Received: 14 April 2022 | Revised: 28 July 2022 | Accepted: 8 December 2022

DOI: 10.1002/widm.1491

FOCUS ARTICLE

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

A geometric framework for outlier detection in high-dimensional data

Moritz Herrmann | **Florian Pfisterer** | **Fabian Scheipl**

Department of Statistics, Ludwig Maximilians University, Munich, Germany

Correspondence
Moritz Herrmann, Department of Statistics, Ludwig Maximilians University Munich, Ludwigstr. 33, D-80539, Munich, Germany.
Email: moritz.herrmann@stat.uni-muenchen.de

Funding information
German Federal Ministry of Education and Research (BMBF), Grant/Award Number: 01IS18036A

Edited by: Sushmita Mitra, Associate Editor, and Witold Pedrycz, Editor-in-Chief

Abstract

Outlier or anomaly detection is an important task in data analysis. We discuss the problem from a geometrical perspective and provide a framework which exploits the metric structure of a data set. Our approach rests on the *manifold assumption*, that is, that the observed, nominally high-dimensional data lie on a much lower dimensional manifold and that this intrinsic structure can be inferred with manifold learning methods. We show that exploiting this structure significantly improves the detection of outlying observations in high dimensional data. We also suggest a novel, mathematically precise and widely applicable distinction between *distributional* and *structural* outliers based on the geometry and topology of the data manifold that clarifies conceptual ambiguities prevalent throughout the literature. Our experiments focus on functional data as one class of structured high-dimensional data, but the framework we propose is completely general and we include image and graph data applications. Our results show that the outlier structure of high-dimensional and non-tabular data can be detected and visualized using manifold learning methods and quantified using standard outlier scoring methods applied to the manifold embedding vectors.

This article is categorized under:
Technologies > Structure Discovery and Clustering
Fundamental Concepts of Data and Knowledge > Data Concepts
Technologies > Visualization

KEYWORDS
anomaly detection, dimension reduction, manifold learning, outlier detection

1 | INTRODUCTION

Detecting atypical observations that deviate substantially from the bulk of the data is an important task in data analysis with applications across domains like, for example, intrusion detection (Zhang & Zulkernine, 2006), medical imaging (Fritsch et al., 2012), or network analysis (Azcorra et al., 2018). The most common terms for this task are *outlier* or *anomaly detection*, but many different terms are used (Zimek & Filzmoser, 2018). Although there is a vast amount of

literature on the topic, there is neither a commonly accepted, precise definition of what exactly constitutes outliers or anomalies, nor agreement on whether these two terms are synonymous. As Unwin (2019, p. 635) puts it:

> "Outliers are a complicated business. It is difficult to define what they are, it is difficult to identify them, and it is difficult to assess how they affect analyses."

Overviews on the topic are given by Zimek et al. (2012) or Goldstein and Uchida (2016) from a computer science perspective, and by Rousseeuw and Leroy (2005) or Unwin (2019) from a statistical perspective. Kandanaarachchi and Hyndman (2020) provide a short summary including both perspectives, while Campos et al. (2016) as well as Marques et al. (2020) focus on the evaluation of unsupervised outlier detection. Zimek and Filzmoser (2018) provide a comprehensive survey bringing together both perspectives with in-depth epistemological discussion. In particular, Zimek and Filzmoser (2018) discuss that there are two different notions of outliers and different terms used to describe these notions—including, for example, *apparent*, *discrepant*, *real*, *contaminating*, or *true* outlier—in the literature. From this discussion, it can be inferred that (1) Zimek & Filzmoser (2018, p. 7) distinguish "true" and "apparent" outliers and consider "those objects as '(true) outliers' that have been 'generated by a different mechanism' than the remainder or major part of the data or than the whatsoever defined reference set", (2) there is neither a clear understanding of how these two notions are different and actually manifest in practice nor (3) a "language" to precisely describe the problem theoretically.

In the more statistically flavored literature, the problem of unsupervised outlier detection is usually tackled by defining outliers based on a single probability distribution P. If P allows for a density, outliers are simply observations in low-density regions. From this perspective, we have *distributional outliers* whose outlyingness is defined relative to a single probability distribution. The notion of *distributional outliers* is easy to define precisely in probabilistic terms, for example, based on minimum level sets (Scott & Nowak, 2006) or M-estimation (Clémençon & Jakubowicz, 2013), and has yielded a multitude of results and algorithms. In practical terms, this requires access to (an estimate of) the underlying density and finding a suitable (local) density level below which observations are to be classified as outliers. Note that both are infeasible for general, non-tabular data types like shapes, functions, or images whose domains frequently do not admit probability densities. However, Zimek and Filzmoser (2018) emphasize that "observations which are in the extremes of the model distribution (i.e., distributional outlier) should be distinguished from 'real' outliers (contaminants)" (Zimek & Filzmoser, 2018, p. 13). This second notion of outliers ("true" or "real" outliers) is not reflected by the statistical concept because such outliers are assumed to be observations generated by a different data-generating process. This is reflected in statements like "different mechanism" or "any observation that is not a realization from the target distribution" (Beckman & Cook, 1983, p. 121). That means, for the second notion ("real outliers") it is implicitly assumed that outliers are not independent and identically distributed (IID) observations. So next to *distributional outliers* there are also *structural outliers* whose outlyingness is caused by the structural differences between the underlying data generating processes. The two outlier types are complementary and both are necessary to fully address the challenges of outlier detection. In contrast to *distributional outliers*, *structural outliers* are much more difficult to formalize, but also more general.

With this work, we intend to "broaden" the view on the problem of unsupervised outlier detection to account for the two notions of outliers present in the literature. We show that a *geometric* approach to the problem, which does not require the availability of probability densities defined over the data space but only some metric structure (i.e., suitable dissimilarity or distance measures), allows for a more precise conceptualization and terminology. To do so, we focus on building up intuition and demonstrating the application of these concepts to diverse and comprehensive practical examples and visualizations. However, to be able to "speak" of this new perspective without referring to vague and subjective perceptions as done previously (see Zimek & Filzmoser, 2018), we also consider it necessary to introduce a certain degree of mathematical terminology. The provided degree of formality is exhausted in the definition of *distributional* and *structural* outliers in precise mathematical terms and thus serves the need for precise terminology but does not overload the work with more formalism than we think necessary to contribute to the overall scope of the study. Readers interested in more rigorous mathematical approaches to infer structures in data may, for example, consult Mordohai and Medioni (2010) for dimensionality estimation and manifold learning based on tensor voting, Niyogi et al. (2011) for a topological perspective on unsupervised learning, Guan and Loew (2021) of a distance-based measure of class separability, and Kandanaarachchi and Hyndman (2020) for outlier detection in tabular data based on dimensionality reduction.

The rest of the paper is structured as follows. Section 2 describes the scope and contribution of the study and outlines its background and related work. The proposed theoretical framework is defined in Section 3 and its practical

relevance is demonstrated in Section 4 using qualitative and quantitative experiments for a variety of data sets of different data types. Section 5 discusses our findings and the resulting conceptual implications, before we conclude in Section 6.

2 | PRELIMINARIES

2.1 | Scope and contribution of the study

With this focus article, we intend to draw connections between different conceptual aspects provided in the overview article by Zimek and Filzmoser (2018) and as proposed by us in a paper focusing on functional data analysis (FDA). Therefore, we recapitulate the underlying conceptualization presented in the earlier paper in a more general form and different terminology in Section 3. This framework builds on principles from *manifold learning* (Lee & Verleysen, 2007; Ma & Fu, 2011), that is, dimension reduction methods that infer the intrinsic lower-dimensional manifold structure of high-dimensional data and yield low-dimensional vector representations of the data. This perspective allows us to formalize structural and distributional outliers jointly in a single mathematical framework, where structural outliers are data that are separate from the main data manifold, and distributional outliers are data that are situated at the periphery of, but still on the main data manifold. While the first paper exclusively focused on the functional data setting, the present focus article generalizes the underlying conceptualization of outlier detection to other data types. This is straightforward theoretically but has important general conceptual implications that have never been described in detail and demonstrated on diverse real data problems before. In particular, we draw connections between and provide a unifying perspective on different data types (functions, images, graphs, tabular data) which were previously often treated separately from a theoretical as well as a practical perspective, in particular when it comes to outlier detection.

The main contribution of this review paper is to discuss and demonstrate two conceptual aspects of outlier detection in general. First of all, as already outlined, there seems to be a lack of clarity about what defines outliers, evidenced also by the plethora of terms used to describe the issue (Zimek & Filzmoser, 2018). Several recent reviews on the topic also point out this conceptual ambiguity (Goldstein & Uchida, 2016; Unwin, 2019; Zimek & Filzmoser, 2018). In particular, the comprehensive overview of Zimek and Filzmoser (2018, p. 4) devotes a complete section to the question of "what an 'outlier' possibly means". Recall that they define "true outliers" as objects "that have been 'generated by a different mechanism' than the remainder or major part of the data or than the whatsoever defined reference set" and distinguish them from "objects that appear to be outliers (independent of whether or not they actually are (true) outliers)" (Zimek & Filzmoser, 2018, p. 7). As we will show, the geometrical framework provides suitable mathematical terminology to delineate "true" and "apparent" outliers much more cleanly and thus reduces the conceptual ambiguity that surrounds the topic: We transfer concepts established in manifold learning to the problem of outlier detection, deriving a novel underlying conceptualization of the problem of outlier detection that (1) is capable of reflecting two types of outliers, (2) replaces vague notions of "real", "contaminant" or "apparent" outliers with a precise definition, and (3) incorporates the well-established concept of *distributional* outliers in a unified fashion. With this, we can abandon vague notions of outlier subtypes in favor of two precisely defined concepts.

Second, our framework also suggests that outlier detection in high-dimensional (and/or non-tabular) data is not necessarily more challenging than in low-dimensional settings once the underlying manifold structure is recovered and exploited. This is important because high dimensionality is often reported to be particularly problematic for outlier detection and many outlier detection methods break down or at least face particular challenges in such settings (e.g., Aggarwal, 2017; Aggarwal & Yu, 2001; Goldstein & Uchida, 2016; Kamalov & Leung, 2020; Navarro-Esteban & Cuesta-Albertos, 2021; Ro et al., 2015; Thudumu et al., 2020; Xu et al., 2018; Zimek et al., 2012).

To highlight these aspects, we again provide simple and easily accessible functional data examples to demonstrate the principal practical implications (in addition to the recapitulation of the theoretical conceptualization). Functional data analysis (Ramsay & Silverman, 2005) deals with data that are (discretized) realizations of stochastic processes over a compact domain. Functional data is well suited to illustrate the underlying conceptualization both practically and theoretically because it is usually highly structured (the manifold assumption is specifically realistic and useful), theoretically/analytically well accessible, and easily visualized in bulk. Beyond the FDA setting, we also use examples of other data types including image, graph, curve, and tabular data.

