Functional Outlier Detection for Density-Valued Data with Application to Robustify Distribution-to-Distribution Regression

Xinyi Lei*abc, Zhicheng Chen oabc, and Hui Liabc

*Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin, China; oKey Lab of Structures Dynamic Behavior and Control of the Ministry of Education, Harbin Institute of Technology, Harbin, China; abcSchool of Civil Engineering, Harbin Institute of Technology, Harbin, China

ABSTRACT

Distributional data analysis, concerned with the statistical analysis of data objects consisting of random probability distributions in the framework of functional data analysis (FDA), has received considerable interest in recent years and is increasingly applied in various fields including engineering. Outlier detection and robustness are of great practical interest; however, these aspects remain unexplored for distributional data. To this end, this study focuses on density-valued outlier detection and its application in robust distributional regression. Specifically, we propose a transformation-based approach for single-dataset outlying density detection with an emphasis on converting the less detectable shape outliers to easily detectable magnitude outliers. We also propose a distributional regression-based approach for detecting the abnormal associations of the density-valued two-tuples associated with two datasets. Then, the proposed outlier detection methods are applied to robustify a distribution-to-distribution regression method used in engineering, and we develop a robust estimator for the regression operator by downweighting the detected outliers. The proposed methods are validated and evaluated via extensive simulation studies. The relevant results reveal the superiority of our method over other competitors in distributional outlier detection. A case study in structural health monitoring demonstrates the great potential of our proposal in engineering applications.

1. Introduction

Driven by the rapid development of functional data analysis (FDA), statistical methods for random probability density function (PDF)-valued data have seen a research boom in recent years as such data can be widely encountered in many disciplines (Petersen, Zhang, and Kokoszka 2022) including engineering, such as quality control (Menafoglio et al. 2018), civil structural health monitoring (Chen et al. 2019). PDF-valued data are special functional data subjected to inherent constraints of taking nonnegative values and integrating into one. Such specificities make the PDF space lack a linear structure in the usual sense, which presents many methodological challenges in analyzing PDF-valued data (Kokoszka et al. 2019; Petersen, Zhang, and Kokoszka 2022). In the statistics literature, PDF-valued data are classified as distributional data to distinguish them from ordinary functional data. To date, distributional data analysis mainly concentrates on principal component analysis (Hron et al. 2016; Petersen and Müller 2016), regression analysis (Talská et al. 2018; Chen et al. 2019; Petersen and Müller 2019; Han, Müller, and Park 2020; Chen, Lin, and Müller 2021, Ghodrati and Panaretos 2022, 2022), time series analysis (Kokoszka et al. 2019; Jiang 2022; Zhang, Kokoszka, and Petersen 2022), correlation analysis (Zhou, Lin, and Yao 2021), causal inference (Lin, Kong, and Wang 2021), change-point detection (Dubey and Müller 2020; Horváth, Kokoszka, and Wang 2021) and so on. Many aspects remain unexplored, such as outlier detection and robustness improvement. The latter is a critical performance requirement in real applications, but it has not received sufficient attention. Outlier detection not only contributes to uncovering abnormal distributional patterns and providing more insights for the data but can also be exploited to robustify various related statistical procedures.

In structural health monitoring (SHM), our application of interest, distributional data arise in many situations, and anomaly detection or robust procedures are of particular practical interest. SHM concerns continuously monitoring the structural responses and service environments of civil infrastructure using various sensors and diagnosing structural health based on the collected data (Farrar and Worden 2013). In damage detection and structural condition assessment, the two core objectives of SHM, the anomalies in the distributions of extracted damage-sensitive features (DSF) are of great concern, since such anomalies may signify damage events. To facilitate tracking the abnormal behavior or temporal evolution in the distributions (related to structural conditions) of the DSF data, the corresponding derived distributional information is usually in the form of density- or histogram-valued streaming data (Li et al. 2018). In...
addition to being derived from monitoring data, distributional data also naturally arise in various SHM applications, typified by the emerging self-powered sensing concept. Some types of self-powered sensors directly record the cumulative durations of overthreshold events and store them in the form of histograms (Alavi et al. 2017; Azimi and Pekcan 2020). In such situations, the collected data are distributional data, and distributional outlier detection methods have great potential in subsequent damage detection. Outlier detection-based approaches form a major branch of unsupervised damage detection techniques. As stated by Farrar and Worden (2013, p. 54), when monitoring data at the damaged stage are not sufficient for training a valid supervised detector, the damage detection for the structure usually relies on outlier detectors. Distributional outlier detection remains an important topic and is seldom studied in both statistical and engineering fields. Meanwhile, different distributional data-analytic methods are increasingly employed for SHM applications (Chen et al. 2019, 2020; Momeni and Ebrahimkhaniou 2022); however, most of the methods lack robustness, seriously limiting their practical performances, as real SHM data usually contain various disturbances and contaminations (Yi, Huang, and Li 2017). Therefore, developing effective and reliable outlier detection and robust methods for distributional data is becoming a particularly pressing task, which is not only an urgent need for the SHM field but also for various other fields involving distributional data analysis.

Similar to ordinary functional data, PDF-valued data also have different types of outlyingness with the shape outlier (the curve possesses a significantly different shape but its “body” is usually masked among the majority of the curves) being the most difficult to detect. In fact, shape outlier detection is a common challenge in functional outlier detection (Arribas-Gil and Romo 2014; Dai et al. 2020). Another major challenge for outlying PDF detection stems from its specialty. Although various outlier detectors have been developed for ordinary functional data, they usually lose their efficacies when applied to PDFs due to the special nature possessed by such data. When considering two correlated PDF-valued datasets, another type of anomaly may exist that can significantly affect regression analysis, namely, an abnormal association of PDF pairs (an in-depth description will be provided in Section 3.2). The PDFs associated with such anomalies may both be normally behaved when viewed from a single dataset. Detecting such outliers is another challenge, as the functional outlier detectors developed in general routine are usually only suitable for the case of a single dataset. To date, functional outlier detectors for outlying PDF-valued two-tuples are still very rare.

Our research contributions are mainly 3-fold. First, we propose a computationally efficient functional outlier detection method for single-dataset distributional anomaly diagnosis. The main innovation is a transformation system for converting the less detectable shape outliers to more detectable magnitude outliers. Second, we propose a novel distributional regression-based diagnostic tool for detecting abnormal associations of density pairs between two datasets; finally, as an application, the proposed outlier detection methods are incorporated into a distributional regression framework to robustify a distribution-to-distribution regression (DtDR) model by downweighting the detected outliers. We derived a robust estimator for the regression operator using reproducing kernel Hilbert space (RKHS) theory, which is another novel aspect of our proposal. Recently, DtDR methods have garnered increased research interest in both the statistical and engineering fields, and representative approaches include the LQD transformation-based method (Chen et al. 2019), Wasserstein regression method (Chen, Lin, and Müller 2021) and optimal transport regression method (Ghedrati and Panaretos 2022, 2022). DtDR methods have great practical usefulness in various SHM applications, such as sensor fault correction (i.e., correct the information of distorted or corrupted data (Yi, Huang, and Li 2017), including missing data (Chen et al. 2019)). Sensor fault correction constitutes an important step of sensor validation (the other two steps are sensor fault detection and isolation; see Yi, Huang, and Li (2017) for more details). Practically, the detected distributional outliers of SHM data may be attributed to sensor faults rather than damages. In such situations, the distorted distributional information (i.e., outliers) is expected to be corrected using data reconstruction methods. The proposed robust distributional regression method can provide a useful tool for such applications.

