite components, as follows:

- Start with a unit-length stick and break it according to β_1, where $\beta_1 \sim \text{Beta}(1, \alpha_0)$, and assign β_1 to π_1;
- Break remaining stick according to the proportion $\beta_k \sim \text{Beta}(1, \alpha_0)$ and assign β_k portion of the remaining stick to π_k.

The sequence $\pi = \{\pi_k\}_{k=1}^\infty$ satisfies $\sum_{k=1}^\infty \pi_k = 1$ and is typically written as $\pi \sim \text{GEM}(\alpha)$.[10]

5.2 | Integrated nonparametric clustering and anomaly detection

We propose an instance of the general Bayesian anomaly detection algorithm described in Section 4 which uses a DPMM as its base distribution, G_0. The generative model (Figure 6) consists of two coupled DPMM models, each corresponding to the normal and anomalous behaviors, respectively, and is equivalent to the following distributions[11]:

$$\pi | \alpha \sim \text{GEM}(\alpha), \tag{31}$$

---

[10]Named after Griffiths, Engen, and McCloskey.

[11]GEM is a recursive process with an infinite number of clusters of which only a finite number of them are populated. The number of the populated clusters as well as the corresponding proportions are learned sequentially as seen in the stick breaking process. Since the true number of clusters is unknown, Dirichlet process priors, like the GEM distribution, are traditionally used to sample the vectors π and π^a. When sampling from the GEM distribution, we generate a vector (of unknown but finite length) from a simplex that sums to one (as seen in the stick breaking approach). The vector length can be regulated using the concentration parameter (large concentration parameter returns more number of populated clusters, i.e., vector of longer length).
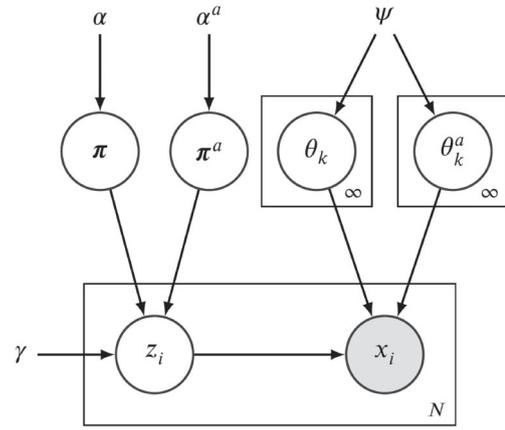

FIGURE 6 Graphical representation of the proposed INCAD model

$$\pi^a | \alpha^* \sim \text{GEM}(\alpha^*), \tag{32}$$

$$\theta_k | \psi \sim \text{G}_0(\psi), \tag{33}$$

$$\theta_k^a | \psi \sim \text{G}_0^{EV}(\psi), \tag{34}$$

$$sign(z_i) | \gamma \sim \text{Bernoulli}(\gamma), \tag{35}$$

$$|z_i| \big| \pi, \pi^a, sign(z_i) \sim \begin{cases} \text{Multi}(\pi) & \text{if } sign(z_i) = 1 \\ \text{Multi}(\pi^a) & \text{if } sign(z_i) = -1 \end{cases}, \tag{36}$$

$$x_i | z_i, \{\theta_k\}_{k=1}^\infty, \{\theta_k^a\}_{k=1}^\infty \sim \begin{cases} \text{G}\left(\theta_{|z_i|}\right) & \text{if } sign(z_i) = 1 \\ \text{G}\left(\theta_{|z_i|}^a\right) & \text{if } sign(z_i) = -1 \end{cases}. \tag{37}$$

The key difference from the model in Section 4 is the additional variable, z_i, that works as the cluster labels as well as anomaly indicator. The $sign(z_i)$ represents the presence of anomalous behavior where anomalous (or non-anomalous) observations are assigned negative (or positive) labels. Based on the observed labels, anomalies can be classified into global, local and group anomalies.

Definition 2. (Global anomalies). A single observation is defined as a group anomaly if it is an observation with distinctly novel behavior. INCAD classifies such observations into singleton clusters with negative cluster labels.

Definition 3. (Group anomalies). Multiple observations with similar behavior that is distinct from existing predominant behaviors (normal clusters) are classified as group anomalies. Such observations are classified into smaller clusters with negative cluster labels.

Definition 4. (Local anomaly). Observations with behaviors that moderately deviate from normal clusters but are

not distinct enough to form individual clusters are defined as local anomalies. Such observations are classified into normal clusters with similar behavior but with negative labels to indicate diverging behavior. Anomalies that originate from an overlapping anomalous cluster are often classified as local anomalies.

Since labels are assigned considering both clustering as well as anomaly detection, we call this model, INCAD (integrated nonparametric clustering and anomaly detection). Based on $sign(z_i)$, z_i is sampled from a multinomial distribution that is either parameterized by $\boldsymbol{\pi}$ (if $sign(z_i) = 1$) or $\boldsymbol{\pi}^a$ (if $sign(z_i) = -1$). The Multinomial parameters, $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^a$ are sampled from the stick breaking construction of a Dirichlet process, that is, $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ and $\boldsymbol{\pi}^a \sim \text{GEM}(\alpha^*)$.

The INCAD model goes beyond the illustrated simple case where we assume multiple anomalous sources, each associated with a different concentration parameter α^*. The generative model can now be seen as a collection of multiple DPMMs of which all but one DPMM can be perceived as sources for anomalous data and the set of concentration parameters for anomalous data, $\{\alpha_d^*\}$, would dictate the corresponding DPMM's cluster proportions $\{\boldsymbol{\pi_d^a}\}$.

Inference for the INCAD model includes inferring posteriors for $(z_i)_{i=1}^n$, $\left(\theta_k, \theta_k^a\right)_{k=1}^\infty$. While this follows the general Gibbs sampling-based scheme discussed in Section 4 (omitting exact details in the interest of space), there are some additional issues that are unique to the INCAD model. In particular, the dependency between z_i and $sign(z_i)$ in Figure 6 means that one cannot consider the model as a straightforward mixture for two DPMMs. However, the relationship between the normal and anomalous model parameters, via the EVT construct, means that we can calculate the posteriors for $sign(z_i)$ using the modification proposed earlier (26).

5.2.1 | Inference when G_0^{EV} is available

MCMC and variational inference based algorithms [7, 40] have been typically used for inference of the computationally expensive infinite mixture models. For INCAD, we adopt an extension of a Gibbs sampling-based method for a fixed mixture model that allows room for additional cluster formation. The algorithm is inspired by the sampling-based MCMC method for conjugate priors (Algorithm 1, [40]). Here, new clusters comprise anomalous observations identified using EVT.

Gibbs sampling: The anomaly classification variable $sign(z_.)$ is a unique feature of INCAD that distinguishes it from traditional DPMM. Thus, the posterior probabilities for the latent variables namely, the number of clusters K, cluster and anomaly indicators $\{z_i\}_{i=1}^N$ are computed using

Markov property and Bayes rule:

$$
\begin{aligned}
P\left(|z_i| = k \,|\, x., z_{-i}, \alpha, \alpha^*, \boldsymbol{\pi}, \boldsymbol{\pi}^a, \psi, \{\theta_k\}, \{\theta_k^a\},\right. \\
\left. sign(z.), \gamma\right) = P\left(|z_i| = k \,\middle|\, x.,\right. \\
\left. z_{-i}, \alpha, \alpha^*, \{\theta_k\}, \{\theta_k^a\}, sign(z_i)\right),
\end{aligned} \tag{38}
$$

$$
\propto \begin{cases}
P(|z_i| = k | z_{-i}, \alpha, \theta_k) & sign(z_i) = 1 \\
\times P(x_i | |z_i| = k, z_{-i}, \theta_k, \alpha), & \\
P(|z_i| = k | z_{-i}, \alpha^*, \theta_k^a) & sign(z_i) = -1 \\
\times P(x_i | |z_i| = k, z_{-i}, \theta_k^a, \alpha^*),
\end{cases} \tag{39}
$$

$$
= \begin{cases}
\frac{n_k}{(n+\alpha-1)} G(x_i \,\middle|\, \theta_k), & sign(z_i) = 1 \\
\frac{n_k}{(n+\alpha^*-1)} G(x_i \,\middle|\, \theta_k^a), & sign(z_i) = -1
\end{cases} \tag{40}
$$

where $\alpha^* = \frac{1}{1-p_i}$, p_i is the probability of x_i being anomalous, n_k is the number of observations in the kth cluster and K is the number of non-empty clusters. In the improved versions of INCAD, p_i is the cumulative density function for the extreme value distribution.