Finally, our framework is fully general and does not rely on a specific combination of manifold learning and outlier detection methods. To demonstrate its practical performance, we show that one of the simplest and most established

manifold learning methods—Multidimensional Scaling (MDS) (Cox & Cox, 2008)—combined with a standard outlier detection algorithm—Local Outlier Factors (LOF) (Breunig et al., 2000)—already yields a flexible, reliable and generally applicable recipe for outlier detection and visualization in complex, high-dimensional data.

2.2 | Background and related work

The fundamental assumption of manifold learning is that the high-dimensional data observed in a D-dimensional space $\mathcal{H}$ actually lie on a d-dimensional manifold $\mathcal{M} \subset \mathcal{H}$ with $d < D$. Manifold learning methods yield an *embedding* function $e : \mathcal{H} \to \mathcal{Y}$ from the high-dimensional data space to a low-dimensional embedding space $\mathcal{Y}$ such that the configuration of embedded data reflects the characteristics of $\mathcal{M}$. The terms manifold learning and nonlinear dimension reduction are often used interchangeably (Lee & Verleysen, 2007; Ma & Fu, 2011). Typically, the fundamental step is to compute distances between the high-dimensional observations. Methods based on this approach are, for example, Multi-dimensional Scaling (MDS) (Cox & Cox, 2008), Isomap (Tenenbaum et al., 2000), diffusion maps (Coifman & Lafon, 2006), local linear embeddings (Roweis & Saul, 2000), Laplacian eigenmaps (Belkin & Niyogi, 2003), t-distributed stochastic neighborhood embeddings (t-SNE) (Maaten & Maaten & Hinton, 2008), and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018), to name only a few. The methods differ in how they infer the manifold structure from these distances and how they obtain low-dimensional embedding vectors from these.

Despite their promising results in other settings, manifold learning methods have not found application for outlier detection to a significant extent so far. Kandanaarachchi and Hyndman (2020) define an outlier detection method explicitly based on dimension reduction, while Pang et al. (2018) make use of ranking model-based representation learning. However, they do not provide a general conceptual framework and focus on tabular data. For functional data, Xie et al. (2017) introduce a geometric approach that decomposes functional observations into amplitude, phase, and shift components in order to identify specific types of outliers. However, the approach is only applicable to functional data and does not make use of the intrinsic structure of the functional observations from a manifold learning perspective. Ali et al. (2019) analyze time series data using 2**D**-embeddings obtained from manifold methods for outlier detection and clustering and Toivola et al. (2010) compare specific dimensionality reduction techniques for outlier detection in structural health monitoring, but both focus on practical considerations and do not provide a theoretical underpinning. Another line of work focuses on projection-based outlier detection, for example for high-dimensional Gaussian data (Navarro-Esteban & Cuesta-Albertos, 2021), financial time series (Loperfido, 2020) or functional data (Ren et al., 2017).

3 | GEOMETRICAL FRAMEWORK FOR OUTLIER DETECTION

The framework we propose generalizes an approach for outlier detection in functional data developed recently (Herrmann & Scheipl, 2021). Since the approach exploits the metric structure of a functional data set, it is straightforward to generalize it to other data types, both from a theoretical as well as a practical perspective. Theoretically, the observation space needs to be a metric space, that is, it needs to be equipped with a metric. Practically, there only needs to be a suitable distance measure to compute pairwise distances between observations. Two assumptions are fundamental for the framework. First of all, the *manifold assumption* that observed high-dimensional data lie on or close to a (low-dimensional) manifold. Note that functional data typically contain a lot of structure, and it is often reasonable to assume that only a few modes of variation suffice to describe most of the information contained in the data, that is, such functional data often have low intrinsic dimension, at least approximately, see Figure 1 for a simple synthetic example. Similar remarks hold for other data types such as image data (Lee & Verleysen, 2007; Ma & Fu, 2011). Second, it is assumed that outliers are either *structural outliers*—or in the terminology of Zimek and Filzmoser (2018, p. 10) "real outliers" stemming from a different data generating process than the bulk of the data—or *distributional* outliers, observations that are structurally similar to the main data but still appear outlying in some sense. We make these notions mathematically precise in the remainder of this section based on the exposition in Herrmann and Scheipl (2021) before we demonstrate the practical relevance of the framework in Section 4 and summarize its general conceptual implications in Section 5.

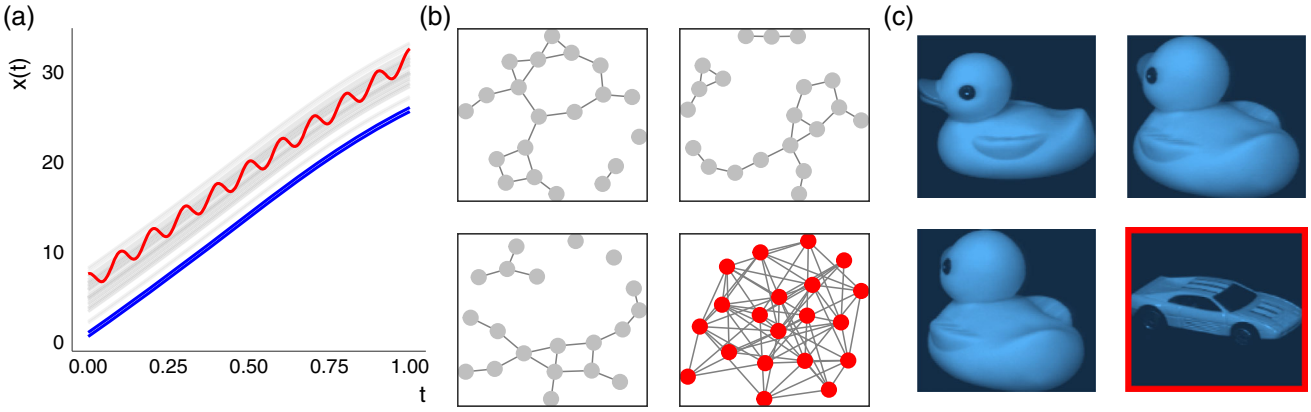

FIGURE 1 Example data types. (a) Functional inliers (gray) with a structural outlier (red) and distributional outliers (blue). (b) Graph data with a structural outlier (lower right graph). (c) Image data with a structural outlier (lower right image).

Given a high-dimensional observation space $\mathcal{H}$ of dimension D, a d-dimensional parameter space $\Theta \subset \mathbb{R}^d$, such that the elements $\theta_i \in \Theta$ are realizations of the probability distribution P over the domain $\mathbb{R}^d$, that is, $\theta_i \sim P$, and given an embedding space $\mathcal{Y} \subset \mathbb{R}^{d'}$, define the mappings ϕ and e so that

$$\Theta \xrightarrow{\phi} \mathcal{M}_{\mathcal{H}} \xrightarrow{e} \mathcal{Y},$$

with $\mathcal{M}_{\mathcal{H}} \subset \mathcal{H}$ a manifold in the observation space. The structure of $\mathcal{M}_{\mathcal{H}}$ is determined by the structure and dimensionality of Θ, P, and the map ϕ, which is isometric for the appropriate metrics in Θ and $\mathcal{H}$. Conceptually, the low-dimensional parameter space Θ represents the modes of variation of the data and the mapping ϕ represents the data generating process that yields high-dimensional data $x_i = \phi(\theta_i) \in \mathcal{M}_{\mathcal{H}}$ characterized by these modes of variation. We assume that low-dimensional representations of the observed data in the embedding space $\mathcal{Y}$, which capture as much of the metric structure of $\mathcal{M}_{\mathcal{H}}$ as possible, can be learned from the observed data. A successful embedding e then also recovers as much of the structure of the parameter space Θ as possible in the low-dimensional representations $y_i = e(x_i) \in \mathcal{Y}$.

In our framework, distributional outliers are defined w.r.t. minimum volume sets (Polonik, 1997) of P in this parameter space Θ:

Definition 1. *Minimum volume set.*

Given a probability distribution P over (a subset of) $\mathbb{R}^d$, a minimum volume set Ω^*_α is a set that minimizes the quantile function $V(\alpha) = \inf_{C \in \mathcal{C}}\{\text{Leb}(C) : P(C) \geq \alpha\}, 0 < \alpha < 1$ for i.i.d. random variables in $\mathbb{R}^d$ with distribution P, $\mathcal{C}$ a class of measurable subsets in $\mathbb{R}^d$ and Lebesgue measure Leb.

So $\Omega^*_{\alpha,P}$ is the smallest region containing a probability mass of at least α. We can now define structural outliers and distributional outliers as follows:

Definition 2. *Structural and distributional outlier.*

Define $\mathcal{M}_{\Theta,\phi}$ as the codomain of ϕ applied to Θ.

Define two such manifolds $\mathcal{M}_a = \mathcal{M}_{\Theta_a,\phi_a}$ and $\mathcal{M}_c = \mathcal{M}_{\Theta_c,\phi_c}$ and a data set $X \subset \mathcal{M}_a \cup \mathcal{M}_c$.

W.l.o.g., let $r = \frac{|\{x_i : x_i \in \mathcal{M}_a \wedge x_i \notin \mathcal{M}_c\}|}{|\{x_i : x_i \in \mathcal{M}_c\}|} \lll 1$ be the *structural outlier ratio*, that is, most observations are assumed to stem from $\mathcal{M}_c$. Then an observation $x_i \in X$ is

- a *structural outlier* if $x_i \in \mathcal{M}_a$ and $x_i \notin \mathcal{M}_c$ and
- a *distributional outlier* if $x_i \in \mathcal{M}_c$ and $\theta_i \notin \Omega^*_\alpha$, where Ω^*_α is defined by the density of the distribution generating Θ_a.

Figure 1 shows examples of three data types with structural outliers (in red) and some distributional outliers for the functional data example. Since distributional outliers are structurally similar to inliers, they are hard to detect visually

for graph and image data, as doing so requires a lot of "normal" data to reference against and we can only display a few example observations here. As outlined, this is one reason why we again use functional data for our exposition in the following.

Summarizing the framework's crucial aspects in less technical terms, we assume that the bulk of the observations comes from a single "common" process, which generates observations in some subset $\mathcal{M}_c$, while some data might come from an "anomalous" process, which defines structurally distinct observations in a different subset $\mathcal{M}_a$. This follows standard notions in outlier detection which often assume (at least) two different data-generating processes (Dai et al., 2020; Zimek & Filzmoser, 2018). Note that this does not imply that structural outliers are in any way similar to each other: P_a could be very widely dispersed or arise from a mixture or several different distributions and/or $\mathcal{M}_a$ could consist of several unconnected components representing various kinds of structural abnormality. The only crucial aspect is that the process from which *most* of the observations emerge yields structurally similar data. We consider settings with a structural outlier ratio $r \in [0,0.1]$ to be suitable for outlier detection. The proportion of *distributional* outliers on $\mathcal{M}_c$, in contrast, depends only on the α-level for Ω^*_{α,P_c}. Practically speaking, neither prior knowledge about these manifolds nor specific assumptions about structural differences are necessary for our approach. The key points are that (1) structural outliers are not on the main data manifold $\mathcal{M}_c$, (2) distributional outliers are at the edges of $\mathcal{M}_c$, and (3) these properties are preserved in the embedding vectors as long as the embedding is based on an appropriate notion of distance in $\mathcal{H}$.