This article is organized as follows: Section 2 briefly reviews the literature on functional outlier detection. Section 3 presents the proposed distributional outlier detection methods, followed by an application to robustify a distribution-to-distribution regression method in Section 4. Using synthetic data, the proposed methods are validated in Section 5. Then, a case study on SHM is presented in Section 6. Finally, the conclusions and discussions are summarized in Section 7.

2. Literature Review

Outlier detection for scalar data has been developed for a long time, but related studies on functional data have only begun in recent years (Kuhnt and Rehage 2016). For ordinary functional data, outliers can be generally divided into two categories, namely, magnitude outliers and shape outliers (Dai et al. 2020; Harris et al. 2021). Magnitude outliers are much easier to detect; however, detecting shape outliers is much more challenging (Arribas-Gil and Romo 2014; Dai et al. 2020).

To date, various functional outlier detection methods have been developed, most of which are either based on graphical detection tools or functional depth. The functional boxplot (Sun and Genton 2011) is the most popular graphical tool for functional outlier detection. Other representative graphical tools include the functional bagplot (Hyndman and Shang 2010), the outliergram (Arribas-Gil and Romo 2014) and the directional quantile envelope (Agarwal and Sun 2021). The standard functional boxplot is more suitable for detecting magnitude outliers rather than shape outliers, while the outliergram (Arribas-Gil and Romo 2014) and the directional quantile envelope (Agarwal and Sun 2021) have some potential for shape outlier detection. Functional depth is a statistical notion that can be used to measure the centrality of curves and has been widely applied in center-outwards rankings as well as outlier detection (López-Pintado and Romo 2009; Claeskens et al. 2014). One popular functional depth is the band depth (López-Pintado and Romo 2009) but it is insensitive to shape outliers. Researchers have also defined functional depths specifically for shape outlier detection, such as the functional tangential angle (FUNTA)
pseudo-depth (Kuhnt and Rehage 2016), modified integrated and infimal depths (Nagy, Gijbels, and Hlubinka 2017), functional directional outlyingness (Dai and Genton 2019), and elastic depth (Harris et al. 2021). The functional directional outlyingness (FDO)-based approach (Dai and Genton 2019) and the elastic-depth-based approach (Harris et al. 2021) use the strategy of decomposing the total information of curves into magnitude and shape information to enhance their effectiveness in shape outlier detection. Other functional depths that have been employed for functional outlier detection include the functional halfspace depth (Claeskens et al. 2014), skew-adjusted projection depth (Hubert, Roussieeuw, and Segaert 2015), total variation depth (Huang and Sun 2019), and others. The shape outliers in real datasets are complex and usually exhibit various patterns; hence, detecting them by using a single feature or single depth is generally difficult.

Recently, Dai et al. (2020) proposed a more effective approach for functional outlier detection by applying sequential transformations to convert shape outliers to easily detected magnitude outliers. However, the transformations introduced by Dai et al. (2020) are relatively scattered and unsystematic, and they have limitations in PDF-outlier detection. As mentioned above, PDF-valued data have their specialties, and ordinary functional outlier detectors usually perform poorly. The literature on PDF-outlier detection is quite rare, and many aspects remain unexplored. To this end, we present an alternative tree-structured transformation system for detecting PDF-outliers. The properties of the distributional data are fully considered to uncover the shape outliers, and computationally efficient outlier detectors are designed for the related transformed data. Moreover, we propose a nonparametric distributional regression-based approach for detecting abnormal associations of PDF pairs.

### 3. Distributional Outlier Detection Methods

#### 3.1. Single-Dataset Outlier Detection

Consider a PDF-valued dataset denoted as \( \{f_i(x)\}_{i=1}^{n} \) consisting of \( n \) different smooth univariate PDFs with support on \( I = [0, 1] \) (PDFs with other finite support can be converted to support on \( I = [0, 1] \)). Generally, such a PDF-valued dataset may contain two types of outliers: (a) the horizontal-shift outlier and (b) the shape outlier (see Figure 1 for a schematic illustration). The former is easier to detect, while detection of the latter is much more challenging because they are masked by the “curve net” of the majority of the data. Compared to the majority of the data, the outliers may tell a different story; thus, they are expected to be detected to further investigate the causes behind them. For instance, as mentioned in the introduction, the occurrence of outliers in SHM data may signify an event of structural damage or sensor fault. In addition, such outliers may also adversely affect the performances of some distributional data analytic methods that are susceptible to outliers.

The shape outliers in a real functional dataset often exhibit various patterns, which are generally difficult to fully screen out by a single feature or single outlyingness measure. Here, we consider a collection of transformations for extracting different features for outlier detection and for converting the less detectable shape outliers into more detectable magnitude outliers. Generally, there are two basic transformations that can effectively expose the abnormal patterns of the curves. The first one is performing a derivative of the functions, which helps to expose curves with abnormal slopes (Dai et al. 2020); the other one is centralization (i.e., shifting the curves along a direction to make the bulk of the curves more gathered), which helps to peel away the masking effects caused by the position variability of the bulk of the curves (Dai et al. 2020). The related illustrations that show how these two basic transformations work can be found in Section S.3.2.1 of the supplementary materials. The proposed transformations for distributional outlier detection are also guided by such basic principles but consider the specialties of the distributional data. One outstanding feature of distributional data is that the inverse functions are available when they are represented as cumulative distribution functions (CDFs), which enables us to introduce the transformations to the data (for exposing shape outliers) along two main directions (i.e., from the angles of the original function and its inverse). By performing derivative operations (in combination with logarithmic or normalized transformations) along these two directions, the resulting functions can be systematically represented as a chain of functions shown in Figure 2(a). The considered transformations for exposing the outlying PDFs are based on this basic transformation chain, and they can be assembled into a tree-structured transformation system with the PDF node as the root, as shown in Figure 2(b). Such a transformation system is called the transformation tree throughout this study. The purpose (or function) of each branch of the tree in outlier detection and its relationship with the transformation chain are detailed as follows:

**Branch I:** This branch contains 3 nodes (corresponding to nodes \( N_4 \sim N_6 \) on the transformation chain), and its main purpose is to expose the shape outliers by performing transformations to the inverse functions of the CDFs. The transformations involved in this branch have several appealing properties in outlier detection, which will be discussed later on.