The posterior probability of forming a new cluster denoted by $K + 1$ is given by:

$$
\begin{aligned}
P\left(|z_i| = K+1 \,|\, x., z_{-i}, \alpha, \alpha^*,\right. \\
\left. \boldsymbol{\pi}, \boldsymbol{\pi}^a, \psi, \{\theta_k\}, \{\theta_k^a\}, sign(z.), \gamma\right) \\
= P(|z_i| = K+1 \,|\, x_i, z_{-i}, \alpha, \alpha^*, \psi, sign(z_i)),
\end{aligned} \tag{41}
$$

$$
\propto \begin{cases}
P(|z_i| = K+1 | z_{-i}, \alpha, \psi) & sign(z_i) = 1 \\
\times P(x_i | |z_i| = K+1, z_{-i}, \alpha, \psi, sign(z_i)), & \\
P(|z_i| = K+1 | z_{-i}, \alpha^*, \psi) & sign(z_i) = -1 \\
\times P(x_i | |z_i| = K+1, z_{-i}, \alpha^*, \psi, sign(z_i)),
\end{cases} \tag{42}
$$

$$
= \begin{cases}
\frac{\alpha}{n+\alpha-1} \int G(x_i | \theta) G_0(\theta | \psi) \, d\theta, & sign(z_i) = 1 \\
\frac{\alpha^*}{n+\alpha^*-1} \int G(x_i | \theta^a) G_0^{EV}(\theta^a | \psi) \, d\theta^a, & sign(z_i) = -1
\end{cases} \tag{43}
$$

Similarly, the parameters for clusters $k \in \{1, 2, \ldots, K\}$ are sampled from:

$$
\theta_k \propto G_0(\theta_k | \psi) \mathcal{L}(\boldsymbol{x}_k | \theta_k) \quad \text{if cluster is not anomalous,} \tag{44}
$$

$$
\theta_k^a \propto G_0^{EV}(\theta_k^a | \psi) \mathcal{L}(\boldsymbol{x}_k | \theta_k^a) \quad \text{if cluster is anomalous.} \tag{45}
$$

where $\boldsymbol{x}_k = \{x_i | |z_i| = k\}$ is the set of all points in cluster k. Finally, to identify the anomaly classification of the data, the posterior probability of $sign(z_i)$ is given by:

$$
\begin{aligned}
P\left(sign(z_i) = -1 \,|\, x., |z.|, \alpha, \alpha^*, \boldsymbol{\pi}, \boldsymbol{\pi}^a, \psi, \{\theta_k\}, \{\theta_k^a\}, \gamma\right) \\
= P\left(sign(z_i) = -1 | x_i, |z.|, \alpha^*, \psi, \{\theta_k^a\}, \gamma\right),
\end{aligned} \tag{46}
$$

$$\propto \sum_{k=1}^{K+1} P\left(sign\left(z_i\right) = -1 \,|\, x_i, |z_i| = k, z_{-i}, \alpha^*, \psi, \left\{\theta_k^a\right\}, \gamma\right)$$

$$* P\left(|z_i| = k \,|\, x_i, z_{-i}, \alpha^*, \psi, \left\{\theta_k^a\right\}, \gamma\right) \qquad (47)$$

$$= \sum_{k=1}^{K} P(x_i \,|\, \theta_k^a) \gamma \frac{n_k}{(n + \alpha^* - 1)}$$

$$+ \left(\int G\left(x_i \,|\, \theta^a\right) \mathrm{G}_0^{\mathrm{EV}}\left(\theta^a \,|\, \psi\right) d\theta^a \right) \gamma \frac{\alpha^*}{n + \alpha^* - 1}. \quad (48)$$

Similarly,

$$P\left(sign\left(z_i\right) = 1 \,\big|\, x_i, |z_.|, \alpha, \psi, \{\theta_k\}, \gamma\right) \qquad (49)$$

$$\propto \sum_{k=1}^{K} P(x_i \,|\, \theta_k)(1 - \gamma) \frac{n_k}{(n + \alpha - 1)}$$

$$+ \left(\int G\left(x_i \,|\, \theta\right) \mathrm{G}_0\left(\theta \,|\, \psi\right) d\theta \right) (1 - \gamma) \frac{\alpha}{n + \alpha - 1}. \tag{50}$$

5.2.2 | Inference when $\mathrm{G}_0^{\mathrm{EV}}$ is not available

Existence of a tail distribution $\mathrm{G}_0^{\mathrm{EV}}$ is not always feasible. As the extreme value distribution might not belong to the family of the conjugate priors of G, we assume $\theta^a \sim \mathrm{G}_0$ for sampling the parameters $\left\{\theta_k^a\right\}_{k=1}^{\infty}$ for anomalous clusters. Here, we perform rejection sampling to sample observations from the tail distribution. For this, we initially sample P observations from G_0 and isolate observations with probability density less than a set threshold[12] $0 < t \ll 1$. The above procedure is repeated M times till sufficient samples S_{tail} from the tail distribution have been identified. The cluster means $\left\{\theta_k^a\right\}_{k=1}^{\infty}$ can be estimated by randomly sampling from the tail observations S_{tail}. However, this could result in potential convergence issues. Thus, we propose the closest observation in S_{tail} to the sample estimate for the respective anomalous cluster.

The pseudo-Gibbs sampling algorithm, presented in Algorithm 2, has been designed to address the cases when $\mathrm{G}_0^{\mathrm{EV}}$ is not available. For such cases, the modified concentration parameter α^* is given by the function f where,

$$f\left(\alpha \,|\, x_n, \mathbf{x}, \mathbf{z}\right) = \begin{cases} \alpha, & \text{if not in tail} \\ \frac{1}{1 - p_n}, & \text{if in tail} \end{cases}, \tag{51}$$

Algorithm 1. Gibbs sampling algorithm when $\mathrm{G}_0^{\mathrm{EV}}$ is available

Given $z^{(t-1)}, \left\{\theta_k^{(t-1)}\right\}, \left\{\theta_k^{a(t-1)}\right\}$ from iteration $(t-1)$. Let K be the total number of clusters at iteration $(t-1)$.

  Set $z_. = |z_.^{(t-1)}|$ and $a_. = sign(z_.^{(t-1)})$
  for each observation i **do**
    Remove x_i from its cluster z_i.
    if x_i is the only point in its cluster **then**
      Remove the cluster and update K to K-1.
    end if
    Drop empty clusters.
    Sample z_i from the Multinomial distribution given by Equations (40) and (43)
    if $z_i = K + 1$ **then**
      Sample new cluster parameters from the following distribution.[14]

$$\theta \,\Big|\, x_i, z_., \left\{\theta_k^{(t-1)}\right\}, \left\{\theta_k^{a(t-1)}\right\}, a_.^{(t-1)} \qquad (52)$$

$$\propto \begin{cases} \alpha \mathrm{G}_0(\theta|\psi) G(x_i|\theta) + \sum_{j \neq i} G(x_i|\theta_{z_j}) \\ \quad \times \delta(\theta - \theta_{z_j}^{(t-1)}) \delta(a_j^{(t-1)}), \quad a_i^{(t-1)} = 1 \\ \alpha^* \mathrm{G}^{\mathrm{EV}}(\theta|\psi) G(x_i|\theta) + \sum_{j \neq i} G(x_i|\theta_{z_j}) \\ \quad \times \delta(\theta - \theta_{z_j}^{(t-1)}) \delta(a_j^{(t-1)} - 1), \quad a_i^{(t-1)} = -1 \end{cases} \qquad (53)$$

      Update $K = K + 1$
    end if
    for each cluster $k \in \{1, 2, \dots, K\}$ **do**
      Sample cluster parameters θ_k and θ_k^a using Equations (44) and (45).
    end for
    Sample the anomaly classification a_i using Equations (48) and (50).
    Set $z_i^{(t)} = z_i * a_i$
  end for

---

where p_n is the cumulative density of x_n for the extreme value distribution of the tail data[13] where, the cumulative density is given by the extended GPD described in Section 3.3.