4 | EXPERIMENTS

This section lays out practical implications of the framework through experiments on several different data types, via a comprehensive qualitative and visual analysis of six examples. In addition, we provide quantitative results for six labeled data sets.

4.1 | Methods

The focus of our experiments is to evaluate a general framework for outlier detection, which is motivated by geometrical considerations. With these experiments, we support the claim that the perspective induced by the framework lets us visualize, detect, and analyze outliers in a principled and canonical way. For this demonstration, we chose Multidimensional Scaling (MDS) (Cox & Cox, 2008) as our embedding method and Local Outlier Factors (LOF) (Breunig et al., 2000) as our outlier scoring method. Note that the experiments are not intended to draw conclusions about the superiority of these specific methods and other combinations of methods may be as suitable or even superior for the purpose (see for example results for Isomap in Herrmann and Scheipl (2021)).

However, more sophisticated embedding methods than MDS require tuning over multiple hyperparameters, whereas MDS has only one—the embedding dimension. Moreover, an advantage of MDS over other embedding methods is that it aims for isometric embeddings, that is, tries to preserve all pairwise distances as closely as possible, which is crucial in particular to reflect structural outlyingness. In fact, Torgerson Multidimensional Scaling (tMDS, i.e., MDS based on L_2 distance)—that is: a simple linear embedding equivalent to standard PCA scores—seems to uncover many outlier structures sufficiently well in many data settings despite its simplicity. For similar reasons, we chose to use LOF as an outlier scoring method. This method also has a single hyperparameter, *minPts*, the number of nearest neighbors used to define a point's (local) neighborhood, which we denote as k in the following. Moreover, in contrast to many other outlier scoring methods such as one-class support vector machines (Muñoz & Moguerza, 2004) which require low-dimensional tabular data as input (i.e., which can only be applied to complex data types indirectly by using embedding vectors as feature inputs), LOF can also be applied to high-dimensional and non-tabular data directly as it only requires a distance matrix as input. Experiments on functional data have shown that LOF applied directly to a distance matrix of functional data and LOF applied to the corresponding embedding vectors yield consistent results (Herrmann & Scheipl, 2021).

Note, however, that beyond the ability to apply outlier scoring methods to low-dimensional embedding vectors of high-dimensional and/or non-tabular data, such embeddings provide additional practical value: In particular, scalar scores or ranks as provided by outlier scoring methods are not able to reflect differences between distributional and structural outliers whereas such differences become accessible and interpretable in visualizations of these embeddings.

This also points to a major caveat of the quantitative (in contrast to the qualitative) experiments, in which we use ROC-AUC to evaluate the accuracy of outlier ranks obtained with LOF with respect to the "outlier" structure defined by the different classes of labeled data. Setting one class as $\mathcal{M}_c$ and contaminating this "normal" class with observations from other classes, which are assumed to be structurally different and thus form $\mathcal{M}_a$, we obtain data sets $X \subset \mathcal{M}_c \cup \mathcal{M}_a$. Although this is a widely used approach (Campos et al., 2016; Goldstein & Uchida, 2016; Pang et al., 2018), such an evaluation only considers outliers as defined by the class labels and poor AUC values may not necessarily imply poor performance if there are observations from the "normal" data class which are (distributionally or structurally) more outlying (and thus obtain higher scores) than some of the "labeled" outliers, see also Campos et al. (2016). This is why we do not merely report potentially problematic quantitative measures (Section 4.3), and instead put more emphasis on qualitative experiments that are much closer to the way we would recommend using these methods in practical applications.

4.2 | Qualitative assessment

In this section, we provide extensive qualitative analyses to demonstrate the practical relevance of the framework. First, we demonstrate that the distinction between structural and distributional outliers is preserved in embeddings using two simulated functional data sets. Second, using two real-world data sets—a functional and an image data set—we show that the approach can be applied flexibly to different data structures. Third, we illustrate the general applicability to more general data types based on synthetic graph data and real-world curves data. In the following, all LOF results are obtained using $k = 0.75n$, where n is the number of observations.

4.2.1 | Demonstrating the framework's practical implications on idealized synthetic data

Figure 2 shows two simulated functional data sets (A & B, left) and their 2**D** PCA/tMDS embeddings (A & B, right). One can observe that data set A is an example with structural outliers in terms of shape and slope. This is an extended version of an example by Hernández and Muñoz (2016) and, following their notation, the two manifolds can be defined as $\mathcal{M}_a = \{x(t) | x(t) = b + 0.05t + \cos(20\pi t), b \in \mathbb{R}\} \cup \{x(t) | x(t) = (c - 0.05t) + e_t, c, e_t \in \mathbb{R}\}$ and $\mathcal{M}_c = \{x(t) | x(t) = a + 0.01t + \sin(\pi t^2), a \in \mathbb{R}\}$ with $t \in [0,1]$ and $a \sim N(\mu = 15, \sigma = 4)$, $b \sim N(\mu = 5, \sigma = 3)$, $c \sim N(\mu = 25, \sigma = 3)$, and $e_t \sim N(\mu = 0, \sigma = 4)$. Note that the structural outliers are not all similar to each other in shape or slope, which is reflected in $\mathcal{M}_a$ being a union of two structurally different manifolds.

In contrast, data set B of various (vertically shifted) Beta-distribution densities is an example where distributional outlyingness is defined by phase—that is horizontal—variation and structural outlyingness by vertical shifts. The respective manifolds are defined as $\mathcal{M}_a = \{x(t) | x(t) = b + B(t, \alpha, \beta), (b, \alpha, \beta)' \in \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+\}$ and $\mathcal{M}_c = \{x(t) | x(t) = B(t, \alpha, \beta), (\alpha, \beta)' \in \mathbb{R}^+ \times \mathbb{R}^+\}$ with $t \in [0,1]$, $\alpha, \beta \sim U[0.1, 2]$, $b \sim U[-5, 5]$ and B the density of the beta distribution. For both, we generate 100 "normal" observations from $\mathcal{M}_c$ and 10 structural outliers from $\mathcal{M}_a$, with $D = 500$ evaluation points in the first and $D = 50$ evaluation points in the latter example.

Structural outliers are clearly separated from observations on $\mathcal{M}_c$ in both cases and appear as outlying in the 2**D** embeddings. Moreover, we see that distributional outliers are embedded at the periphery of $\mathcal{M}_c$. Numbers in the figures are ascending LOF score ranks of the outliers. Note that $\mathcal{M}_c \subset \mathcal{M}_a$ in data set B. Nevertheless, most structurally outlying observations from $\mathcal{M}_a$ are clearly separated in the embedding. Two structural outliers are in or very close to $\mathcal{M}_c \cup \mathcal{M}_a$ and thus appear in the main bulk of the data. The LOF scores also reflect this, as one of the distributional outliers is ranked as even more outlying.

Summarizing, we see that in these simulated situations, practically relevant outlier sub-structure—deviations in terms of functional shape, slope, or vertical shifts—are represented accurately by low-dimensional embeddings learned from the observed high-dimensional data. In particular, structural outliers do not need to be similar to each other as Example A demonstrates. Also, note that Example B illustrates as a by-product that there can be situations where the approach yields meaningful results even though the two manifolds are not completely disjoint. However, this does not necessarily hold in general. See Souvenir and Pless (2005) for an approach to disentangle intersecting manifolds. Moreover, we see that situations where distributional outliers appear "more" outlying than structural outliers are captured as well. Note that this is a crucial aspect. Although this aspect is quantified correctly by an outlier scoring method such as LOF, the two outlier types can be distinguished only if visualizations, as provided by embedding methods, are

considered. Consider that evaluation of unsupervised outlier detection is often performed using a labeled data set, setting observations from one class as inliers and sampling observations from another class as outliers, and then computing binary classification performance measures such as the AUC (Campos et al., 2016; Goldstein & Uchida, 2016; Pang et al., 2018). Different class labels do not guarantee that the classes do not overlap, that is, that the respective manifolds are disjoint in $\mathcal{H}$, nor that there are no distributional outliers appearing more outlying than structural outliers. Thus, there may be distributional outliers among the inliers which are scored as more outlying than structural outliers (see data set B) and a purely quantitative assessment is likely to mislead. Being able to create faithful visualizations of such more complex outlier structures for high-dimensional data is a crucial benefit of the proposed approach.

4.2.2 | Demonstrating flexibility on real functional and image data

Of course, real-world data settings are usually more complicated than our simulated examples. First of all, real data are much more difficult to assess since the underlying manifolds are usually not directly accessible, so it is impossible to define the exact structure of the data manifolds like in the simulated examples. In addition, some data sets may not contain any clear *structural* outliers, while others may not contain any clear *distributional* outliers, or both. A crucial aspect of the approach is that, although it is based on a highly abstract conceptualization involving unobservables like the parameter space Θ and its probability measure P, it is not at all necessary to come up with any such formalization of the data generating process to put the approach into practice and obtain meaningful results, as will be demonstrated in the following.

Consider Figure 3, which shows a real functional data set of 591 ECG measurements (Dau et al., 2019; Goldberger et al., 2000) with 82 evaluation points per function, that is, a $D = 82$ dimensional data set (a), and a sample of the COIL20 data (Nane et al., 1996) (b). It is impossible to define the exact structure of the ECG data manifold. However, the visualizations of the functions on the left-hand side suggest that there are no observations with clear structural differences in functional form: none of the curves are clearly shifted away from the bulk of the data, nor are there any curves with isolated peaks, or observations with clearly different shapes. In accordance with this observation, there is also no clearly separable structure in the embedding. However, observations that appear in low-density regions of the embedding can be regarded as distributional outliers in terms of horizontal shift, that is, phase variation, like the three observations with the earliest minima colored in blue. This is also reflected in the scoring of the embeddings, as the observations with the lowest LOF ranks are clear distributional outliers in function space. However, the embedding provides much more complete information in this example than LOF ranks and the functional visualization alone. For example, they also pinpoint a *vertical shift* outlier in the first and last thirds of the domain (green curve, which would be hard to detect based on its functional representation alone). This apparently represents a second "dimension" of distributional outlyingness.