---

**Figure 1.** (a) The horizontal-shift outlier (represented by dotted line) and (b) shape outlier (represented by dashed line).
Branch I starts with transforming the density functions \( \{f_i(x)\}_{i=1}^n \) into quantile functions (QFs) denoted as \( \{Q_i(t)\}_{i=1}^n \) (\( Q_i = F_i^{-1} \) is the inverse function of the CDF \( F_i(x) = \int_{-\infty}^x f_i(\tau) \, d\tau \)). As illustrated in Figure 3(b), in the QF-space, the horizontal-shift outliers have been converted to magnitude outliers; however, the shape outliers may still hide in the majority of the data. One appealing property of such a quantile-function-based transformation is that it provides a natural alignment for disordered functional data. To uncover the hidden shape outliers, we further perform a derivative calculation on the quantile functions and take the logarithm of the results, that is, \( \psi_i(t) = \log \left( \frac{dQ_i(t)}{dt} \right) = \log(q_i(t)) \), with \( q_i(t) \) being the quantile density function (this is actually the log quantile density (LQD) transformation defined by Petersen and Müller (2016)).

Note that \( q_i(t) \) has a constraint of \( q_i(t) \geq 0 \), and the logarithmic operation aims to release this constraint. Consequently, the resulting data \( \{\psi_i(t)\}_{i=1}^n \) are ordinary functions residing in a functional space possessing a linear structure (i.e., closed under addition and scalar multiplication); thus, they can be conveniently embedded into an appropriate metric space to perform outlier detection using the associated metric (which will be detailed later). Benefiting from the derivative operation involved in the LQD transformation, the shape outliers with abnormal slopes can significantly stand out in comparison to the majority of the data, as shown in Figure 3(c). Moreover, such an LQD transformation also has an appealing advantage in exposing the hidden shape outliers attributed to the horizontal variability of the PDF-valued data. The horizontal shift of a PDF (e.g., \( f_i(x) = f(x-c) \), where \( c \) stands for any constant) is equivalent to the vertical shift of the corresponding quantile function (e.g., \( Q_i(t) = Q(t) + c \)). The quantile density function \( q(t) \) in the LQD transformation is independent of the vertical shift of \( Q(t) \) due to \( \frac{dQ(t)}{dt} = \frac{dQ(t+c)}{dt} \). Consequently, the resulting functional data in the LQD node of the transformation tree are independent of the horizontal shifts of the PDFs (Kokoszka et al. 2019). Thus, the LQD transformation can peel away the influences caused by horizontal shifts of the PDFs, which has a similar effect as horizontally centralizing the PDFs (or vertically centralizing the QFs). In this sense, the LQD transformation combines the merits of derivative operations and centralization, which makes it a powerful tool for revealing shape outliers. However, for the distributional regression application involved later, the insensitivity of the LQD transformation to horizontal shift is required to be remedied by a data preprocessing described in Section S.2.1 of the supplementary materials.

Moreover, we recommend normalizing \( \psi_i(t) \) as \( \psi_{\text{norm},i}(t) = \psi_i(t) / \int_0^1 \psi_i(\tau) \, d\tau \). With this normalization, the majority of the data with relatively similar features gather more closely, making the outlying curves more distinguishable, as shown in Figure 3(d). According to our experience, this processing usually has favorable effects on shape outlier detection, especially for some real datasets.

Branch II: The purpose of this branch is to detect functional outliers for the PDF-valued data in their own space. More specifically, the PDFs are embedded into the Bayes space (a detailed introduction to this space can be found in Section S.1.1 of the supplementary materials) and borrow the associated Bayes metric to perform outlier detection. The Bayes space, denoted as \( \mathfrak{B}^2(I) \), is a linear space formed by the positive functions defined on \( I \) with square integrable logarithms (Van den Boogaart, Egozcue, and Pawlow sky-GLahn 2014; Hron et al. 2016), and a PDF supported on \( I \) is in itself an element of \( \mathfrak{B}^2(I) \). In essence, this branch is to perform outlier detection in node \( N_2 \) on the transformation chain.

It is worth noting that the Bayes space embedding imposes no transformation to the PDF-valued data, as they can be naturally embedded into the Bayes space without any change. To expose shape outliers, we conduct a horizontal centralization processing (implemented in the H-CENTR node) to the PDFs in the Bayes space. Specifically, all PDFs \( f_i(x)s \) are shifted along the \( x \)-direction to align them according to a pre-specified feature point (e.g., the median, mean, mode, or a combination of them). Let \( [u, v] \) be the new common support of the shifted PDFs denoted as \( \{\text{Hshift}f_i\}_{i=1}^n \), and we truncate each shifted PDF...
using the common support \([u, v]\) to obtain the function \(\tilde{f}_i\) in the H-CENTR node, that is, \(\tilde{f}_i(x) = \chi_{[u,v]}(x) \cdot \text{Hshift}[f_i](x), x \in [u, v]\), where \(\chi_{[u,v]}(\cdot)\) is an indicator function. The resulting functions \(\tilde{f}_i\)'s are also positive-valued with square-integrable logarithms, which are elements of the Bayes space \(\mathcal{B}_2^2([u,v])\). The Bayes space is a quotient space, where the two elements are identical if they are proportional (Hron et al. 2016), that is, \(\tilde{f} = \gamma \cdot \tilde{f}, \forall \gamma \in \mathcal{B}_2^2([u,v])\) with \(\gamma\) standing for an arbitrary positive constant and \(\equiv\) standing for the equal sign of the Bayes space. Therefore, although the resulting function \(\tilde{f}\) in the H-CENTR node may no longer be a PDF, it is equivalent to the PDF \(\tilde{f}/c_0\) with \(c_0 = \int_{[u,v]} \tilde{f}(x) \, dx\) if viewed in the Bayes space.