5.2.3 | Non-exchangeability and evolution detection in stream

Exchangeable models are robust to alterations in the order of the sequence of observations. However, for streaming

---

[12]The choice of threshold governs the range of values that can be considered in the tail. Larger threshold allows wider sample range and therefore, better parameter estimation. However, collecting extreme tail samples using rejection sampling could be difficult when using larger thresholds. It must be noted that optimal choice specific to the data can be made based on the data distribution. In our analysis, we set the threshold to 15% (probability density) for ease of sampling.

[13]The left and right continuous inverses of the function $\frac{1}{1 - \mathrm{G}_0^{\mathrm{EV}}(.)}$ are broadly studied in EVT to understand the behavior of the tail distributions.

[14]It must be noted that the above posterior distribution was derived under the assumption of independence and exchangeability of priors for mathematical ease.

Algorithm 2. Gibbs sampling algorithm when G_0^{EV} is not available

---

Given $z^{(t-1)}_.$, $\left\{ \theta_k^{(t-1)} \right\}$, $\left\{ \theta_k^{a(t-1)} \right\}$ from iteration $(t-1)$. Let K be the total number of clusters at iteration $(t-1)$.

    Set $z_. = |z^{(t-1)}_.|$ and $a_. = sign(z^{(t-1)}_.)$
    for each observation i **do**
        Remove x_i from its cluster z_i.
        if x_i is the only point in its cluster **then**
            Remove the cluster and update K to K-1.
        end if
        Drop empty clusters.
        Sample z_i from the Multinomial distribution given by Equations (40) and (43)
        if $z_i = K + 1$ **then**
            Set the cluster distribution to be multivariate normal with the new cluster mean as x_i and cluster variance as Σ which is pre-defined.
            Update K = K+1.
        end if
        for each cluster $k \in \{1, 2, \ldots, K\}$ **do**
            Sample cluster parameters θ_k and θ_k^a using Equation (44).
        end for
        Sample the anomaly classification a_i from the Binomial(p_i) where p_i is given by

$$p_i = p(x_i) = \begin{cases} \text{Probability of } x_i & x_i \text{ in tail} \\ \quad \text{being anomalous,} & \\ 0, & \text{otherwise} \end{cases} \quad (54)$$

        if most cluster instances are classified as anomalous **then**
            Classify all cluster's instances as anomalies.
        end if
        Set $z_i^{(t)} = z_i * a_i$
    end for

---

data that evolves over time, it can be costly to assume exchangeability among the observations. The instances that mark the beginning of an evolution are captured and monitored in INCAD. Additionally, relapse of outdated and non-prevalent behaviors are identified and evaluated. These features are possible due to the non-exchangeable nature of the INCAD model.

To further understand the non-exchangeable nature of INCAD, one can look at the joint probability of the cluster assignments for the INCAD model,

$$P(z_1, z_2, .z_n | \mathbf{x}) = P(z_1 | \mathbf{x}) P(z_2 | z_1, \mathbf{x}) . P(z_n | z_{1:n-1}, \mathbf{x}). \quad (55)$$

Without loss of generality, let us assume there are K clusters. Let, for any, the joint probability of all the points in cluster k be given by

$$\left(\frac{\alpha * p_{k,1}}{I_{k,1} + \alpha - 1} + \frac{\alpha^* * (1 - p_{k,1})}{I_{k,1} + \alpha^* - 1} \right) \prod_{n_k=2}^{N_k}$$
$$\times \left(\frac{(n_k - 1) * p_{k,n_k}}{I_{k,n_k} + \alpha - 1} + \frac{(n_k - 1) * (1 - p_{k,n_k})}{I_{k,n_k} + \alpha^* - 1} \right), \quad (56)$$

where N_k is the size of the cluster k, $I_{k,i}$ is the index of the ith instance joining the kth cluster and $p_{k,i} = p_{I_{k,i}}$. Thus, the joint probability for complete data is then given by

$$\frac{\prod_{k=1}^{K} \left[(I_{k,1} - 1) p_{k,1} (\alpha - \alpha^*) + \alpha^* (I_{k,1} + \alpha - 1) \right.}{\prod_{i=1}^{N} ((i + \alpha - 1)(i + \alpha^* - 1))}, \quad (57)$$

which is dependent on the order of the data. This shows that the model is not exchangeable unless $\alpha = \alpha^*$ or $p_{k,n_k} = 0$ or $p_{k,n_k} = 1$. These conditions effectively reduce the prior distribution to a traditional CRP model. Hence, it can be concluded that the INCAD model cannot be modified to be exchangeable.

The non-exchangeable and nonparametric prior in the INCAD model serves as an excellent platform to capture drift or evolution in the behavior(s) locally and globally. Such prior can detect the following trends:

1. Instances that signify new evolutionary behavior are captured and classified as anomalous.
2. Increased prevalence in a previously rare behavior can be reevaluated and conceived as normal.[15]
3. Outdated behaviors that are no longer prevalent would be classified as anomalous. Additionally, relapse of such behaviors are also branded as anomalous till sufficient popularity is reached.

A clear streaming extension of the INCAD model involves exclusive reevaluation of the tail instances as opposed to updating with entire data. The Gibbs sampling algorithm for the streaming INCAD model is given in Algorithm 3.

5.2.4 | Choice of priors

For computational ease, the base distribution that generates the parameters for the normal clusters, G_0, is chosen

---

[15] As an alternate frame of reference, one can say that with sufficient surge in the instances, group anomalies can eventually grow to become normal clusters.

Algorithm 3. Algorithm for streaming extension

Perform clustering on a small portion of the data (∼20%) using non-streaming model

> **for** each new data point x_N **do**
>> Compute the mixture proportions m_para and the mixture density for all the data.
>>> Compute $t_1 = q^{th}$ percentile pdf value to identify the tail points
>>> For each x_i s.t. $g(x_i) < t_1$ repeat steps 3 → 18 of Algorithm 2
>>> If cluster size $\leq 0.05 * N$ then, classify all the cluster points as anomalies.
>
> **end for**

---

to be the conjugate of the generative distribution for the actual data, G. This makes the inference task considerably simpler, though approximate methods have been discussed for non-conjugate prior choices as well [32, 40]. In this paper, we use a multivariate normal distribution (MVN) as the data distribution, G, and the Normal Inverse Wishart (NIW) as the base distribution, G_0. It must be noted that the model is not limited to MVN distribution. In particular, any univariate data distribution that satisfies the necessary conditions in Theorem 2 could be used. For multivariate data, distributions from the exponential family satisfy the necessary conditions needed for the Ext-GPD approach. The required conditions for the multivariate case have been presented in the Supporting Information and in Theorem 3.

The concentration parameter, α, and the prior for the base distributions, ψ, are treated as hyper-parameters, though suitable vague priors maybe set to make the model more robust to the choice of the hyper-parameters. α controls the final number of normal clusters, while α_d^* controls the final number of anomalous clusters from the dth DPMM. To ensure that a larger number of populated non-anomalous clusters are formed with few instances assigned to them, α's can be typically set to a higher values.

The parameter γ influences the number of anomalous instances in the dataset, and is initialized based on the expected proportion of anomalies in the given context. For the results listed in this paper, we have used a standard set of the parameter and hyper-parameter choices to show the results in a generalized setting (detailed in Section 6.1). But in other contexts, one can use the information from the data to determine the hyper-parameters. For instance, the γ value can be initially set to the proportion of anomalies known in the data, and the concentration parameter α can be set higher if the true number of clusters is known to be high. It must be noted that the choice of hyper-parameters $\{\alpha_d^*\}$ and parameter γ is updated and optimized using

Extreme Value distributions and Bayesian updates over iterations.

6 | EXPERIMENTAL SETUP

To comprehensively evaluate the capabilities of the proposed INCAD model, results on both synthetically generated and publicly available benchmark datasets are provided. We evaluate the ability of the proposed model to identify both clusters and anomalies, in both batch and streaming settings. We also compare the model performance with existing methods for anomaly detection and clustering. Additionally, we study the role of various user-defined parameters on the model performance.