The COIL20 data (Nane et al., 1996) consists of 1440 pictures (128×128, grayscale) of 20 different objects. The 72 pictures in each class depict one and the same object at different rotation angles with a picture taken at every 5° within $[0°, 355°]$. We use all 72 pictures of a rubber duck to represent observations from $\mathcal{M}_c$ and randomly sample 7 observations (i.e., $r \approx 0.1$) from the 72 pictures of a toy car as structural outliers from $\mathcal{M}_a$. We compute L_2 distances of the vectorized pixel intensities ($D = 128^2 = 16384$). Figure 3b, left column, shows a sample of 6 inlier and 3 structural outlier pictures, the right column shows embeddings of all 79 images. Since the inlier data are images of a rotated object, $\mathcal{M}_c$ is the image of a one-dimensional closed and circular parameter space defining the rotation angle (c.f. Ma & Fu, 2011), that is, other than in the ECG example substantial considerations yield at least some knowledge about the specific structure of the data manifold(s) in this case.

The 2**D** embedding reflects the expected structure of our COIL20 subset very well, with clear separation of the 7 pictures of the toy car as structural outliers. In addition, the embedding of $\mathcal{M}_c$ indeed yields a closed, but not quite a circular loop, as does the embedding of the seven rotated images from $\mathcal{M}_a$. The corresponding 3D embedding (not shown) reveals that the embeddings of the inliers lie on a rough circle folded over itself. In summary, in the ECG example there seem to be no clearly separable, structurally different outliers that could be detected with tMDS, but only distributional outliers, whereas in the COIL data there are clearly separate structural outliers, but no distributionally outlying observations. These two examples with very different intrinsic structures (single connected manifold with distributional outliers versus disconnected manifolds without clear distributional outliers) illustrate that it is not necessary to have explicit prior knowledge about the data generating process or its outlier characteristics for the approach to work and that it is able to handle different data manifold structures flexibly and successfully.

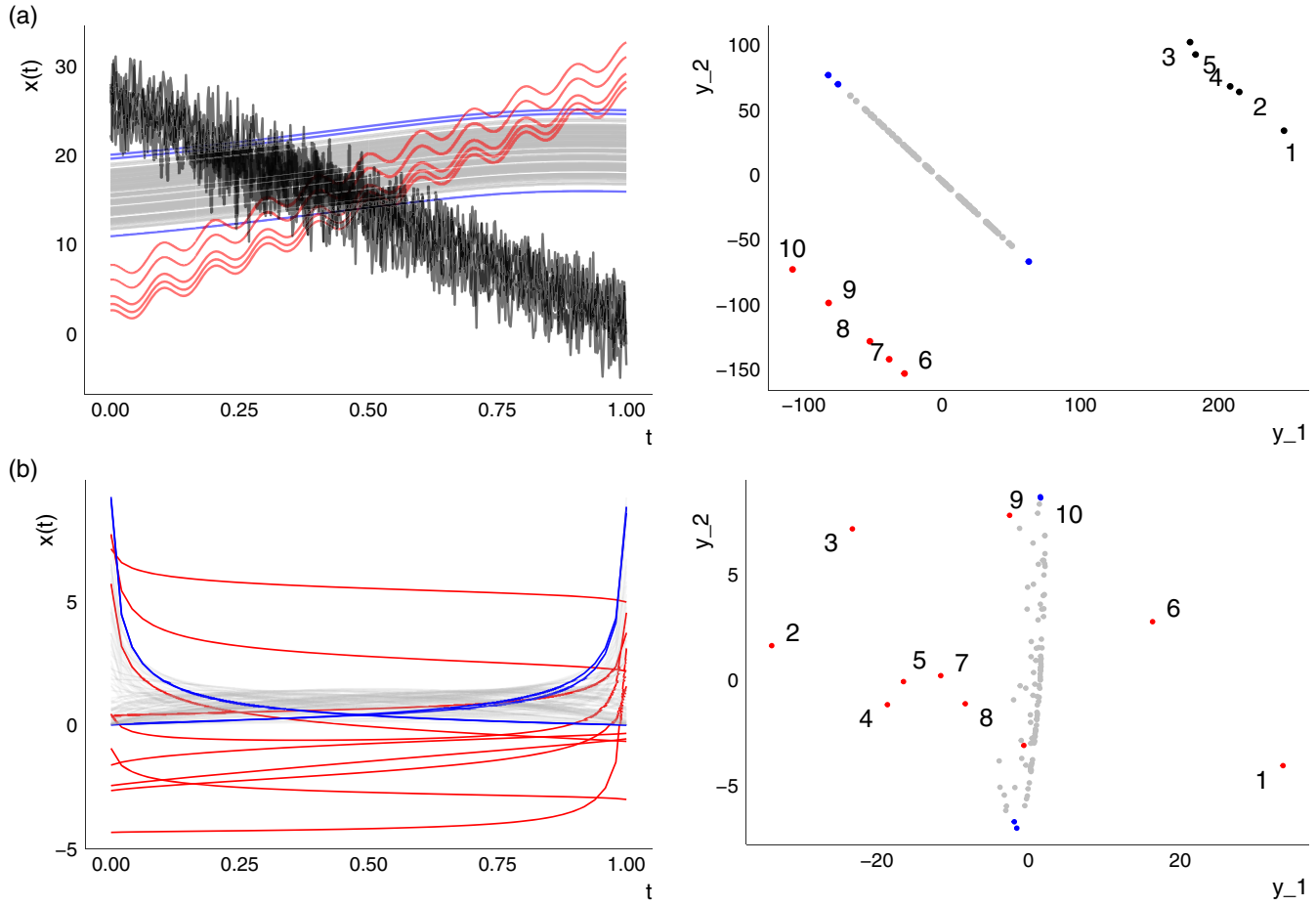

FIGURE 2 Simulated functional data and their 2**D** embeddings. Numbered labels are ascending LOF score ranks of the outliers ($k = 0.75n$).

4.2.3 | Demonstrating generalizability on graph and curve data

Note that the COIL example illustrates that the framework also works in image data and that a fairly simplistic approach of computing L_2 distances between vectorized pixel intensities yields very reasonable results in this example. The framework is, however, not at all restricted to these two data types nor such a simple distance metric. Recall that the approach can be applied to any data type whatsoever as long as a suitable distance metric is available. Beyond 1**D** functional and image data, the framework can also be extended to more general and complex data types, for example, graphs or 2**D** curves as depicted in Figure 4. We use more specialized distance measures to show that good results can also be obtained on such data.

We simulate two structurally different classes of Erdös-Rényi graphs with 20 vertices (see Figure 4a). This structural difference results from different edge probabilities p_v that two given vertices of the graph are connected, setting $p_v = 0.1$ for $\mathcal{M}_c$ and $p_v = 0.4$ for $\mathcal{M}_a$. We randomly sample 100 observations from $\mathcal{M}_c$ and 10 from $\mathcal{M}_a$, that is, $r = 0.1$, and obtain a pairwise distance matrix by computing the Frobenius distances between the graph Laplacians.

The curves data (Figure 4b) consists of spiral curve drawings from an Archimedes spiral-drawing test that is used to diagnose patients with Parkinson's disease (Alty et al., 2017; Steyer et al., 2022). Taking data from the dynamic version of the test (Isenkul et al., 2014), we use 15 curves drawn by healthy controls not suffering from Parkinson's disease and two curves drawn by Parkinson patients to represent potential structural outliers, where each curve is evaluated on 200 points. Previous investigations have shown that an elastic shape distance is better suited than L_2 distances to discriminate between the two groups (Steyer et al., 2022).

So, in contrast to the previous examples, we use more specialized distance measures to capture the relevant structures in these settings. This illustrates that the approach is not only flexible with respect to the actual structure present

in a given data set as demonstrated in the previous section but that it is also very generally applicable to a variety of data types. The approach can be used for any kind of data simply by defining an appropriate (data-specific) distance measure. In both, the embeddings of the graphs, as well as the embeddings of the curves, structurally different observations (in red) are clearly separated from the observations on $\mathcal{M}_c$. This is also reflected by their LOF scores. Moreover, in both settings, there are observations from $\mathcal{M}_c$ (in blue) which appear in peripheral, sparser regions of the "normal" data and thus can be considered distributional outliers. Note that it is not always immediately obvious on the level of the original data why observations appear distributionally outlying. For example, in the graph data, note that other than in previous examples (e.g., Figure 2a) comparing them to a few inliers does not reveal a striking difference at first (in contrast to the structural outliers!): Figure 4a, left column, shows six inlier graphs in the 1st and 2nd row, the three distributional outlier graphs in the 3rd row, and three structural outlier graphs in the 4th row.

Nevertheless, the embedding vectors and their LOF ranks indicate that the distributionally outlying observations have obtained some specific characteristics setting them apart from most inlying observations. For example, further analysis reveals that the graph with LOF rank 11 contains the node with maximum connectedness of all nodes in all inlier graphs. Its degree is 8 (i.e., it is directly connected to 8 other nodes), while the average of the maximum degree in the graphs on $\mathcal{M}_c$ is just 4.39. In contrast, the graph with LOF rank 13 contains 8 isolated nodes of degree 0, while the average number of nodes with degree 0 is only 2.47 on $\mathcal{M}_c$. The respective values of the graph with LOF rank 10 are above the upper quartile for both of these metrics, with four unconnected nodes and a maximally connected node with degree 6.