The Hilbert space structure of the Bayes space \(\mathcal{B}_2^2([u,v])\) naturally induces a metric for measuring the dissimilarity between two PDFs (Talská et al. 2018), that is, \(d_{\mathcal{B}_2^2} (\tilde{f}, \tilde{g}) = \left\| \tilde{f} \oplus (1 - \tilde{g}) \right\|_{\mathcal{B}_2^2}, \forall \tilde{f}, \tilde{g} \in \mathcal{B}_2^2([u,v]); \text{see Section S.1.1 of the supplementary materials for more details about this metric.}\) The Bayes metric \(d_{\mathcal{B}_2^2}\) can be employed to detect functional outliers in the Bayes space. However, directly computing the operations \(\oplus\) and \(\odot\) using their definitions (given in Equations (S-2a) and (S-2b) of the supplementary materials) is inconvenient. Fortunately, \(\mathcal{B}_2^2([u,v])\) is isometrically isomorphic to \(L^2([u,v])\) under the centered log-ratio (CLR) transformation, that is, \(d_{\mathcal{B}_2^2} (\tilde{f}, \tilde{g}) = d_{L^2} (\text{CLR}[\tilde{f}], \text{CLR}[\tilde{g}])\) (Talská et al. 2018), where \(d_{L^2}\) denotes the \(L^2\) distance and \(\text{CLR}[\tilde{f}](x) = \log \tilde{f}(x) - \frac{1}{\pi} \int_0^\infty f_u \log \tilde{f}(\tau) \, d\tau\) is the CLR transformation. Therefore, we add the CLR node to transform the data \(\{\tilde{f}_i\}_{i=1}^n\) (residing in \(\mathcal{B}_2^2([u,v])\)) into \(L^2([u,v])\) using the CLR transformation to facilitate the calculation of the Bayes metric (which will be used in the functional outlier detection procedure).

Branch III: This branch consists of only one node (i.e. DIFF, corresponding to node N1 on the transformation chain), which directly computes the derivatives to the PDFs. Such a transformation can help to expose the outlying PDFs with significant abnormal slopes in the PDF space.

Branch IV: Instead of exposing shape outliers as did Branches I–III, this branch mainly aims at detecting the horizontal-shift outlier (as illustrated in Figure 1(a)) using the median information. Given a PDF \(f_i\) on \([0, 1]\), the median is defined as \(\text{med}(f_i) = \inf\{x \in [0, 1] : \int_0^x f_i(t) \, dt \geq \frac{1}{2}\}\). The horizontal-shift outlier usually possesses an abnormal median due to its significant deviation along the \(x\)-direction; thus, it can be identified by performing outlier detection on the median dataset. The data in the MED node are scalar data; hence, the outliers can be directly detected by using the two-sided boxplot detector given in Equation (S-14) of the supplementary materials.

Remark. In addition to the transformations considered above, another possible transformation is the warping transformation. Given the CDFs \(\{F_i(x)\}_{i=1}^n\) associated with node N3 on the transformation chain shown in Figure 2(a), we can warp them to be aligned to a specified reference CDF \(F_{TP}(x)\), that is, \(F_i(g_i(x)) = F_{TP}(x), \) where \(g_i = F_i^{-1} \circ F_{TP}\) denotes the warping function that characterizes the phase of \(F_i\) with respect to \(F_{TP}\). The shape outliers (with abnormal phases) can be detected from the warping functions \(\{g_i(x)\}_{i=1}^n\) by using the phase distance defined by Harris et al. (2021). Our simulation study conducted later (i.e., Simulation study I in the online supplement, and the relevant results are reported in Table S-5) shows that such a strategy performs poorly with a high risk of false detection; thus, it is not adopted in our proposal. Next, we present the detection methods for the exposed shape outliers. For convenience, let \(\phi_i(t)\) be the transformed cur-
(Royden and Fitzpatrick 2010, p. 142). Moreover, due to the smoothness assumption of the PDFs, the data in nodes nLQD and DIFF are continuous functions that can be treated as elements of the $C(A)$ space endowed with the sup distance (see Table 1). Consequently, the $L^2$, $L^1$ and sup distances listed in Table 1 are all valid for performing outlier detection for the data in nodes nLQD and DIFF. Generally, the $L^2$ or $L^1$ distance is mainly for quantifying the global dissimilarities of the functional data, while the sup distance is mainly for quantifying the local dissimilarities; see Section S.3.2.3 in the supplementary materials for demonstrations and comparisons. Practically, the global and local dissimilarity measures have strong complementarity in functional outlier detection. For global dissimilarity quantification, the $L^2$ distance and $L^1$ distance have similar performances; the main simulation studies conducted later and presented in Section S.5.1 of the supplementary materials show that the $L^1$ distance performs slightly better. Hence, in this study, the $L^1$ distance and the sup distance are used as the default distance combination (one for global dissimilarity quantification and the other for local) to detect the outlying curves in nodes nLQD and DIFF. For the same node, the outlier detection using different distances is conducted independently, and then we merge the results. We now specify the distance for the CLR node. As noted above, the purpose of Branch II of the transformation tree is to perform outlier detection for the PDF-valued data in their own space, namely, the Bayes space. The Bayes space is a metric space endowed with the Bayes distance $d_{\text{Bayes}}$ and is isometrically isomorphic to the $L^2$ space (where the CLR-transformed data reside), that is, $d_{\text{Bayes}}(f_1, f_2) = d_{L^2}^{\text{CLR}}(f_1, f_2)$. Therefore, the $L^2$ distance is selected for the outlier detection in the CLR node to be equivalent to the original intention of performing outlier detection in the Bayes space for the centralized PDFs (after processing by the H-CENTR node).

For convenience, such a distance-based detection approach is referred to as the Tree-Distance detection method throughout this study. With the recommended distances, the related default argument setting of the Tree-Distance detection method is listed in Table S-1 in the supplementary materials.

### 3.2. Abnormal Association Detection for PDF Pairs between Two Datasets

We now consider another type of anomaly, that is, the abnormal association of PDF pairs between two datasets. Consider $n$ PDF-valued two-tuples, denoted as $\mathcal{T} = \{g_i(x), f_i(x)\}_{i=1}^n$, formed by elements from two PDF-valued datasets $\{g_i(x)\}_{i=1}^n$ and $\{f_i(x)\}_{i=1}^n$ supported on the common interval $I = [0, 1]$, and assume that $f_i(x)$ depends on $g_i(x)$ in some manner. Distributional correlations (or dependences) can be widely encountered in many practical situations; for instance, in bridge structural health monitoring, the distributions of the structural responses (e.g., strain, displacement) caused by the same excitation sources (e.g., traffic loads, temperature) can usually be highly correlated among the different nearby measurement points (Chen et al. 2019, 2020). We say that the association of a PDF two-tuple is abnormal if it significantly violates the dependence pattern, followed by the majority of the data. It is worth noting that a PDF pair with an abnormal association can behave as either abnormal or normal in their respective datasets (see Section S.3.3.1 in the supplementary materials for illustrations); thus, the single-dataset outlier detection methods described above have limitations in detecting such abnormal associations. If an appropriate distribution-to-distribution regression (DtDR) model has been used to fit the dependence of the PDF-valued two-tuples, and it is assumed that the DtDR model has approximated the dependence pattern of the majority of the data, then the outlying two-tuple might stand out as extraordinary residuals. With a slight abuse of terminology, we call such outliers regression outliers, and a DtDR-based approach will be developed for detecting them. Regression-based outlier diagnostics for scalar data started very early (Rousseeuw and Leroy 1987); however, related research contributions for functional outliers still appear to be quite rare.