6.1 | Model initialization

The INCAD model has the following user-defined hyper-parameters: the initial number of clusters (K), the concentration parameter (α), the initial mean and covariance matrices for the clusters, and the prior for the proportion of anomalies (γ). For the experiments, we set K to 10 and α to 1. For each dataset, the sample mean and covariance are used as the initial values for the cluster parameters. The proportion of anomalies (γ) is set to 0.1. In the batch phase, the model is run until convergence is achieved, with a maximum iteration limit of 1000.

6.2 | Data description

We consider a variety of publicly available benchmark datasets from different domains (Table 2) for the experimental evaluation. Additionally, a synthetically generated 2D dataset, SD, with 4 normal clusters and scattered anomalies was generated to evaluate the joint clustering and anomaly detection performance. Each cluster consisted of 100 observations, sampled from a 2D Gaussian distribution with means in $\{(-40, -40), (-30, 10), (40, -60), (45, 30)\}$, for each cluster, respectively. The covariance matrix for each cluster was set to $5I$, where I is the 2×2 identity matrix. Twenty-three anomalies were added by sampling from a Gaussian distribution with mean at $(0, 0)$ and covariance as $100I$. For a qualitative evaluation of the joint clustering and anomaly detection performance, we use the MNIST handwritten digits dataset [37], which consists of 60,000 28×28 images, corresponding to 10 digits (clusters). We use a 10% sample of the original dataset and use principal component analysis (PCA) to reduce the dimensionality of the data from 784 to 25.

Finally, we use the gas sensor array drift dataset [56] to understand the performance of the INCAD model in

TABLE 2 Description of the benchmark datasets used for evaluation of the clustering (*Source*: UCI-ML repository [14]) and anomaly detection (*Source*: Outlier Detection DataSets/ODDS [47]) capabilities of the proposed model

(a) Clustering

Name	N	d	c
Pageb	5473	11	2
Wine-Cluster	6497	12	2
Heart Statlog	270	13	2
Zoo	101	16	7
Abalone	4177	8	2
Magic Gamma	19,020	10	2
Iono	351	33	2
Ecoli	336	7	8
Haberman	306	2	12
Concrete	1030	9	2
German	1000	7	9
Segment	2310	18	7
Iris	150	4	3
Yeast	1484	8	10
WDBC	569	31	2
Vehicle	846	18	4
Glass	214	9	6
Tae	151	3	3
Balance Scale	625	4	3
Vowel	990	10	11

(b) Anomaly detection

Name	N	d	a
Annthyroid	7200	6	7.42%
Pen Global	809	16	11.12%
Cardio	1831	21	9.61%
Mammography	11,183	6	2.32%
Letter	1600	32	6.25%
Seismic Bumps	2584	11	6.58%
Cover	217	10	9.22%
Breast Cancer	367	30	2.72%
Smtp	113	3	11.5%
Wine-AD	129	13	7.75%
Pendigits	6870	16	2.27%

Abbreviations: a, fraction of known anomalies in the dataset; c, number of true clusters; d, number of attributes; N, number of instances.

a streaming setting. The dataset consists of 470 readings from an array of 16 chemical sensors exposed to gas mixtures at three different concentration levels. First two concentration levels were used as the batch dataset and the third concentration level was injected in a streaming fashion.

6.3 | State-of-the-art methods

We compare the performance of INCAD with several existing state-of-art anomaly detection and clustering methods, as well as one method that has been proposed for joint clustering and anomaly detection [11].

Anomaly detection: For anomaly detection, we consider four existing methods: k nearest neighbor outlier detection (kNN) [45], local outlier factor (LOF) [8], one-class support vector machines (oc-SVM) [54], and k-means— [11]. The first two methods assign an anomaly score for each data instance, while the last two methods assign an anomaly label. Both kNN and LOF have been previously shown to outperform other existing methods [30], and are considered state-of-art methods. The k-means— method performs joint clustering and anomaly detection, and thus is the most similar to INCAD. All methods have one or more user-defined parameters. We investigated a range of values for each parameter, and report the mean results.

Clustering: We compare the clustering performance of INCAD with k-means, k-means—, and a Bayesian Gaussian Mixture model with a Dirichlet prior (BGM-DP). While both k-means and k-means— are hard clustering algorithms that require specifying the number of clusters as a user-defined parameter, BGM-DP is a soft clustering
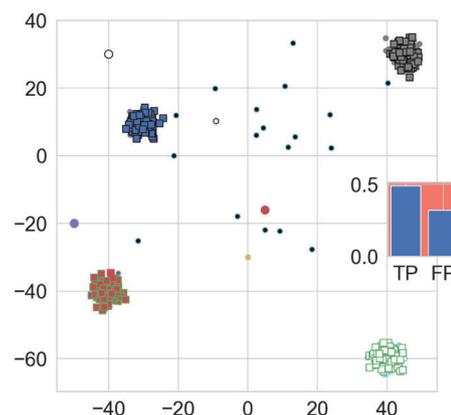

FIGURE 7 INCAD output for the synthetic data, SD. Instances belonging to the normal clusters are shown as □ and instances belonging to anomalous clusters are shown as ○. The size of the anomalous instances indicates the probabilistic anomaly score. *Inset*: the average anomaly score for truly anomalous instances (TP) and false positives (FP)

TABLE 3 Comparing INCAD with existing anomaly detection algorithms using f-measure on the anomaly class as the evaluation metric

Dataset	LOF	KMeans—	KNN	OCSVM	INCAD	INCAD (score)
COVER	**0.36 (± 0.0331)**	0.15 (± 0.0316)	0.15 (± 0.0)	0.15 (± 0.0554)	0.3 (± 0.1613)	0.18 (± 0.0714)
WINE	0.24 (± 0.08)	0.3 (± 0.0)	0.23 (± 0.0943)	0.1 (± 0.0419)	**0.41 (± 0.1941)**	0.1 (± 0.0)
SMTP	**0.59 (± 0.1674)**	0.54 (± 0.0)	0.53 (± 0.0921)	0.21 (± 0.0915)	0.31 (± 0.0669)	0.32 (± 0.102)
PENDIGITS	0.08 (± 0.0075)	**0.19 (± 0.1537)**	0.1 (± 0.0152)	0.06 (± 0.0124)	0.09 (± 0.0365)	0.07 (± 0.0138)
BREAST-CANCER	0.44 (± 0.0165)	**0.6 (± 0.0)**	0.39 (± 0.0598)	0.05 (± 0.0479)	0.19 (± 0.0638)	0.4 (± 0.015)
LETTER	0.44 (± 0.0409)	0.07 (± 0.04)	0.4 (± 0.0779)	0.11 (± 0.0162)	0.28 (± 0.0354)	**0.45 (± 0.0265)**
ANNTHYROID	0.21 (± 0.0121)	0.17 (± 0.0817)	0.3 (± 0.0084)	0.11 (± 0.019)	0.36 (± 0.0254)	**0.39 (± 0.0455)**
PEN-GLOBAL	0.23 (± 0.0365)	0.34 (± 0.0627)	0.25 (± 0.0278)	0.21 (± 0.0497)	**0.53 (± 0.0662)**	0.25 (± 0.0358)
CARDIO	0.21 (± 0.0173)	**0.36 (± 0.3145)**	0.31 (± 0.0772)	0.15 (± 0.0297)	0.2 (± 0.1045)	0.2 (± 0.0838)
MAMMOGRAPHY	0.19 (± 0.0455)	0.12 (± 0.1276)	0.22 (± 0.03)	0.05 (± 0.0354)	0.12 (± 0.0131)	**0.24 (± 0.0216)**
SEISMIC-BUMPS	0.07 (± 0.0113)	0.1 (± 0.0766)	0.15 (± 0.0068)	0.13 (± 0.0304)	**0.23 (± 0.0191)**	0.17 (± 0.0189)

Note: For scoring based methods, instances with top k scores are labeled as anomalous, where k is the actual number of anomalies in the dataset. The average precision and recall on the anomaly class, across all datasets, is shown in the last two rows. Bold values indicate the best performance (in terms of f-measure) across of the models (in each row).