4.3 | Quantitative assessment

In order to provide less subjective experimental results, we assess the approach quantitatively, using labeled data with at least two classes. For each data set, we consider four outlier ratios $r \in \{0.01, 0.025, 0.5, 0.1\}$. Setting one class as $\mathcal{M}_c$, with $n_{\text{in}} = |\mathcal{M}_c|$, and contaminating this "normal" class with $n_{\text{out}} = r \cdot n_{\text{in}}$ "structural" outliers from other classes, which form $\mathcal{M}_a$, we obtain data sets $X \subset \mathcal{M}_c \cup \mathcal{M}_a$ with $n = n_{\text{in}} + n_{\text{out}}$. For each setting, we repeat the contamination process 50 times, sampling outliers at random from $\mathcal{M}_a$. Based on outlier ranks computed with LOF, we use ROC-AUC as a performance measure and report the mean AUCs over the 50 replications for each combination of settings. Note that we only use the labels of the "structural" outliers for computing this performance measure, not for the unsupervised learning of the embeddings themselves. For all data sets considered in this section, plots of typical embeddings for $r = 0.05$ can be found in Figures A1 and A2. We consider three additional functional data sets for this experiment:

TABLE 1 Mean ROC-AUC values over 50 replications based on the ranks as assigned by LOF

A		**Dodgers**				**Phoneme**				**Starlight**			
		$n_{in} = 97, D = 289$				$n_{in} = 400, D = 150$				$n_{in} = 6656, D = 1025$			
	k	$0.01n$	$0.1n$	$0.75n$	$0.9n$	$0.01n$	$0.1n$	$0.75n$	$0.9n$	$0.01n$	$0.1n$	$0.75n$	$0.9n$
r : 1.0%		0.78	0.98	0.96	0.96	0.78	1.00	0.99	0.99	0.96	1.00	0.69	0.78
r : 2.5%		0.62	0.97	0.96	0.96	0.54	1.00	0.99	0.99	0.55	1.00	0.88	0.88
r : 5.0%		0.59	0.97	0.96	0.96	0.56	0.99	0.99	0.99	0.53	1.00	0.92	0.92
r : 10%		0.54	0.84	0.97	0.96	0.57	0.75	0.99	0.99	0.56	0.98	0.95	0.87
B		Iris				Wisconsin breast cancer				Simulated data			
		$n_{in} = 50, D = 4$				$n_{in} = 357, D = 30$				$n_{in} = 750, D = 1000$			
	k	$0.01n$	$0.1n$	$0.75n$	$0.9n$	$0.01n$	$0.1n$	$0.75n$	$0.9n$	$0.01n$	$0.1n$	$0.75n$	$0.9n$
r : 1.0%		1.00	1.00	1.00	1.00	0.76	0.96	0.94	0.89	0.66	1.00	1.00	1.00
r : 2.5%		0.71	1.00	1.00	1.00	0.64	0.97	0.95	0.92	0.56	1.00	1.00	1.00
r : 5.0%		0.52	1.00	1.00	1.00	0.60	0.97	0.94	0.91	0.58	1.00	1.00	1.00
r : 10%		0.61	0.69	1.00	1.00	0.58	0.96	0.94	0.92	0.56	1.00	1.00	1.00

Note: Each data set consists of n observations, n_{in} from $\mathcal{M}_c$ and $n_{\text{out}} = n_{\text{in}} \cdot r$ from $\mathcal{M}_a$. $\mathcal{M}_a$ and $\mathcal{M}_c$ are defined by classes of the original labeled data sets. D is the dimensionality of a data set (i.e., evaluations per function for functional data) and k the number of nearest neighbors used in the LOF algorithm. (A) Functional data. (B) Tabular data.

dodgers (Dau et al., 2019), a set of times series of daily traffic close to Dodgers Stadium, with days on weekends forming $\mathcal{M}_a$ and weekdays forming $\mathcal{M}_c$; *phoneme* (Febrero-Bande & Oviedo de la Fuente, 2012), discretized log-periodograms of five different phonemes, with phoneme "dcl" forming $\mathcal{M}_a$ and phonemes "sh", "iy", "aa", and "ao" forming $\mathcal{M}_c$; *starlight* (Dau et al., 2019; Rebbapragada et al., 2009), phase-aligned light curves of Eclipsing Binary, Cepheid, and RR Lyrae stars, the first forming $\mathcal{M}_a$ and the latter two forming $\mathcal{M}_c$. All results are based on simple, linear tMDS/PCA embeddings with the LOF algorithm applied to the resulting 2**D** embedding vectors.

In addition, we consider three tabular data sets, two real and one simulated. This includes the well-known Iris data (Anderson, 1935; Fisher, 1936) where class *Setosa* forms $\mathcal{M}_c$ and the other two classes $\mathcal{M}_a$. Moreover, we use the Wiscon Breast Cancer (wbc) data (Street et al., 1993) as provided by the UCI Machine Learning repository (Dua & Graff, 2017). This tabular data set comprises 30 features containing information about the cell nuclei of breast tissue and has been used by Goldstein and Uchida (2016) for outlier detection before. Following their approach, the healthy patients form $\mathcal{M}_c$ and patients with malignant status form $\mathcal{M}_a$. Yet, other than Goldstein and Uchida (2016), we do not fix outliers to the first 10 observations from the latter class but—as outlined—repeatedly sample outliers at random from $\mathcal{M}_a$. Finally, we include a simple simulated example where $\mathcal{M}_c = \{x : x \sim \mathcal{N}_{1000}(\mathbf{0}, \Sigma)\}$ and $\mathcal{M}_a = \{x : x \sim \mathcal{N}_{1000}(\mathbf{1}, \Sigma)\}$, $\Sigma = \text{diag}(\mathbf{1})$. That is, 1000-dimensional data with observations sampled from two multivariate normal distributions where the class difference stems from the difference in the mean vectors, $\mathbf{0}$ for $\mathcal{M}_c$ and $\mathbf{1}$ for $\mathcal{M}_a$.

The results depicted in Table 1 show that outlier detection does not need to be specifically challenging in nominally high-dimensional data. In each of the data sets, which have very different numbers of observations and numbers of dimensions, high ROC-AUC ≥ 0.95 can be achieved for all considered outlier ratios r. This indicates that most of the observations from $\mathcal{M}_a$ indeed appear to be outlying in the embedding space and thus obtain high LOF scores. Furthermore, as in the qualitative analysis, a global setting of $k = 0.75n$ seems to be a reasonable default for the LOF algorithm. Only for $r = 0.01, 0.025$ in the starlight data, we see a large improvement (AUC $= 1.00$) with $k = 0.1n$. For small $r < 0.1$, in all other settings the achieved ROC-AUC is very robust against changes in this tuning parameter.

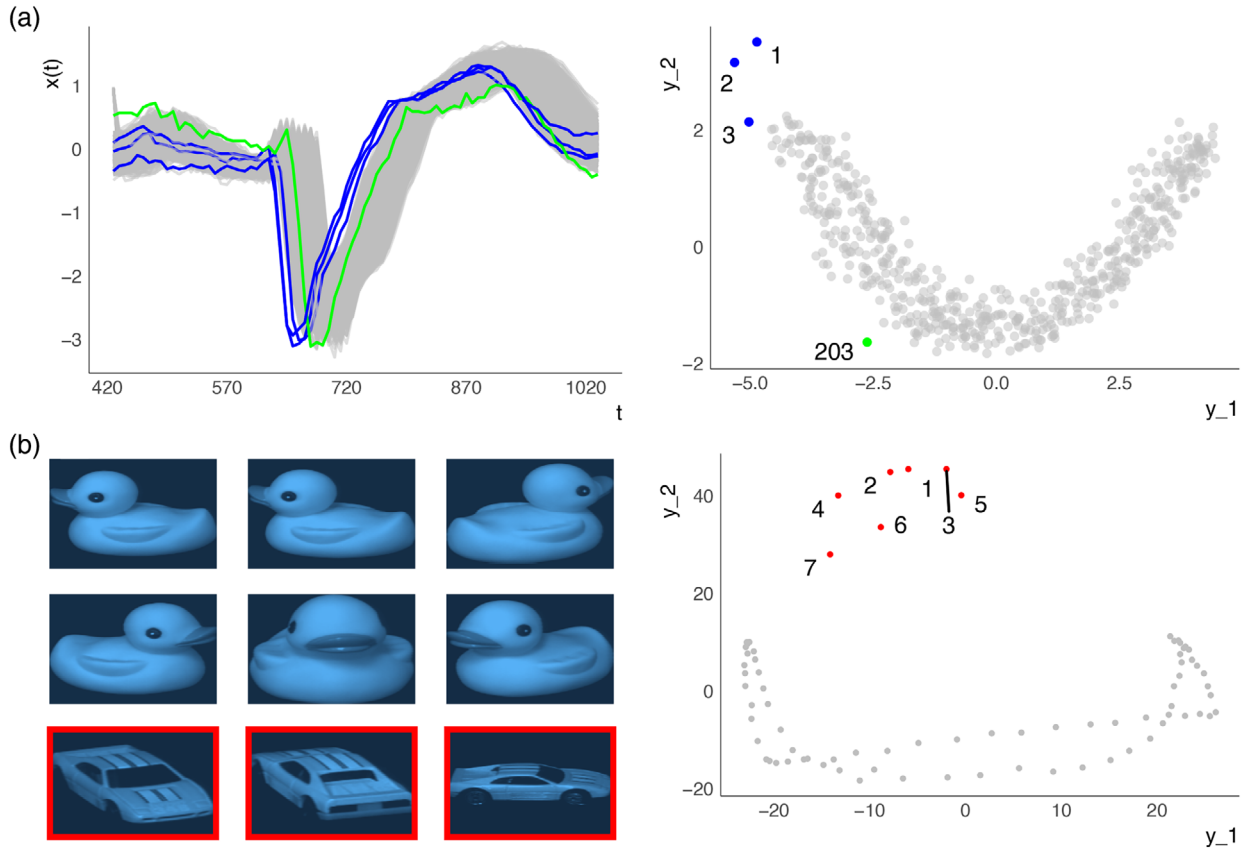

FIGURE 3 Real functional and image data and their 2**D** tMDS embeddings. Numbered labels are ascending LOF score ranks of the outliers ($k = 0.75n$).

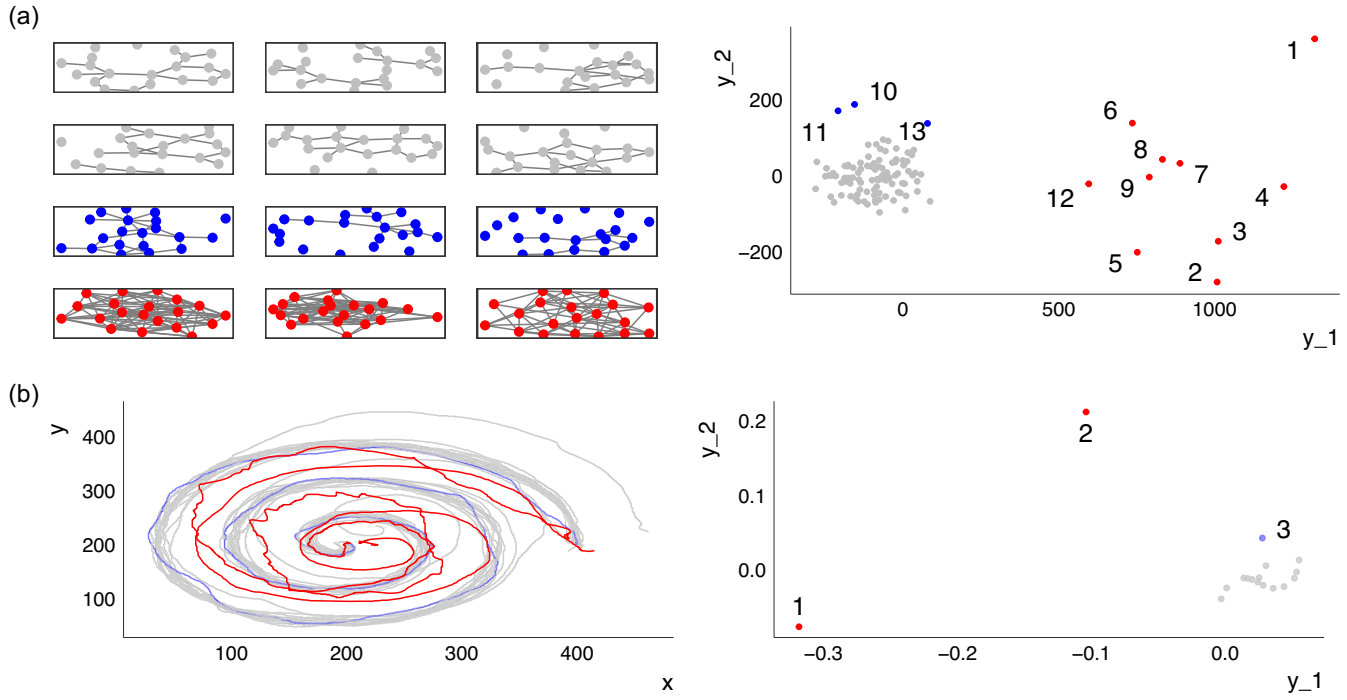

FIGURE 4 Curve and graph data as further examples to demonstrate the flexibility and general applicability of the approach, and their 2**D** MDS embeddings based on Frobenius (graphs) and Elastic shape distances (curves). Numbered labels are ascending LOF score ranks of the outliers ($k = 0.75n$).