In this study, the LQD-RKHS DtDR model proposed by Chen et al. (2019) is employed for this task. The main reasons for this choice are 2-fold: (a) the LQD-RKHS DtDR is a nonlinear and nonparametric model, which is more flexible in modeling the different dependence patterns of the distributional data; thus, the risk of model misspecification in outlier detection is much lower; (b) it is highly efficient because the regression operator has an explicit expression. At present, only the standard version of the LQD-RKHS DtDR model is available for regression outlier detection. Its robust version will be developed later but it is not suitable for outlier diagnosis, as its robustness relies on downweighting the outliers detected in advance.

#### 3.2.1. Preliminary: LQD-RKHS Distribution-to-Distribution Regression

Given the two correlated PDF-valued datasets $\{g_i(x)\}_{i=1}^n$ and $\{f_i(x)\}_{i=1}^n$, the general form of the distribution-to-distribution regression model for relating $g_i(x)$ to $f_i(x)$ can be formulated as $f_i(x) = \Psi(g_i(x)) + e_i(x)$ with $\Psi$ and $e_i$ being the regression operator and random error term, respectively. In the LQD-RKHS distributional regression framework (Chen et al. 2019), the PDFs are first transformed into the Hilbert space $L^2[0, 1]$ to release their inherent constraints by using the log-quantile-density (LQD) transformation (Petersen and Müller 2016), that is,

| Space | Elements | Metric (Distance) |
|-------|----------|------------------|
| $L^2(A)$ | Square integrable functions on $A$ | $L^2$ distance: $d^2_{L^2} (\phi_1, \phi_2) = \left( \int_A (\phi_1 (t) - \phi_2 (t))^2 \, dt \right)^{1/2}$ |
| $L^1(A)$ | Integrable real functions on $A$ | $L^1$ distance: $d^1_{L^1} (\phi_1, \phi_2) = \int_A |\phi_1 (t) - \phi_2 (t)| \, dt$ |
| $C(A)$ | Continuous real functions on $A$ | sup distance: $d^\text{sup} (\phi_1, \phi_2) = \sup_{t \in A} |\phi_1 (t) - \phi_2 (t)|$ |
| $\mathfrak{B}^2(A)$ | Positive functions on $A$ with square integrable logarithms | Bayes distance: $d_{\text{Bayes}} (f_1, f_2) = d^2_{L^2} (\phi_1, \phi_2)$, where $\phi_1 = \text{CLR}[f_1]$ and $\phi_2 = \text{CLR}[f_2]$ |
\[ \psi_i^s(t) = -\log \left( f_i^s \left( Q_{i,j}^s(t) \right) \right) \text{ with } \]
\[ f_i^s = \left( 1 - \alpha_{\text{mix}}^{LQD} \right) f_i + \alpha_{\text{mix}}^{LQD} i = 1, \ldots, n \] (2a)
\[ \psi_i^\varepsilon(t) = -\log \left( g_i^\varepsilon \left( Q_{i,j}^\varepsilon(t) \right) \right) \text{ with } \]
\[ g_i^\varepsilon = \left( 1 - \alpha_{\text{mix}}^{LQD} \right) f_i + \alpha_{\text{mix}}^{LQD} i = 1, \ldots, n \] (2b)

where \( Q_{i,j}^s \) and \( Q_{i,j}^\varepsilon \) are the quantile functions corresponding to \( f_i^s \) and \( g_i^\varepsilon \), respectively, and \( \alpha_{\text{mix}}^{LQD} \in [0.2, 0.5] \) is the PDF preprocessing parameter (such a processing is aimed at curing the “blindness” of the LQD transformation to horizontal translations of PDFs (see Section S.2.1 of the supplementary materials for more details). Then, a functional principal component analysis (FPCA) is conducted to reduce the Hilbertian data \( \psi_i^s(t) \) to finite-dimensional vector-valued data formed by the FPC scores denoted by \( \xi_i^s = [\xi_i^s1, \xi_i^s2, \ldots, \xi_i^sm] \in \mathbb{R}^m \) on the \( m \) dominant FPCs. Consequently, the original DtDR model is converted to a function-to-vector regression model
\[ \xi_i^s = F_{\text{reg}} \left( \psi_i^s(t) \right) + \varepsilon_i, \psi_i^s \in \mathbb{L}^2 [0, 1] \text{ and } \xi_i^s, \varepsilon_i \in \mathbb{R}^m \] (3)

where \( F_{\text{reg}} \) is the regression operator, and \( \varepsilon \) is the error term. Such dimension reduction processing mainly aims at decreasing the computational burden in estimating the regression operator; see Chen et al. (2019) for more details. \( F_{\text{reg}} \) is assumed to reside in a reproducing kernel Hilbert space (RKHS) \( \mathcal{H}(K) \) and is estimated by solving the following minimization problem:
\[ \min_{F_{\text{reg}} \in \mathcal{H}(K)} \left\{ \sum_{i=1}^{n} \left\| \xi_i^s - F_{\text{reg}} \left( \psi_i^s \right) \right\|_2^2 + \lambda_1 \left\| F_{\text{reg}} \right\|_{\mathcal{H}(K)}^2 \right\} \] (4)

where \( \lambda_1 \) is the regularization parameter. With this regression model, the prediction for a PDF can first be achieved in the \( \mathbb{L}^2 [0, 1] \) space and then is transformed back to the density space through the inverse LQD transformation. It is important to keep in mind that the effect of the added uniform distribution (represented by \( \alpha_{\text{mix}}^{LQD} \) in Equation (2)) should be cleared by using eq. (7) in Chen et al. (2019).

### 3.2.2. Abnormal Association (Regression Outlier) Detection Method

Based on the LQD-RKHS distributional regression method described above, we can construct two DtDR models for regression outlier detection, that is,

**Forward regression model:**
\[ f_i = \Gamma_i^{df} \left( \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD}, m, \lambda_1, \psi_i^s \right) \]

**Reverse regression model:**
\[ g_i = \Gamma_i^{rf} \left( \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD}, m, \lambda_1, \psi_i^s \right) \]

where \( \Gamma_i^{df} \left( \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD}, m, \lambda_1, \psi_i^s \right) \) denotes the regression model for regressing \( \{f_i\} \) to \( \{g_i\} \), \( \Gamma_i^{rf} \left( \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD}, m, \lambda_1, \psi_i^s \right) \) denotes the regression model for regressing \( \{f_i\} \) to \( \{g_i\} \), \( \mathcal{L}^{\text{index}} \) is the index set for training samples, \( \alpha_{\text{mix}}^{LQD} \) is the PDF preprocessing parameter in Equation (2), \( m \) is the truncation order of FPCA and \( \lambda_1 \) (or \( \lambda_2 \)) corresponds to the regularization parameter \( \lambda_1 \) in Equation (4). In regression outlier detection, \( \lambda_1 \) and \( \lambda_2 \) are adaptively determined using the generalized cross-validation strategy (Golub, Heath, and Wahba 1979; Lian 2007); see Appendix 1 in the supplementary materials for computational details. For more effective regression outlier detection, the forward and reverse regression models should be used jointly. The detection procedure using the forward regression model is summarized as follows:

1. Fit the DtDR model \( \Gamma_i^{df} \left( \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD} \right) \) using the training dataset \( \{g_i, f_i\} : f_i \in \mathcal{L}^{\text{index}} \}; 
2. Use the fitted DtDR model to predict \( \{f_i, \ldots, f_n\} \), and the results are denoted as \( \{\hat{f}_1, \ldots, \hat{f}_n\} \}; 
3. Compute the residuals using Algorithms S.1 and S.2 in the supplementary materials separately to obtain the following residual datasets:
\[ \mathcal{E}^{\text{df}}_{\mathcal{L}^{\text{index}}} = \left\{ \epsilon_{\mathcal{L}^{\text{index}}}, \epsilon_{\mathcal{L}^{\text{index}}} \right\}, \epsilon_{\mathcal{L}^{\text{index}}} = d_{\mathcal{L}^{\text{index}}} \left( f_i, f_i \right) / \theta_i, \alpha_{\text{mix}}^{LQD}, \right\} \]
\[ i = 1, \ldots, n \]
\[ \mathcal{E}^{\text{LQD}} = \left\{ \epsilon_{\mathcal{L}^{\text{index}}}, \epsilon_{\mathcal{L}^{\text{index}}} \right\}, \epsilon_{\mathcal{L}^{\text{index}}} = d_{\mathcal{L}^{\text{index}}} \left( f_i, f_i \right) / \alpha_{\text{mix}}^{LQD}, \right\} 
\[ i = 1, \ldots, n \]

where \( \epsilon_{\mathcal{L}^{\text{index}}} \) is the residual calculated by embedding the PDFs into the Bayes space, \( \theta_i \) is a user-specified parameter for alleviating the horizontal deviation effects (see Section S.3.3.2 in the supplementary materials for more details), and \( \epsilon_{\mathcal{L}^{\text{index}}} \) is the residual calculated using the LQD-transformed PDFs.

4. Perform outlier detection for the residual datasets \( \mathcal{E}^{\text{df}}_{\mathcal{L}^{\text{index}}} \) and \( \mathcal{E}^{\text{LQD}} \) by using the one-sided boxplot-based detector given in Equation (S-15) of the supplementary materials.

For convenience, the outlier detection procedure summarized in Steps 1–4 is referred to as the forward-regression-outlier-detector (FROD) and denoted as
\[ \mathcal{O}^{\text{df}}_{\mathcal{L}^{\text{index}}} = \mathcal{O}^{\text{rf}}_{\mathcal{L}^{\text{index}}} \left( g_i, f_i \right) / \mathcal{L}^{\text{index}}, \alpha_{\text{mix}}^{LQD}, \epsilon_{\mathcal{L}^{\text{index}}} \]

where \( \mathcal{O}^{\text{df}}_{\mathcal{L}^{\text{index}}} \) is the index set for detected outliers, and \( \epsilon_{\mathcal{L}^{\text{index}}} \) is the whisker parameter of the one-sided boxplot-based detector involved in Step 4. The implementation of the reverse-regression-outlier-detector (ROD), denoted as \( \mathcal{O}^{\text{rf}}_{\mathcal{L}^{\text{index}}} \), is similar to \( \mathcal{O}^{\text{df}}_{\mathcal{L}^{\text{index}}} \) (just switch \( g_i \) and \( f_i \)).

The regression operator in the standard LQD-RKHS DtDR is estimated by solving a regularized least squares (LS) problem (see Equation (4)). As noted in Rousseeuw and Leroy (1987), a regression model estimated by LS may mask some outliers, as the regression model itself is not robust. To reduce this risk, we employ an iterative detect-and-delete procedure, as summarized in Algorithm 1, for regression outlier detection. In each iteration, the detected outliers are removed from the training dataset (implemented by \( \mathcal{L}^{\text{index}} = \mathcal{L}^{\text{index}} \setminus \mathcal{O}^{\text{index}} \)); then, we use the remaining training data to fit the forward and reverse regression models and reperform the outlier detections using the new regression models. After several iterations, the regression models at the last iteration are much more likely to be fitted by using “good” data; then, we output the detected
regression outliers as the final results. Moreover, before performing the regression outlier detection using Algorithm 1, one can also conduct a first-round outlier filtration processing to the raw PDF-valued datasets by using the single-dataset detection methods discussed above to filter out the extremely anomalous functional samples that will potentially cause the fitted regression model to misbehave grossly.

Algorithm 1 Stepwise Detect-and-Delete Regression Outlier Detector

1: Determine the regularization parameters \( \lambda^\text{reg} \) and \( \lambda^\text{est} \) using the generalized cross-validation (GCV) procedure described in Appendix 1 in the supplementary materials.

2: Set \( T_{\text{index}} = \{1, 2, \ldots, n\} \), \( \lambda^\text{est} = \max\{\lambda^\text{est}, \theta_h\} \) and \( \lambda^\text{reg} = \max\{\lambda^\text{reg}, \theta_h\} \), where \( \theta_h \) is a user-specified threshold for preventing the regularization parameter from taking a too small value. Combined with the GCV, this is a reinforced strategy to avoid overfitting.

3: for \( k = 1 \) to \( N_{\text{reg}} \) do (\( N_{\text{reg}} \) is the user-specified number of iterations)

   a: Outlier detection using the forward and reverse regression models, respectively

   \[
   \mathcal{O}^\text{est}_{\text{index}} = \mathcal{D}^\text{est}_{\text{LQD-RKHS}} \left( g_{i,j} | T_{\text{index}}, \alpha^{\text{mix}}, \alpha^{\text{mix}}, m, \lambda^\text{est}, \theta_h, \lambda^\text{reg}, \rho, \alpha^{\text{mix}}, \lambda^\text{reg}, \rho \right)
   \]

   \[
   \mathcal{O}^\text{reg}_{\text{index}} = \mathcal{D}^\text{reg}_{\text{LQD-RKHS}} \left( f_{i,j} | T_{\text{index}}, \alpha^{\text{mix}}, \alpha^{\text{mix}}, m, \lambda^\text{reg}, \theta_h, \lambda^\text{reg}, \rho, \alpha^{\text{mix}}, \lambda^\text{reg}, \rho \right)
   \]

   b: Set \( \mathcal{O}_{\text{index}} = \mathcal{O}^\text{est}_{\text{index}} \cup \mathcal{O}^\text{reg}_{\text{index}} \) and \( T_{\text{index}} = T_{\text{index}} \setminus \mathcal{O}_{\text{index}} \)