TABLE 4 Comparing INCAD with existing clustering algorithms using purity score as the evaluation metric

Dataset	k-means	k-means—	BGM (DP prior)	INCAD
PAGEB	0.9	0.9	0.94	**0.99 (± 0.0114)**
ABALONE	0.75	**0.81**	0.76	**0.81 (± 0.0139)**
ZOO	**0.87**	0.41	0.64	0.79 (± 0.0913)
WINE	0.63	0.63	0.69	**0.79 (± 0.0719)**
HEART-STATLOG	**0.84**	0.71	0.61	0.79 (± 0.033)
IONO	0.71	0.64	**0.83**	0.79 (± 0.0156)
MAGIC.GAMMA	0.65	0.73	0.77	**0.78 (± 0.0103)**
ECOLI	**0.83**	0.43	0.57	0.76 (± 0.0079)
HABERMAN	**0.75**	0.74	**0.75**	**0.75 (± 0.0069)**
SEGMENT	0.55	0.14	0.52	**0.71 (± 0.0989)**
GERMAN	**0.7**	**0.7**	**0.7**	**0.7 (± 0.0036)**
CONCRETE	0.6	**0.87**	0.65	0.69 (± 0.0324)
IRIS	**0.81**	0.33	0.76	0.67 (± 0.0096)
YEAST	**0.66**	**0.66**	**0.66**	**0.66 (± 0.002)**
WDBC	**0.91**	0.91	0.82	0.63 (± 0.0021)
GLASS	**0.56**	0.36	0.51	0.55 (± 0.0296)
TAE	0.44	0.4	0.44	**0.54 (± 0.0145)**
VEHICLE	0.37	0.35	**0.5**	0.49 (± 0.0416)
BALANCE-SCALE	**0.65**	**0.65**	0.59	0.46 (± 0.0016)
VOWEL	0.33	0.09	0.34	**0.37 (± 0.0587)**
Average purity	0.68	0.57	0.66	**0.69**

FIGURE 8 Output of INCAD for the MNIST 10% sample data. (A) Clusters: Cluster centers identified by INCAD. Note that the number of clusters (18) is automatically inferred by the model. (B) Anomalies: Anomalies identified by INCAD

(A)

(B)
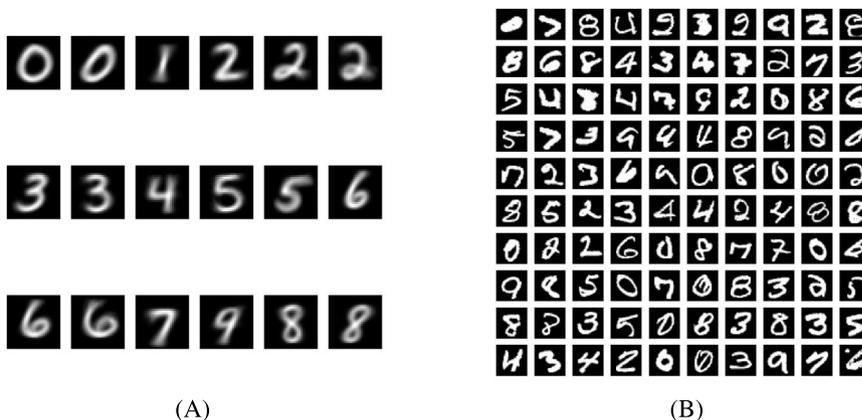
algorithm that does not need the number of clusters to be provided in advance. Thus, it is similar to INCAD in that regard.

6.4 | Evaluation metrics

For the anomaly detection methods that assign an anomaly label to a test instance, that is, oc-SVM, k-means—, and INCAD, f-measure[16] on the anomaly class is used as the evaluation metric. For the scoring methods, that is, kNN, LOF, and the scoring version of INCAD, the instances with top p anomaly scores are labeled as anomalies, and these labels are then used to calculate the f-measure. For the clustering evaluation, we use average cluster purity [11], as the evaluation metric, where the purity of a cluster is defined as the fraction of the majority class of the cluster with respect to the size of the cluster.

7 | RESULTS

In this section, we discuss the overall performance of the INCAD model against the state-of-the-art algorithms with respect to clustering and anomaly detection, in both streaming and batch settings on simulated as well as benchmark datasets.

7.1 | Simulated data

Batch scenario: For a given batch dataset, INCAD produces two types of outputs. First, it assigns every data instance to either a normal cluster (with a positive index) or an anomalous cluster (with a negative index). The sign of the cluster index is used as the anomaly label. Additionally, the method also assigns a probability for each instance to be in the tail of the overall data distribution, which is used as the probabilistic anomaly score. For the SD dataset,

the identified normal and anomalous clusters, as well as the anomaly scores, are shown in Figure 7. We first note that INCAD identifies the four main clusters in the data, without the need to initially specify the number of clusters. Additional anomalous clusters, with negative index, were identified as well. While the method correctly labels all the 23 anomalous instances, it also identified some peripheral instances of the normal clusters as anomalies; these would constitute false positives. However, the probability score is higher for the true anomalies (Figure 7, inset). Thus, simple heuristics, such as a low threshold on the anomaly probability, can be potentially employed, as a post-processing step, to filter out these false positives.

Streaming scenario: To study the performance of INCAD in a streaming mode, we simulate the following streaming scenario: We first create a batch of data consisting of instances belonging to three of the four clusters in SD and present it to INCAD for batch learning. INCAD identifies the three primary clusters, and some of the peripheral instances as local anomalies, after the batch phase (Figure 1A). The instances belonging to the fourth cluster and the anomalies are sequentially presented to the model. With each incoming streaming instance, the tail data is reevaluated and the overall identified data distribution is updated. In the beginning of the streaming phase, the new instances are identified as group anomalies, as shown in Figure 1B. However, a fourth normal cluster is identified after a sufficient number of instances belonging to the fourth cluster are observed in the stream, as shown in Figure 1C. Finally, the remaining truly anomalous instances are identified as global anomalies, as they do not form a tight enough group to become a normal cluster, as shown in Figure 1D.

7.2 | Anomaly detection performance on benchmark datasets

The f-measure performance of INCAD and the competing algorithms is shown in Table 3. For all the listed

---

[16]The class-specific f-measure is defined as the harmonic mean of the recall and precision on the given data set for that class.
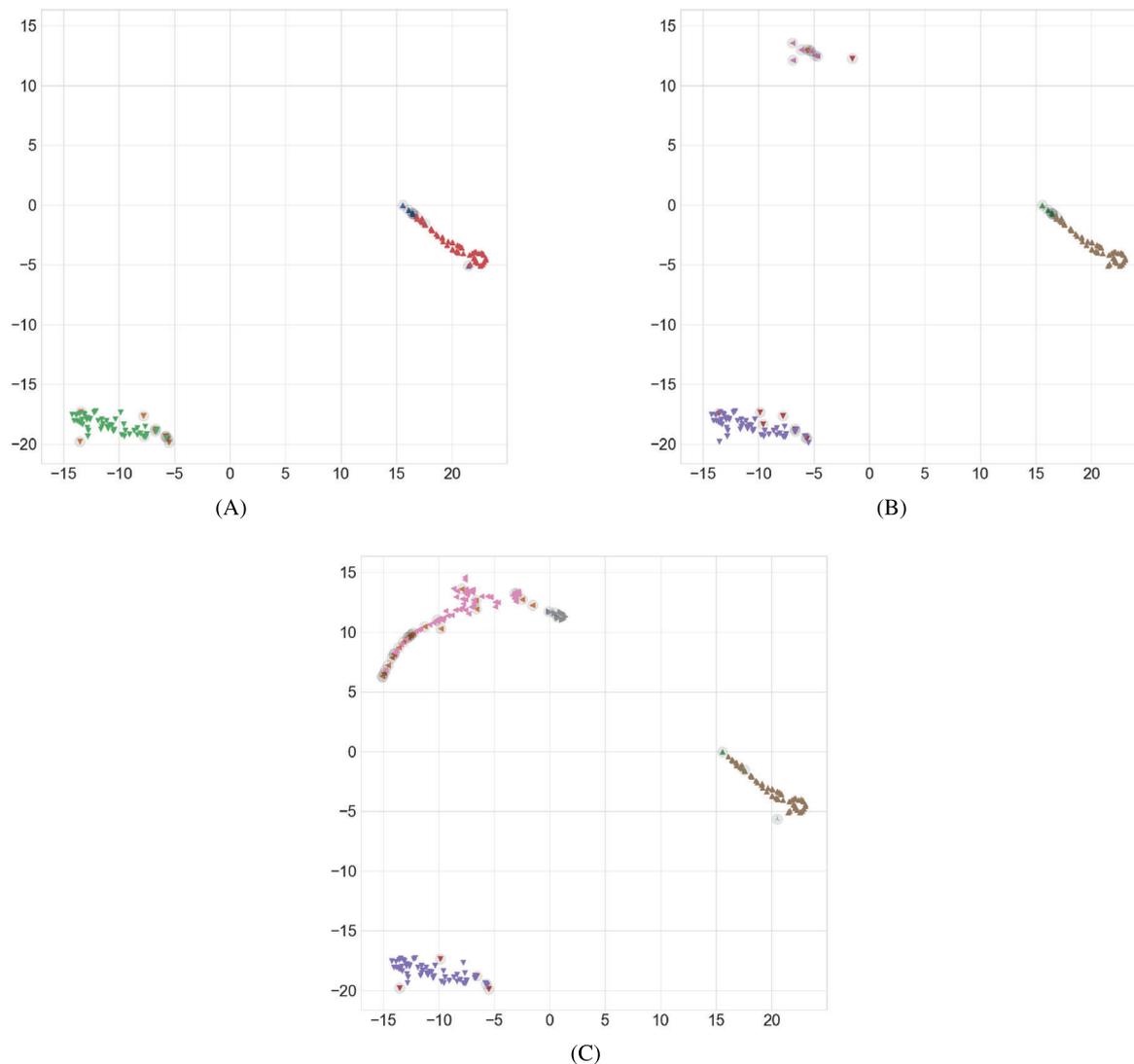
FIGURE 9 Evolving anomalies and clusters identified by INCAD for the gas sensor array drift data. Cluster assignments are shown using colored symbols, anomalous observations are labeled using colored circles. While the original data has 16 dimensions, the data is mapped to 2D using the t-SNE algorithm [38]. (A) Before streaming. (B) After adding 5 streaming observations. (C) After adding all streaming observations