5 | DISCUSSION

5.1 | Summary

We propose a geometrically motivated framework for outlier detection, which exploits the metric structure of a (possibly high-dimensional) data set and provides a mathematically precise distinction between *distributional* outliers and *structural* outliers. Experiments show that the outlier structure of high-dimensional and non-tabular data can be detected, visualized, and quantified using established manifold learning methods and standard outlier scoring. The decisive advantage of our framework from a theoretical perspective is that the resulting embeddings make subtle but important properties of outlier structure explicit and—even more importantly—that these properties are made accessible based on visualizations of the embeddings. From a more practical perspective, our proposal requires no prior knowledge nor any specific assumptions about the actual data structure in order to work, an important aspect since data generating processes are usually inaccessible. This is highly relevant in practice, in particular since a well-established, computationally cheap combination of widely used and fairly simple methods like (t)MDS and LOF proved to be a strong baseline that yields fairly reliable results without the need for tuning hyperparameters. In addition, the proposed framework has several more general conceptual implications for outlier detection which will be summarized in the following.

5.2 | Implications

5.2.1 | Outlier taxonomy

We propose a clear taxonomy to distinguish between frequently interchangeably used terms *anomalies* and *outliers* in a canonical way: we regard *anomalies* as observations from a different data generating process than the majority of the data (i.e., as observations that are on $\mathcal{M}_a$ but not on $\mathcal{M}_c$), which can be more precisely identified as *structural* outliers. Recall that Zimek and Filzmoser (2018, p. 10) refer to such observations as "real" outliers that need to be distinguished

from "observations which are in the extremes of the model distribution". On the other hand, regarding *outliers* as observations from low-density regions of the underlying "normal" data manifold $\mathcal{M}_c$, they can be more precisely identified as *distributional* outliers. Based on our reading of the literature, this distinction is usually not made explicit. Since there is rarely a practical reason to assume that a given data set contains only *distributional* or only *structural* outliers, some of the confusion surrounding the topic (Goldstein & Uchida, 2016; Unwin, 2019; Zimek & Filzmoser, 2018) might be because such conceptual differences have not been made sufficiently clear. As outlined, the concept of structural difference is very general. For example, structural differences in functional data may appear as shape anomalies in data mainly characterized by vertical shift variation (see Figure 1a) or as vertical shift anomalies in data dominated by shape variation, as phase anomalies in data with magnitude variation or magnitude anomalies in data with phase variation, etc.

In real unlabeled data, there may not always be a clear distinction between somewhat structurally anomalous observations with "off-manifold" embeddings and merely distributionally outlying observations with embeddings on the periphery of the data manifold, as in the ECG data in Figure 3a. Nevertheless, the theoretical distinction between these two kinds of outliers adds conceptual clarity even if the practical application of the categories may not be straightforward.

5.2.2 | Curse of dimensionality

As outlined in Section 2.1, outlier detection is often reported to suffer from the curse of dimensionality. For example, Goldstein and Uchida (2016) show that most outlier detection methods under consideration break down or perform poorly in a data set with 400 dimensions and conclude that unsupervised outlier detection is not possible in such high dimensions. Some (e.g., Aggarwal, 2017) attribute this to the fundamental problem that distance functions can lose their discriminating power in high dimensions (Beyer et al., 1999), which is linked to the concentration of measure effect (Pestov, 2000). However, this effect occurs only under fairly specific conditions (Zimek et al., 2012), which means that outlier detection does not have to be affected by the curse of dimensionality: In addition to the effects of dependency structures and signal-to-noise ratios (Zimek et al., 2012), the necessary conditions for concentration of measure are not fulfilled if the intrinsic dimensionality of the data is smaller than the actually observed dimensionality, or if the data is distributed in clusters that are relatively well separable (Beyer et al., 1999). Exactly these two characteristics are reflected in our framework in the form of (1) the manifold assumption, which implies low-ish intrinsic dimensionality and (2) the assumption that structural outliers come from different manifolds than the rest of the data, that is, from different "clusters" in $\mathcal{H}$. This has two important consequences: First of all, the geometric perspective our framework is based on makes these important aspects for outlier detection in high-dimensional data explicit, while a purely probabilistic perspective obscures them. Second, it mitigates many of the problems associated with high-dimensional outlier detection: any outlier detection method that performs well in low dimensions becomes—in principle—applicable in nominally high-dimensional and/or complex non-tabular data when applied to suitable low-dimensional embedding coordinates. In addition, our results show that outlier sub-structure, specifically the differences between distributional and structural outliers, can be detected and visualized with manifold methods. This opens new possibilities for descriptive and exploratory analyses:

5.2.3 | Visualizability of outlier characteristics

If the embeddings provided by manifold methods are restricted to two or three dimensions, they also provide easily accessible visualizations of the data. In fact, manifold learning is often used in applications specifically to find two- or three-dimensional visualizations reflecting the essential intrinsic structure of the high-dimensional data as faithfully as possible. Consequently, structural and distributional outliers, which are rather glaring data characteristics if the manifolds are well separable, can often be separated clearly even in two- or three-dimensional representations as long as the embedding is (approximately) isometric with respect to a suitable dissimilarity measure. This is specifically important for complex non-tabular or high-dimensional data types such as images or graphs, where at most a few observations can be visualized and perceived simultaneously. In the same vein, substructures and notions of data depth are reflected in the embeddings, making the approach also useful as an exploration tool for settings with unclear structure.

5.2.4 | Generalizability

Since the central building block of the proposed framework is to capture the metric structure of data sets using distance measures, the framework is very general and applicable to any data type for which distance metrics are available. In Section 4.2, we illustrated this generalizability using high-dimensional as well as non-tabular data; in particular, we applied it to functional, curve, graph, and image data. This also makes the framework very flexible as one can make use of non-standard and customized dissimilarity measures to emphasize the relevant structural differences in specific situations based on domain knowledge: Representing image data as vectors of pixel intensities, we computed distances between those vectors, for example. Dissimilarities between different graphs were captured, for example, by constructing their graph Laplacians and computing Frobenius distances between them, and we used a specific elastic depth distance for the spiral curve data as suggested by earlier results in Steyer et al. (2022).

5.3 | Limitations and outlook

If in an exploratory setting, observations appear clearly separated in the (first few) embedding dimensions, we can be sure they are structural outliers. Note that if D-dimensional data actually live in a d'-dimensional subspace, constructing a d'-dimensional embedding with MDS based on L_2 distances will lead to an embedding with a distance matrix exactly matching the distance matrix in the D-dimensional space, that is, MDS is isometric by design. If other than L_2 distances are used this still holds approximately (Young & Householder, 1938; see also Cox & Cox, 2008; Torgerson, 1952). Note that this is an important difference from many other dimension reduction methods. For example, UMAP is based on a local connectivity constraint (McInnes et al., 2020) which ensures that each point is at least connected to its nearest neighbor and which runs counter to a reliable embedding of structural outliers. In addition, more sophisticated methods require parameter tuning for any given setting, which is inherently difficult for unsupervised tasks, and it is not always clear how to tune other embedding methods so that they yield (approximately) isometric embeddings.

Clear structural outliers are the source of large variation in data sets with low intrinsic dimensionality. Since MDS embedding dimensions are sorted according to the decreasing variation, they will be reflected in the first few embedding dimensions. It may be that some of the distributional outliers are masked due to projection if the embedding dimension d is smaller than d' but following Zimek and Filzmoser (2018) we consider faithfully reflecting structural outliers more important. However, inliers, that is, observations on $\mathcal{M}_c$, may show large "within class" variation and/or may be spread over several disconnected clusters in some situations. For example, object images on $\mathcal{M}_c$, which are structurally similar in terms of the depicted objects' shape, may vary in rotation, scale, or location, and may have different colors or textures. In functional data, observations on $\mathcal{M}_c$ may show phase and amplitude variation and form clusters due to different shapes. In such settings, $\mathcal{M}_c$ can yield complex substructure and highly dispersed observations and it may be hard to distinguish whether separable structures observed in embeddings are due to groups of homogeneous structural outliers or due to multimodality in $\mathcal{M}_c$ in which some modes are sparsely sampled. Moreover, in such cases, the dispersion of $\mathcal{M}_c$ accounts for large parts of the data's variability, and two- or three-dimensional MDS embeddings may not be sufficient to also faithfully represent structural outliers, since MDS embedding vectors are sorted decreasingly by explained "variance". However, this does not mean that structural outliers are not necessarily separable. Instead, they appear as outliers in higher embedding dimensions, requiring higher order embeddings to reflect the outlier structure. That means, if in an exploratory setting, there are no clearly separated observations in the (first few) embedding dimensions, there are either no clear structural outliers or they appear in later embedding dimensions if there are sources in $\mathcal{M}_c$ that induce more variation than the structural outliers. For example, objects in images may be structurally different in texture but not in color, orientation, and scale. In such a case—all observations differ in color, orientation, and scale but only some observations in texture—these other aspects can induce large variation within observations on $\mathcal{M}_c$, and the structural difference in texture is loaded on latter embedding dimensions. In such a situation, one can use scatterplot matrices and *Scagnostics* (scatterplot diagnostics, Wilkinson et al., 2005) for visual inspection. In addition, one can check out the kurtosis of the LOF scores in different embedding dimensions or *high contrast subspaces for density-based outlier ranking* (HiCS, Keller et al., 2012), to find pairs of dimensions that are "interesting" in terms of structural outliers. Moreover, techniques from multi-view learning such as "distance-learning from multiple views" may likely yield better results, because different structures (e.g., structure induced by color vs. structure induced by texture) should be "treated separately as they are *semantically* different" (Zimek & Vreeken, 2015, p. 128).