end for

4. Output \( \mathcal{O}_{\text{index}} \)

4. Application to Robustify Distributional Regression

We apply the proposed distributional outlier detection methods to robustify the LQD-RKHS distribution-to-distribution regression model proposed by Chen et al. (2019). The main steps of such a distributional regression method have been briefly recalled earlier in Section 3.2.1. The proposed robust regression procedure is in two-stages: (a) detect the outlying PDFs and (b) downweight the detected outliers in the regression operator estimation, similar to the strategy of Martínez-Hernández, Genton, and González-Farías (2019). Compared to the “crude” hard rejection, in terms of directly removing the detected outliers, such a downweighting approach belongs to a more flexible soft rejection strategy. The latter enables us to treat the outliers separately by designing different weights for the different types of outliers. For instance, in structural health monitoring, outliers are not always attributed to meaningless external interferences, they may also be induced by changes in the structural conditions; the latter scenario may contain useful information. Indiscriminately removing all the outliers would result in information loss as well as reducing the efficiency of the estimate (Gervini 2012).

In the RKHS-based functional regression framework (Lian 2007), the unknown regression operator \( F_{\text{reg}} \) is assumed to live in an RKHS denoted as \( \mathcal{H} (K_r) \) associated with the operator-valued reproducing kernel \( K_r \) (the related notations and basic properties of the RKHS that are used in this study are provided in Section S.1.2 of the supplementary materials). To cope with the adverse influence of the outlying training samples, the regression operator is estimated by solving the following weighted regularized least squares problem:

\[
\min_{F_{\text{reg}} \in \mathcal{H} (K_r)} \left\{ \sum_{i=1}^{n} w_i \left[ \psi_i^T - F_{\text{reg}} \left( \psi_i^x \right) \right]_2^2 + \lambda_s \left\| F_{\text{reg}} \right\|^2_{\mathcal{H} (K_r)} \right\}
\]

(6)

where \( \lambda_s \) is the regularization parameter, \( \| \cdot \|_2 \) and \( \| \cdot \|_{\mathcal{H} (K_r)} \) are the norms induced by the inner products of the Hilbert spaces \( \mathbb{R}^m \) and \( \mathcal{H} (K_r) \), respectively, and \( w_i \) is the weight corresponding to the \( i \)-th PDF pair \( \{g_i,f_i\} \). Detailed discussions about how to design the weights based on the detected outliers are provided in Section S.4.1 of the supplementary materials.

Proposition 1. The solution of \( F_{\text{reg}} \) in Equation (6) obeys the following general structure (proof in Section S.4.2 of the supplementary materials):

\[
F_{\text{reg}} \left( \cdot, \psi_i^x \right) \alpha_j = F_{\text{reg}} \left( \psi_i^x \right) \alpha_j
\]

(7)

where \( \alpha_j \) is the undetected vector-valued coefficient assumed to reside in a Hilbert space, and \( K_r \left( \cdot, \cdot \right) \) is the operator-valued reproducing kernel associated with \( \mathcal{H} (K_r) \).

A commonly used operator kernel is the Gaussian operator kernel (Lian 2007), given by

\[
K_r \left( \psi_i^x, \psi_j^x \right) = \exp \left\{ -\frac{1}{2\sigma^2} \int \left\| \psi_i^x (\tau) - \psi_j^x (\tau) \right\|^2 d \tau \right\}
\]

(8)

where \( I_{Id} \) denotes the identity operator. The coefficient \( \alpha_j \) given in Equation (7) is also assumed to reside in an RKHS \( H(k_r) \) with real reproducing kernel \( k_r \). Thus, according to the representer theorem of the real RKHS theory (Schölkopf, Herbrich, and Smola 2001), \( \alpha_j \) has the following representation:

\[
\alpha_j \left( \cdot \right) = \sum_{i=1}^{m} b_j k_r (\cdot, l) \mapsto \alpha_j (k)
\]

(9)

where \( \alpha_j (k) \) denotes the \( k \)-th element of the \( m \)-dimensional row vector \( \alpha_j \). Let

\[
\mathbf{K} = \{ k_r (k,l) \}_{k,l=1}^{k,m} \text{ and } \mathbf{b}_j = [b_{j1}, \ldots, b_{jm}], j = 1, \ldots, n
\]

(10)

where \( \mathbf{K} \) is the Gram matrix with elements \( K (k,l) = k_r (k,l) \). According to RKHS theory, \( \mathbf{K} \) is a symmetric and positive semidefinite matrix. Hence, Equation (9) can be expressed in the matrix form

\[
\alpha_j = \mathbf{b}_j \mathbf{K}^T = \mathbf{b}_j \mathbf{K}, j = 1, \ldots, n.
\]

The vector-valued coefficient \( \alpha_j \) lives in the Euclidean space \( \mathbb{R}^m \), which
can be extended to an RKHS by endowing it with a suitable reproducing kernel. It is worth noting that the reproducing kernel depends, not only on the vector space, but also on the equipped inner product (Wu and Lin 2012). The commonly used inner product for \( \mathbb{R}^m \) is \( \langle u, v \rangle = \sum_{i=1}^{m} u_i v_i, \forall u, v \in \mathbb{R}^m \), which is selected as the default inner product for \( \mathbb{R}^m \) unless otherwise stated. Then, the corresponding reproducing kernel is \( k_r(k, l) = \delta_{lk}, k, l \in \{1, 2, \ldots, m\} \), where \( \delta_{lk} \) is the Kronecker delta. Hence, the Gram matrix given in Equation (10) is the identity matrix \( E \) of size \( m \times m \). The inner product for \( \mathbb{R}^m \) can also be defined in other forms, and the Gram matrix \( K \) might no longer be an identity matrix. To ensure that our method can accommodate more general cases, the robust estimator for the regression operator is derived for a general form of the Gram matrix \( K \) (i.e., \( K \) is symmetric and positive semidefinite but not necessarily the identity matrix). For convenience, we define the following matrices:

\[
A = \{a_{ij}\}_{i=1}^{n} \in \mathbb{R}^{n \times n}, \quad W = \text{diag} (\sqrt{w_1}, \ldots, \sqrt{w_n}), \quad B = [\beta_1^T, \ldots, \beta_n^T]^T, \quad Y = \begin{pmatrix} (\xi_1^n)^T \\ \vdots \\ (\xi_n^n)^T \end{pmatrix}^T
\]

where \( a_{ij} \) and \( \beta_i \) are given in Equations (8) and (10), respectively. The following proposition presents the condition for the coefficient matrix \( B \) when Equation (6) reaches its minimum value.