algorithms, results for the best parameter settings are reported. The proposed INCAD model outperforms other methods on 4 out of 11 datasets. While other methods, especially LOF and KNN are better on other datasets, it should be noted that these methods are highly sensitive to the parameter settings. The k-means— method, which is capable of both clustering and anomaly detection, shows the best average performance. However, this model requires specifying the proportion of true anomalies in the dataset, which might not be feasible in a real-world setting.[17]

A specific behavior noticed in the score based INCAD model is the ranking of the anomalies. As INCAD is a conservative algorithm that identifies more anomalies, it can be seen that the model recall is relatively higher than the rest of the methods. However, the true anomalies might not always be ranked as the most anomalous observations. This behavior can be best observed in two particular datasets, namely Pen-Global and Wine data where the score based model has failed to rank most true anomalies in the top while, the classification model still identified some of the true anomalies.

7.3 | Clustering performance on benchmark datasets

Table 4 summarizes the performance of INCAD and other competing clustering methods on the benchmark datasets.

---

[17]For some real datasets with >30% anomalies, smaller clusters identified by INCAD can be manually reclassified as anomalous.

Overall, INCAD has the best average performance compared to others, which is significant, despite not having to provide a prior specification of the expected number of clusters, unlike k-means and k-means—. Looking at both anomaly detection and clustering performance, it is clear that INCAD is effective in detecting both anomalies and clusters in the data, and is superior to k-means—, which also does the joint detection.

To further show the effectiveness of INCAD for the joint detection task, we visualize the detected clusters and anomalies for the MNIST handwritten digit dataset. INCAD identified 18 clusters in the data. The cluster centroids are shown in Figure 8A. The most interesting outcome of clustering using INCAD was the identification subtle writing behaviors identified in the data. For instance, three different writing styles of digits '2' and '6' were identified, which corresponded to distinctive slants, presence of loops, and so forth. The anomalous digits (Figure 8B) identified by INCAD include unrecognizable and ill-written digits.

7.4 | Streaming anomaly detection and clustering: Gas sensor array drift data

The experiment for the gas sensor array drift dataset, simulates a streaming scenario in which a gas at different concentrations is being introduced into a chamber and the concentration levels are being measured by an array of 16 chemical sensors. For these experiments, the observations corresponding to two concentration levels are provided for batch learning, and observations corresponding to the third concentration level are added as a stream. The monitoring outputs of INCAD, at different phases of the stream, are shown in Figure 9. At the end of the batch learning, INCAD is able to identify the two gas concentrations (Figure 9A) present in the batch dataset. After the start of the streaming phase, the new instances are identified as anomalies (Figure 9B), as they belong to a previously unseen concentration. However, as more data is observed in the stream, a new novel cluster is identified (Figure 9C), and all the instances belonging to the third concentration are now considered normal.

7.5 | Sensitivity to batch proportion

Previous results on streaming data show that INCAD can identify anomalies and new clusters in a stream. The performance, however, depends on the size of the initial batch dataset. Figure 10 shows the performance of the model, both in terms of computing time and accuracy in identifying anomalies for the synthetic dataset, SD. While the total size of the dataset is fixed, the proportion of the
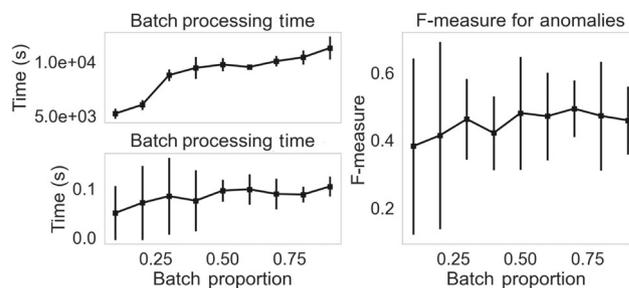
FIGURE 10 Impact of the size of the batch dataset on INCAD performance on the synthetic dataset (SD). For each batch size, mean and standard deviation across five different runs are shown

instances in the batch is varied from 10% to 90%. The computing time[18] for processing the batch increases linearly with the increase in the batch size. At the same time, the time taken to process a single stream instance also increases as the size of the batch increases. This is because the INCAD model has to update the tail probabilities for the data observed so far. The quality of the detected anomalies (shown using the f-measure for the anomalies detected after all of the data is observed), improves as the size of the batch increases. Additionally, the performance is more stable (lower variance across multiple runs) when the batch size is higher because the batch phase is able to learn a stable clustering structure in the data.

8 | CONCLUSIONS AND FUTURE WORK

We have introduced a Bayesian framework for anomaly detection that explicitly models the normal and anomalous data. While in the past, lack of labeled anomalies has prevented such solutions, we adopt concepts from EVT, to model the anomalous data with respect to the extremes of the model for the normal data. This is a fundamental breakthrough in anomaly detection as it permits probabilistic reasoning for both types of instances, without the need for a nonintuitive threshold, as is the case for existing methods. Additionally, the proposed INCAD algorithm combines EVT with another powerful modeling tool—DPMM which allows identifying clusters and anomalies at the same time. The nonparametric prior on the number of clusters ensures that the model is not handicapped by the need to know the exact number of clusters. Moreover, this sets the model up to be adapted for a streaming scenario, where the number of clusters can change over the stream.

---

[18]All the methods are implemented in Python and all experiments were conducted on a 2.7 GHz Quad-Core Intel Core i7 processor with a 16 GB RAM.

As the results show, INCAD outperforms existing methods that have been proposed exclusively for anomaly detection or clustering, on each of the tasks, for most of the datasets (Tables 3 and 4). Moreover, while existing methods rely on carefully specified, problem-specific, parameters, INCAD requires specifying relaxed Bayesian priors, and infers key parameters, such as the number of clusters, from the data. Additionally, the probabilistic output of INCAD allows for an interpretable setting of thresholds on the anomaly score, something that is not possible with most of the existing score based anomaly detection algorithms. INCAD is especially effective in dealing with streaming data, where the notion of normal clusters and anomalies evolve over the duration of the stream, as shown in Figure 9. This makes INCAD highly suitable for monitoring the behavior of complex systems over time, without the need to explicitly retrain the underlying model.

One of the key shortcomings of the model is the complexity of the iterative Gibbs algorithm. Variational inference methods that have been proposed for inference in DPMM clustering [7, 35] can be used to improve the complexity, and will be explored in the future.

DATA AVAILABILITY STATEMENT
Data openly available in a public repositories that issues datasets with DOIs.