Note, however, that suitable inductive biases can also be brought to bear in our framework fairly easily. If substantial considerations suggest that specific structural aspects are important, specifying dissimilarity metrics focused on these aspects allows to emphasize the relevant differences. For example, if isolated outliers in functional data (i.e., functions which yield outlying behavior only over small parts of the domain such as isolated peaks) are of most interest, higher order L_p metrics such as L_{10} will be much more sensitive to such structural differences than general L_2 distances. If phase variation should be ignored, the unnormalized L_1-Wasserstein or the Dynamic Time Warping (DTW) distance can be used. Such problem-specific distance measures can reduce the number of MDS embedding dimensions necessary for faithful embeddings of structural outliers (Herrmann & Scheipl, 2021). In future work, we will investigate these aspects and possible extensions w.r.t. to multi-view learning approaches. Moreover, we will elaborate more on the specifics of other data types, in particular, image data.

6 | CONCLUSION

In conclusion, our illustration suggests that the proposed geometric conceptualization, which distinguishes *distributional* and *structural* outliers on a general level, provides a more precise terminology and shows that outlier detection in high-dimensional and complex non-tabular data does need to be specifically challenging per se. Convincing results could be achieved in a wide range of settings and data types by a combination of the simple methods MDS for dimension reduction and visualization and LOF for outlier scoring. We hope that the proposed framework contributes to a better understanding of unsupervised outlier detection and provides some guidance to practitioners as well as methodological researchers in this regard.

AUTHOR CONTRIBUTIONS

Moritz Herrmann: Conceptualization (lead); data curation (lead); formal analysis (lead); methodology (lead); visualization (lead); writing – original draft (lead); writing – review and editing (equal). **Florian Pfisterer:** Conceptualization (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (equal). **Fabian Scheipl:** Conceptualization (supporting); funding acquisition (lead); methodology (supporting); supervision (lead); writing – review and editing (equal).

ACKNOWLEDGMENT

The authors thank Almond Stöcker for his helpful advice regarding the spiral curve data. Open Access funding enabled and organized by Projekt DEAL.

FUNDING INFORMATION

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibility for its content.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [[insert provided URL from Open Research Disclosure Form]].

DATA AVAILABILITY STATEMENT

The data and code to reproduce the findings of this study are openly available on GitHub at: https://github.com/HerrMo/geo-outlier-framework

ORCID

Moritz Herrmann https://orcid.org/0000-0002-4893-5812

RELATED WIREs ARTICLES

There and back again: Outlier detection between statistical reasoning and data mining algorithms

REFERENCES

Aggarwal, C. C. (2017). *Outlier analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-47578-3

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. *SIGMOD Record*, *30*(2), 37–46. https://doi.org/10.1145/376284.375668

Ali, M., Jones, M. W., Xie, X., & Williams, M. (2019). TimeCluster: Dimension reduction applied to temporal data for visual analytics. *The Visual Computer*, *35*(6), 1013–1026. https://doi.org/10.1007/s00371-019-01673-y

Alty, J., Cosgrove, J., Thorpe, D., & Kempster, P. (2017). How to use pen and paper tasks to aid tremor diagnosis in the clinic. *Practical Neurology*, *17*(6), 456–463. https://doi.org/10.1136/practneurol-2017-001719

Anderson, E. (1935). The irises of the gaspé peninsula. *Bulletin of the American Iris Society*, *59*, 2–5.

Azcorra, A., Chiroque, L. F., Cuevas, R., Anta, A. F., Laniado, H., Lillo, R. E., Romo, J., & Sguera, C. (2018). Unsupervised scalable statistical method for identifying influential users in online social networks. *Scientific Reports*, *8*(1), 1–7. https://doi.org/10.1038/s41598-018-24874-2

Beckman, R. J., & Cook, R. D. (1983). Outlier ... ... s. *Technometrics*, *25*(2), 119–149. https://doi.org/10.1080/00401706.1983.10487840

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396. https://doi.org/10.1162/089976603321780317

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In C. Beeri & P. Buneman (Eds.), *Database theory: ICDT'99* (pp. 217–235). Springer. https://doi.org/10.1007/3-540-49257-7_15

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying density-based local outliers. *SIGMOD Record*, *29*(2), 93–104. https://doi.org/10.1145/335191.335388

Campos, G. O., Zimek, A., Sander, J., Campello, R. J., Micenková, B., Schubert, E., Assent, I., & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery*, *30*(4), 891–927. https://doi.org/10.1007/s10618-015-0444-8

Clémençon, S., & Jakubowicz, J. (2013). Scoring anomalies: A M-estimation formulation. In C. M. Carvalho & P. Ravikumar (Eds.), *Proceedings of the sixteenth international conference on artificial intelligence and statistics* (Vol. 31, pp. 659–667. PMLR). JMLR. https://proceedings.mlr.press/v31/clemencon13a.html

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5–30. https://doi.org/10.1016/j.acha.2006.04.006

Cox, M. A. A., & Cox, T. F. (2008). Multidimensional scaling. In C. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualization* (pp. 315–347). Springer. https://doi.org/10.1007/978-3-540-33037-0_14

Dai, W., Mrkvička, T., Sun, Y., & Genton, M. G. (2020). Functional outlier detection and taxonomy by sequential transformations. *Computational Statistics & Data Analysis*, *149*, 106960. https://doi.org/10.1016/j.csda.2020.106960

Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., & Keogh, E. (2019). The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, *6*(6), 1293–1305. https://doi.org/10.1109/JAS.2019.1911747

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. University of California, Irvine, School of Information. http://archive.ics.uci.edu/ml

Febrero-Bande, M., & Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, *51*(4), 1–28. https://doi.org/10.18637/jss.v051.i04

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann Eugen*, *7*(2), 179–188.

Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.-B., & Thirion, B. (2012). Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. *Medical Image Analysis*, *16*(7), 1359–1370. https://doi.org/10.1016/j.media.2012.05.002

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS One*, *11*(4), e0152173. https://doi.org/10.1371/journal.pone.0152173

Guan, S., & Loew, M. (2021). A novel intrinsic measure of data separability. arXiv:2109.05180 [Cs, Math, Stat]. http://arxiv.org/abs/2109.05180

Hernández, N., & Muñoz, A. (2016). Kernel depth measures for functional data with application to outlier detection. In A. E. P. Villa, P. Masulli, & A. J. Pons Rivero (Eds.), *Artificial neural networks and machine learning: ICANN 2016* (pp. 235–242). Springer. https://doi.org/10.1007/978-3-319-44781-0_28

Herrmann, M., & Scheipl, F. (2021). A geometric perspective on functional outlier detection. *Stats*, *4*(4), 971–1011. https://doi.org/10.3390/stats4040057

Isenkul, M., Sakar, B., Kursun, O., et al. (2014). Improved spiral test using digitized graphics tablet for monitoring parkinson's disease. In *The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014), 5* (pp. 171–175).

Kamalov, F., & Leung, H. H. (2020). Outlier detection in high dimensional data. *Journal of Information & Knowledge Management*, *19*(1), 2040013. https://doi.org/10.1142/S0219649220400134

Kandanaarachchi, S., & Hyndman, R. J. (2020). Dimension reduction for outlier detection using DOBIN. *Journal of Computational and Graphical Statistics*, *1–16*, 204–219. https://doi.org/10.1080/10618600.2020.1807353

Keller, F., Muller, E., & Bohm, K. (2012). HiCS: High contrast subspaces for density-based outlier ranking. In *2012 IEEE 28th international conference on data engineering* (pp. 1037–1048). IEEE Computer Society. https://doi.org/10.1109/ICDE.2012.88

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction* (1st ed.). Springer Science & Business Media. https://doi.org/10.1007/978-0-387-39351-3

Loperfido, N. (2020). Kurtosis-based projection pursuit for outlier detection in financial time series. *The European Journal of Finance*, *26*(2–3), 142–164. https://doi.org/10.1080/1351847X.2019.1647864

Ma, Y., & Fu, Y. (Eds.). (2011). *Manifold learning theory and applications* (1st ed.). CRC press. https://doi.org/10.1201/b11431

Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

Marques, H. O., Campello, R. J., Sander, J., & Zimek, A. (2020). Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *14*(4), 1–42. https://doi.org/10.1145/3394053

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*. https://doi.org/10.48550/ARXIV.1802.03426

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv: 1802.03426 [Cs, Stat]. http://arxiv.org/abs/1802.03426

Mordohai, P., & Medioni, G. (2010). Dimensionality estimation, manifold learning and function approximation using tensor voting. *Journal of Machine Learning Research*, *11*(12), 411–450. http://jmlr.org/papers/v11/mordohai10a.html

Muñoz, A., & Moguerza, J. M. (2004). One-class support vector machines and density estimation: The precise relation. In A. Sanfeliu, J. F. Martínez Trinidad, & J. A. Carrasco Ochoa (Eds.), *Progress in pattern recognition, image analysis and applications* (pp. 216–223). Springer. https://doi.org/10.1007/978-3-540-30463-0_27

Nane, S., Nayar, S., & Murase, H. (1996). Columbia object image library: COIL-20. Dept. Comp. Sci., Columbia University, New York, Tech. Rep.