ORCID
Sreelekha Guggilam https://orcid.org/0000-0002-7795-2945
Varun Chandola https://orcid.org/0000-0001-8990-1398

REFERENCES
1. S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, *Unsupervised real-time anomaly detection for streaming data*, Neurocomputing 262 (2017), 134–147.
2. H. Al-Behadili, A. Grumpe, L. Migdadi, and C. Wöhler, *Semi-supervised learning using incremental support vector machine and extreme value theory in gesture data*, in *2016 UKSim-AMSS 18th Int. Conf. Comput. Model. Simul. (UKSim)*, IEEE, Cambridge, England, 2016, 184–189.
3. C. E. Antoniak, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, Ann. Stat. 2 (1974), 1152–1174.
4. A. Bendale and T. Boult, *Towards open world recognition*, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, IEEE Computer Society, Los Alamitos, CA, 2015.
5. D. M. Blei and P. I. Frazier, *Distance dependent Chinese restaurant processes*, in *Int. Conf. Mach. Learn.*, Omnipress, Madison, WI, 2010, 87–94.
6. D. M. Blei, T. L. Griffiths, and M. I. Jordan, *The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies*, J. ACM 57 (2010), 7.
7. D. M. Blei and M. I. Jordan, *Variational methods for the Dirichlet process*, in Int. Conf. Mach. Learn., Association for Computing Machinery, New York, NY, 2004.
8. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, *Lof: Identifying density-based local outliers*, in *Proceedings of 2000 ACM SIGMOD Int. Conf. Manage. Data*, Association for Computing Machinery, New York, NY, 2000, 93–104.
9. V. Chandola, A. Banerjee, and V. Kumar, *Anomaly detection: A survey*, ACM Comput. Surv. 41 (2009) no. 3. 58.
10. M. Charras-Garrido and P. Lezaud, *Extreme value analysis: An introduction*, J. Soc. Fr. Stat. 154 (2013), 66.
11. S. Chawla and A. Gionis, k-means−: A unified approach to clustering and outlier detection, in *Proc. 2013 SIAM Int. Conf. Data Mining*, SIAM, Austin, TX, 2013, 189–197.
12. D. A. Clifton, L. Clifton, S. Hugueny, and L. Tarassenko, *Extending the generalised Pareto distribution for novelty detection in high-dimensional spaces*, J. Signal Process. Syst. 74 (2014), 323–339.
13. L. De Haan and A. Ferreira, *Extreme value theory: An introduction*, Springer Science & Business Media, Springer, New York, NY, 2007. https://doi.org/10.1007/0-387-34471-3
14. D. Dheeru and E. Karra Taniskidou, *UCI machine learning repository*. http://archive.ics.uci.edu/ml, 2017.
15. E. Di Lello, T. De Laet, and H. Bruyninckx, *Hierarchical Dirichlet process hidden markov models for abnormality detection in robotic assembly*. Neural Inf. Process. Syst., 2012/12/03–2012/12/08, Lake Tahoe, Nevada, 2012.
16. E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, *A geometric framework for unsupervised anomaly detection*, in *Applications of data mining in computer security*, Springer, Boston, MA, 2002, 77–101. https://doi.org/10.1007/978-1-4615-0953-0_4.
17. T. S. Ferguson, *A bayesian analysis of some nonparametric problems*, Ann. Stat. 1 (1973), 209–230.
18. R. A. Fisher and L. H. C. Tippett, *Limiting forms of the frequency distribution of the largest or smallest member of a sample*, Math. Proc. Camb. Philos. Soc. 24 (1928), 180–190. https://doi.org/10.1017/S0305004100015681
19. J. Franzen, *Bayesian inference for a mixture model using the gibbs sampler*, Research Report RR 2006:1, Stockholm University, 2006.
20. J. French, P. Kokoszka, S. Stoev, and L. Hall, *Quantifying the risk of heat waves using extreme value theory and spatio-temporal functional data*, Comput. Stat. Data Anal. 131 (2019), 176–193.
21. B. A. Frigyik, A. Kapila, and M. R. Gupta, *Introduction to the Dirichlet distribution and related processes*, University of Washington, Seattle, Washington, 2010, 206.

22. R. Fujimaki, T. Yairi, and K. Machida, *An approach to space-craft anomaly detection problem using kernel feature space*, in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Association for Computing Machinery, New York, NY, 2005, 401–410.

23. T. Fuse and K. Kamiya, *Statistical anomaly detection in human dynamics monitoring using a hierarchical Dirichlet process hidden markov model*, IEEE Trans. Intell. Transp. Syst. 18 (2017), 3083–3092.

24. G. Gan and M. K.-P. Ng, *K-means clustering with outlier removal*, Pattern Recogn. Lett. 90 (2017), 8–14.

25. Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, 2017: *Generative openmax for multi-class open set classification*.

26. I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, *Em algorithms for weighted-data clustering with application to audio-visual scene analysis*, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016), 2402–2415.

27. M. Ghesmoune, M. Lebbah, and H. Azzag, *State-of-the-art on clustering data streams*, Big Data Anal. 1 (2016), 13.

28. B. Gnedenko, *Sur la distribution limite du terme maximum d'une serie aleatoire*, Ann. Math. 44 (1943), 423–453.

29. N. Goix, A. Sabourin, and S. Clemencon, *Sparse representation of multivariate extremes with applications to anomaly ranking*, in *Proc. 19th Int. Conf. Artif. Intell. Stat. PMLR, Cadiz, Spain*, Vol 51, A. Gretton and C. C. Robert, Eds., Proceedings of Machine Learning Research, Cadiz, Spain, 2016, 75–83.

30. M. Goldstein and S. Uchida, *A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data*, PLoS One 11 (2016), e0152173.

31. N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, *Toward supervised anomaly detection*, J. Artif. Intell. Res. 46 (2013), 235–262.

32. D. Görür and C. E. Rasmussen, *Dirichlet process Gaussian mixture models: Choice of the base distribution*, J. Comput. Sci. Technol. 25 (2010), 653–664.

33. S. Guggilam, S. Zaidi, V. Chandola, and A. K. Patra, *Integrated clustering and anomaly detection (incad) for streaming data (revised)*, arXiv preprint arXiv:1911.00184, 2019.

34. Z. He, X. Xu, and S. Deng, *Discovering cluster-based local outliers*, Pattern Recogn. Lett. 24 (2003), 1641–1650.

35. V. Huynh, D. Phung, and S. Venkatesh, *Streaming variational inference for Dirichlet process mixtures*, in *Asian Conf. Mach. Learn.*, Proceedings of Machine Learning Research, Hong Kong, 2016, 237–252.

36. S. Jiang, X. Song, H. Wang, J.-J. Han, and Q.-H. Li, *A clustering-based method for unsupervised intrusion detections*, Pattern Recogn. Lett. 27 (2006), 802–810.

37. Y. LeCun and C. Cortes, *MNIST handwritten digit database*. http://yann.lecun.com/exdb/mnist/, 2010.

38. L.v.d. Maaten and G. Hinton, *Visualizing data using t-sne*, J. Mach. Learn. Res. 9 (2008), 2579–2605.

39. D. J. Marchette, *A statistical method for profiling network traffic*, in *Workshop Intrusion Detect. Netw. Monit.*, USENIX Association, Santa Clara, CA, 1999, 119–128.

40. R. M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, J. Comput. Graph. Stat. 9 (2000), 249–265.

41. H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng, *A survey on data stream clustering and classification*, Knowl. Inf. Syst. 45 (2015), 535–569.

42. L. Ott, L. X. Pang, F. T. Ramos, and S. Chawla, *On integrated clustering and outlier detection*, NeurIPS Proceedings 27 (2014), 1359–1367.

43. N. Pandeeswari and G. Kumar, *Anomaly detection system in cloud environment using fuzzy clustering based ann*, Mobile Netw. Appl. 21 (2016), 494–505.

44. J. Pickands, *Statistical inference using extreme order statistics*, Ann. Stat. 1 (1975), 119–131.

45. S. Ramaswamy, R. Rastogi, and K. Shim, *Efficient algorithms for mining outliers from large data sets*, in *Proc. 2000 ACM SIGMOD Int. Conf. Manage. Data*, ACM Press, Dallas, TX, 2000, 427–438.

46. C. E. Rasmussen, *The infinite Gaussian mixture model*, in *Advances in neural information processing systems (NIPS)*, MIT Press, Cambridge, MA, 2000, 554–560.

47. S. Rayana, *ODDS library*. http://odds.cs.stonybrook.edu, 2016.

48. S. Sadik and L. Gruenwald, *Research issues in outlier detection for data streams*, ACM SIGKDD Explor. Newsl. 15 (2014), 33–40.

49. M. S. Shotwell and E. H. Slate, *Bayesian outlier detection with Dirichlet process mixtures*, Bayesian Anal. 6 (2011), 665–690.

50. A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet, *Anomaly detection in streams with extreme value theory*, in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Association for Computing Machinery, New York, NY, 2017, 1067–1075.

51. M. Smith, S. Reece, S. Roberts, and I. Rezek, *Online maritime abnormality detection using Gaussian processes and extreme value theory*, in *2012 IEEE 12th Int. Conf. Data Mining*, IEEE, Washington, D.C., 2012, 645–654.

52. P. Talagala, R. Hyndman, K. Smith-Miles, S. Kandanaarachchi, M. Munoz. *Anomaly detection in streaming nonstationary temporal data*, 29, 1, Taylor & Francis, Oxfordshire, England, 2020, 13–27.

53. S. C. Tan, K. M. Ting, and T. F. Liu, *Fast anomaly detection for streaming data*, in *22nd Int. Joint Conf. Artif. Intell*, AAAI Press, Catalonia, Spain, 2011.

54. D. M. J. Tax and R. P. W. Duin, *Support vector data description*, Mach. Learn. 54 (2004), 45–66.

55. J. Varadarajan, R. Subramanian, N. Ahuja, P. Moulin, and J.-M. Odobez, *Active online anomaly detection using Dirichlet process mixture model and Gaussian process classification*, in *2017 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, IEEE, Washington, D.C., 2017, 615–623.

56. A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, *Chemical gas sensor drift compensation using classifier ensembles*, Sensors Actuators B Chem. 166 (2012), 320–329.

57. A. Whitmore, A. Agarwal, and L. Da Xu, *The internet of things—A survey of topics and trends*, Inf. Syst. Front. 17 (2015), 261–274.

58. K. Wu, K. Zhang, W. Fan, A. Edwards, and S. Y. Philip, *Rs-forest: A rapid density estimator for streaming anomaly detection*, in *2014 IEEE Int. Conf. Data Mining*, IEEE, Washington, D.C., 2014, 600–609.

59. L. Xiong, B. Póczos, J. Schneider, A. Connolly, and J. VanderPlas, *Hierarchical probabilistic models for group anomaly detection*, in *Proc. 14th Int. Conf. Artif. Intell. Stat.*, PMLR, Fort Lauderdale, FL, 2011, 789–797.

60. H. Z. Yerebakan, B. Rajwa, and M. Dundar, *The infinite mixture of infinite Gaussian mixtures*, Adv. Neural Inf. Proces. Syst. 27 (2014), 28–36.

APPENDIX

Proof of Ext-GPD for n-D case

Let $X \in \mathbb{R}^n$ be the data space with pdf $g_X : \mathbb{R}^n \to \mathbb{R}^+$. Let $Y \in \mathbb{R}^+$ be the corresponding image space.

Definition 5. $\forall y \in Y$, G_Y is defined as

$$G_Y(y) = \int_{g_X^{-1}([0,y])} g_X(x)dx. \tag{A.1}$$

Claim 5. G_Y is a CDF.

As the limit distribution of the minima of Y is of interest, we wish to study the limit distribution of maxima of $Z = Y_m - Y$. Then the CDF of Z is given by G_Z is

$$
\begin{aligned}
G_Z(z) &= P(Z \le z) \\
&= P(Y_m - Y \le z) \\
&= P(Y \ge Y_m - z) \\
&= 1 - G_Y(Y_m - z) \\
&= \int_{g_X^{-1}([Y_m-z,Y_m])} g_X(x)dx.
\end{aligned}
\tag{A.2}
$$

$\forall z \in [0, Y_m]$.

For, G_Z, the corresponding maximum value, $x^* = Y_m$.

We need the necessary conditions for the above distribution to be in the domain of attraction of a GEV distribution. By von Mises' condition, if we can prove that G_Z' is positive and G_Z'' exists in some neighborhood of Y_m, then G_Z is in domain of attraction of G_γ.

Proof Let $\vec{X} \in \mathbf{R}^n$ and $g_{\vec{X}}^{-1}([0, Y_m - z]) = D(Y_m - z)$ be a n-manifold with a boundary $\partial D(Y_m - z)$. Then,

$$
\begin{aligned}
G_Z'(z) &= \frac{d}{dz} \int_{g_{\vec{X}}^{-1}([Y_m-z,Y_m])} g_{\vec{X}}(\vec{x})d\vec{x} \\
&= \frac{d}{dz} \left[1 - \int_{g_{\vec{X}}^{-1}([0,Y_m-z])} g_{\vec{X}}(\vec{x})d\vec{x} \right] \\
&= -\frac{d}{dz} \int_{D(Y_m-z)} g_{\vec{X}}(\vec{x})d\vec{x},
\end{aligned}
\tag{A.3}
$$

where $d\vec{x} = dx_1 \wedge dx_2 \wedge \ldots \wedge dx_n$.

Then, using Reynolds transport theorem, we get,

$$
\begin{aligned}
\frac{d}{dz} G(z) &= \frac{d}{dz} \int_D g_X(\vec{x})d\vec{x} \\
&= \int_{D(Y_m-z)} \frac{\partial}{\partial z} g_X(\vec{x}) dV + \int_{\partial D(Y_m-z)} g_X(\vec{x})\vec{v}_b \cdot d\Sigma,
\end{aligned}
\tag{A.4}
$$

where $g_X(\vec{x})$, $D(Y_m - z)$ and $\partial D(Y_m - z)$ are as defined above, $\vec{v}_b = \frac{dD(Y_m-z)}{dz}$ is the Eulerian velocity of the boundary, $\mathbf{n}$ is the outward unit normal, dS is the surface element in $\mathbf{R}^d$ and $d\Sigma = \mathbf{n}dS$.

Since, $\frac{\partial}{\partial z} g_X(\vec{x}) = 0$,

$$G_Z'(z) = \int_{\partial D(Y_m-z)} g_X(\vec{x})\vec{v}_b \cdot d\Sigma \tag{A.5}$$

Claim 6: G_Z'' exists.

$$
\begin{aligned}
G_Z''(z) &= \frac{d}{dz} G_Z'(z) \\
&= \frac{d}{dz} \int_{\partial D(Y_m-z)} g_X(\vec{x})\vec{v}_b \cdot d\Sigma.
\end{aligned}
\tag{A.6}
$$

Since, $\partial D(Y_m - z)$ an $(n - 1)$-closed manifold, that is, $(n - 1)$-manifold without a boundary, we use the general statement of the Leibniz integral rule to compute the second order derivative,

$$
\begin{aligned}
G_Z''(z) &= \frac{d}{dz} G_Z'(z) \\
&= \frac{d}{dz} \int_{\partial D(Y_m-z)} g_X(\vec{x})\vec{v}_b \cdot d\Sigma \\
&= \int_{\partial D(Y_m-z)} i_{\vec{v}} \left(d_x \left[g_X(\vec{x})\vec{v}_b \cdot d\Sigma \right] \right),
\end{aligned}
\tag{A.7}
$$

where $d_x f$ is the exterior derivative of f w.r.t space variables only, $\vec{v} = \frac{\partial \vec{x}}{\partial z}$ is the vector field of the velocity and $i_{\vec{v}}$ denotes the interior product with $\vec{v}$.

Thus, it can be seen that G_Z is in the maximum domain of attraction of a GEV distribution iff:

1. $D(Y_m - z)$ is an n-manifold with a boundary $\partial D(Y_m - z)$,
2. The Eulerian velocity of the boundary $\vec{v}_b = \frac{dD(Y_m-z)}{dz}$ exists,
3. $d_x \left[g_X(\vec{x})\vec{v}_b \cdot d\Sigma \right]$ exists, and
4. $i_{\vec{v}} \left(d_x \left[g_X(\vec{x})\vec{v}_b \cdot d\Sigma \right] \right)$ exits.