Navarro-Esteban, P., & Cuesta-Albertos, J. A. (2021). High-dimensional outlier detection using random projections. *TEST*, *30*(4), 908–934. https://doi.org/10.1007/s11749-020-00750-y

Niyogi, P., Smale, S., & Weinberger, S. (2011). A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, *40*, 646–663. https://doi.org/10.1137/090762932

Pang, G., Cao, L., Chen, L., & Liu, H. (2018). Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2041–2050). Association for Computing Machinery. https://doi.org/10.1145/3219819.3220042

Pestov, V. (2000). On the geometry of similarity search: Dimensionality curse and concentration of measure. *Information Processing Letters*, *73*(1–2), 47–51. https://doi.org/10.1016/S0020-0190(99)00156-8

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Processes and their Applications*, *69*(1), 1–24. https://doi.org/10.1016/S0304-4149(97)00028-8

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer. https://doi.org/10.1007/b98888

Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. (2009). Finding anomalous periodic time series. *Machine Learning*, *74*(3), 281–313. https://doi.org/10.1007/s10994-008-5093-3

Ren, H., Chen, N., & Zou, C. (2017). Projection-based outlier detection in functional data. *Biometrika*, *104*(2), 411–423. https://doi.org/10.1093/biomet/asx012

Ro, K., Zou, C., Wang, Z., & Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika*, *102*(3), 589–599. https://doi.org/10.1093/biomet/asv021

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. John Wiley & Sons. https://doi.org/10.1002/0471725382

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326. https://doi.org/10.1126/science.290.5500.2323

Scott, C. D., & Nowak, R. D. (2006). Learning minimum volume sets. *The Journal of Machine Learning Research*, *7*(24), 665–704. http://jmlr.org/papers/v7/scott06a.html

Souvenir, R., & Pless, R. (2005). Manifold clustering. Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 648–653. https://doi.org/10.1109/ICCV.2005.149

Steyer, L., Stöcker, A., & Greven, S. (2021). Elastic analysis of irregularly or sparsely sampled curves. *Biometrics*, 1–13. https://doi.org/10.1111/biom.13706

Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). Nuclear feature extraction for breast tumor diagnosis. *Biomedical Image Processing and Biomedical Visualization*, *1905*, 861–870. https://doi.org/10.1117/12.148698

Tenenbaum, J. B., Silva, V. d., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Thudumu, S., Branch, P., Jin, J., & Singh, J. J. (2020). A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, *7*(1), 1–30. https://doi.org/10.1186/s40537-020-00320-x

Toivola, J., Prada, M. A., & Hollmén, J. (2010). Novelty detection in projected spaces for structural health monitoring. In P. R. Cohen, N. M. Adams, & M. R. Berthold (Eds.), *Advances in intelligent data analysis IX* (pp. 208–219). Springer. https://doi.org/10.1007/978-3-642-13062-5_20

Torgerson, W. S. (1952). Multidimensional scaling: I. *Theory and Method. Psychometrika*, *17*(4), 401–419. https://doi.org/10.1007/BF02288916

Unwin, A. (2019). Multivariate outliers and the O3 plot. *Journal of Computational and Graphical Statistics*, *28*(3), 635–643. https://doi.org/10.1080/10618600.2019.1575226

Wilkinson, L., Anand, A., & Grossman, R. (2005). Graph-theoretic scagnostics. INFOVIS 2005. *IEEE Symposium on Information Visualization, 2005*, Minneapolis, MN, USA, 2005, (pp. 157–164). https://doi.org/10.1109/INFVIS.2005.1532142

Xie, W., Kurtek, S., Bharath, K., & Sun, Y. (2017). A geometric approach to visualization of variability in functional data. *Journal of the American Statistical Association*, *112*(519), 979–993. https://doi.org/10.1080/01621459.2016.1256813

Xu, X., Liu, H., Li, L., & Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. *International Journal of Computational Intelligence Systems*, *11*(1), 652–662. https://doi.org/10.2991/ijcis.11.1.50

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*(1), 19–22. https://doi.org/10.1007/BF02287916

Zhang, J., & Zulkernine, M. (2006). Anomaly based network intrusion detection with unsupervised outlier detection. *2006 IEEE International Conference on Communications*, *5*, 2388–2393. https://doi.org/10.1109/ICC.2006.255127

Zimek, A., & Filzmoser, P. (2018). There and back again: Outlier detection between statistical reasoning and data mining algorithms. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(6), e1280. https://doi.org/10.1002/widm.1280

Zimek, A., Schubert, E., & Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *5*(5), 363–387. https://doi.org/10.1002/sam.11161

Zimek, A., & Vreeken, J. (2015). The blind men and the elephant: On meeting the problem of multiple truths in data from clustering and pattern mining perspectives. *Machine Learning*, *98*(1), 121–155. https://doi.org/10.1007/s10994-013-5334-y

How to cite this article: Herrmann, M., Pfisterer, F., & Scheipl, F. (2023). A geometric framework for outlier detection in high-dimensional data. *WIREs Data Mining and Knowledge Discovery*, *13*(3), e1491. https://doi.org/10.1002/widm.1491

APPENDIX A

A.1 | EXAMPLE VISUALIZATIONS OF THE DATA USED IN THE QUANTITATIVE EXPERIMENTS

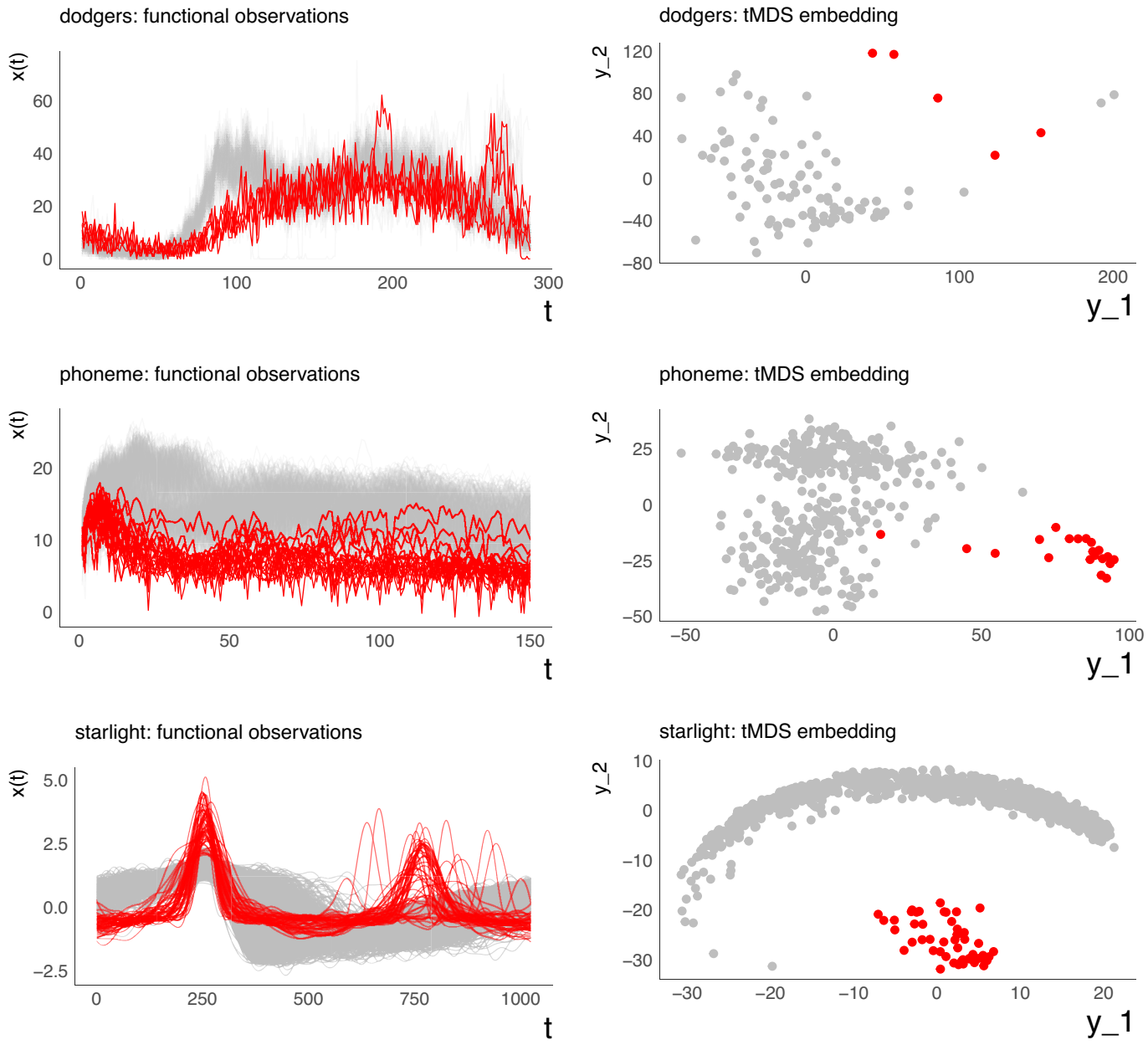

FIGURE A1 Plots to Table 1A: Functional data and tMDS embeddings. Inlier class in grey, outlier class in red. $r = 0.05$

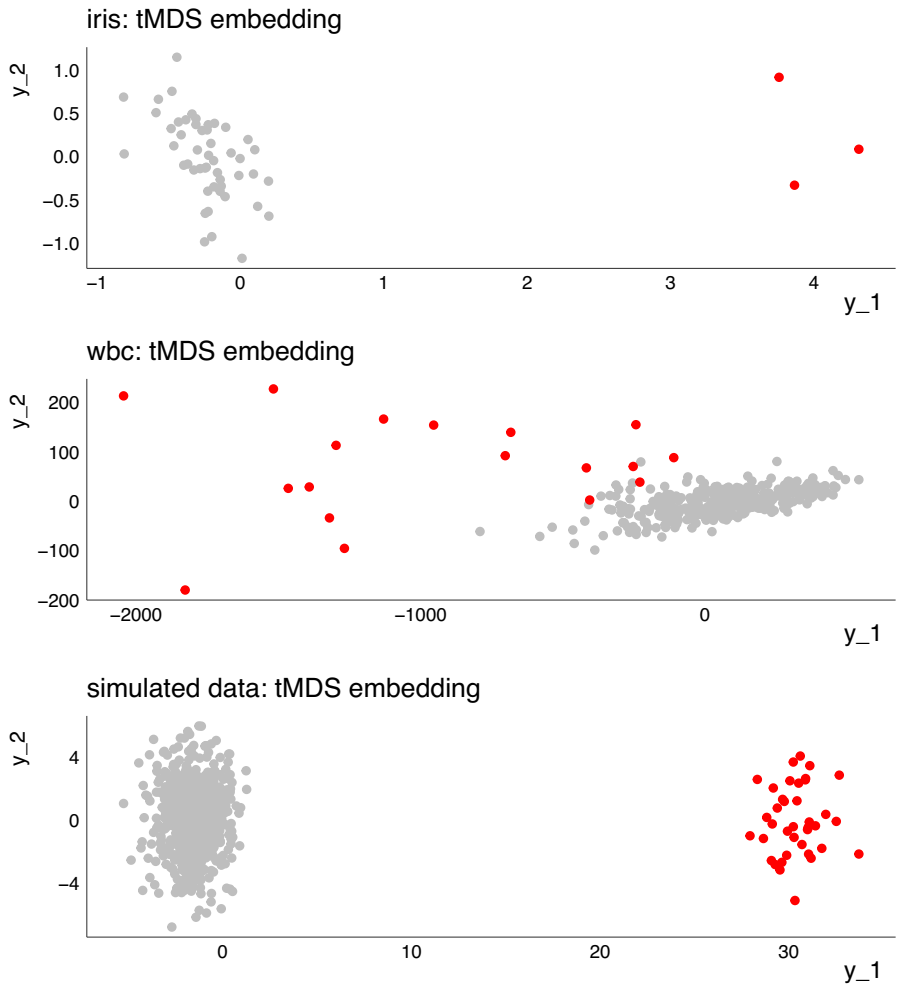

FIGURE A2 Plots to Table 1B: tMDS embeddings of tabular data. Inlier class in grey, outlier class in red. $r = 0.05$