**Proposition 2.** Under conditions (6), (7), (8), and (9), the optimal solution of \( B \) obeys

\[
(C_1 C_2 + \lambda_3 (K \otimes A)) \text{vec}(B) = C_1 \text{vec}(WY)
\]

where \( C_1 = K \otimes (AW) \) and \( C_2 = K \otimes (WA) \).

The proof is given in Section S.4.3 of the supplementary materials. Consequently, the coefficient matrix \( B \) corresponding to the robust regression operator can be estimated as

\[
\text{vec}(\hat{B}) = \left( (K \otimes (AW)) (K \otimes (WA)) + \lambda_3 (K \otimes A) \right)^{-1}(K \otimes (AW)) \text{vec}(WY)
\]

where \( D^{-1} \) denotes the Moore-Penrose pseudoinverse of matrix \( D \).

### 5. Simulation Studies

Extensive simulation studies have been conducted to validate and evaluate the performances of the proposed distributional outlier detection method; however, due to space limitations, the simulation studies are all reported in the supplementary materials; see Sections S.5–S.6 in the supplementary materials.

For the proposed Tree-Distance method, a comparative study is conducted in Simulation study I. Two state-of-the-art ordinary functional outlier detection methods (i.e., the functional directional outlyingness-based approach (Dai and Genton 2019) and the elastic depth-based approach (Harris et al. 2021) are considered as competitors. The detection performances are quantified by the average correct and false detection rates over 1000 repeated detection experiments, and the relevant results are reported in Tables S.3–S.5, showing that our method significantly outperforms the two competitors. The detection results shown in Table S.3 of Simulation study I and Table S-9 of the Additional simulation study I also demonstrate that the considered transformations on the transformation tree have good complementarity in outlier detection.

For the proposed distributional regression-based abnormal association detection method, we conduct two simulation studies (i.e., Simulation study II and the Additional simulation study II). The performances are evaluated based on the average correct and false detection rates over 500 repeated detection experiments, and several different contamination scenarios have been considered. In Simulation study II, the average correct detection rates are all greater than 98%, with 92.24% being the best (see Table S-8). In the Additional simulation study II, the average correct detection rates are all greater than 98% (see Table S-10). These results validate the effectiveness of the proposed method.

### 6. Application in Structural Health Monitoring (SHM)

An engineering application is provided to illustrate the practical utility of our methods. Specifically, the proposed robust distributional regression-based abnormal association detection method is applied to reconstruct the distributional information of the sensor data collected by an SHM system under severe contamination. The SHM system of a civil structure usually operates in complex and harsh environments, and sensor malfunction and external interference are very frequent, leading to data missing and corruption. Generally, the lifespans of sensors consisting of electronic components are much shorter than the service life of the structure itself, and many sensors have been embedded inside the structure, resulting in difficulties in repairing or replacing them once they fail; this is the major cause of massive data being lost. In a broad sense, the data that are severely corrupted by meaningless noises or outliers can also be regarded as lost. Both the missing and corrupted data can be treated as sensor faults, which can bring adverse impacts on SHM applications from many aspects (Sun et al. 2020). Thus, missing data reconstruction (including corrupted data correction) is an important research topic in SHM, especially for data preprocessing (Sun et al. 2020) and sensor validation (Yi, Huang, and Li 2017). This real data study will illustrate the potential of our method in robust distributional information reconstruction.

The investigated data are strain measurements collected by two strain gauges installed on the bottom plate of a steel box of a long-span bridge. For convenience, these two sensors are called Sensor A and Sensor B (the layout of these two sensors are presented in Figure S-28 of the supplementary materials). A total of 8 days of data collected on August 4, 5, 9, 10, 11, 12, 14, and September 10 of 2012 are selected for investigation. The selected days are not continuous because the data of Sensor B in the other days of the two months are either almost completely contaminated by meaningless outliers or are missing. For each sensor, we merge the selected data to form a single time series, and the results are visualized in Figure 4. Even in these selected data with relatively “higher quality,” there are still many abnormally large values contained in the data of Sensor B. Such a phenomenon is caused by the intermittent sensor faults rather than structural damages. Using the data processing procedures in Section S.7.2 of the supplementary materials, we obtain 120 pairs of PDFs, denoted as \( \{\hat{g}_i\}_{i=1}^{120} \) (Sensor A) and...
In the distribution reconstruction problem, the distributions of missing data are reconstructed by the predicted distributions obtained by the distributional regression model. In this case study, several PDFs from \( \{ \hat{f}_i \}_{i=1}^{120} \) (Sensor B) will be assumed to be missing, and then the regression model for relating PDFs \( \hat{g}_k \) to \( \hat{f}_i \) will be built for reconstructing the missing distributions. Before regression, we perform a two-stage initial outlier detection on the datasets \( \{ \hat{g}_i \}_{i=1}^{120} \) and \( \{ \hat{f}_i \}_{i=1}^{120} \) : (i) Single dataset outlier detection: Outlier detections for the datasets \( \{ \hat{g}_i \}_{i=1}^{120} \) and \( \{ \hat{f}_i \}_{i=1}^{120} \) are conducted independently by using the Tree-Distance method. (ii) Regression outlier detection: Outlier detection for datasets \( \{ \hat{g}_i \}_{i=1}^{120} \) and \( \{ \hat{f}_i \}_{i=1}^{120} \) is conducted jointly by using the method described in Section 3.2 after the outliers detected in the first stage have been removed. The argument settings for the detectors are listed in Table S-11 (in the supplementary materials). The outliers detected in the first and second stages are called Type I and Type II outliers, and the relevant detection results are presented in Figure 5. The investigated PDFs contain all types of outliers described in this study, namely, shape outliers, horizontal-shift outliers and regression outliers. The detected Type II outliers correspond to the abnormal associations of the PDF-valued two-tuples; the three pairs of PDFs (represented by bold curves in Figure 5(b)) detected as Type II outliers are associated with the 35 test PDFs denoted as \( \{ \hat{g}_k, \hat{f}_k \} \), the prediction error is quantified by the integrated absolute error (IAE), defined as \( \epsilon_{IAE} = \int \left| \hat{f}_k(\tau) - \hat{f}_k(\tau) \right| dt \), where \( \hat{f}_k \) denotes the prediction of \( \hat{f}_k \) using \( \hat{g}_k \) as the predictor. The boxplots for the prediction errors associated with the 35 test PDFs are plotted in Figure 6(a) for the standard and robust versions. As expected, the robust approach is significantly more accurate than the standard approach. For comparison, 10 representative predicted PDFs obtained by the two distributional regression methods are shown in Figure 6(b). Most of the predicted PDFs obtained by the standard LQD-RKHS method exhibit significant deviations from the true PDFs to the right side, indicating that the estimated regression operator has been seriously distorted by the outlying PDFs. In contrast, the proposed robust estimator performs much better, indicating that the robust version is more suitable for distribution reconstruction when the data are seriously contaminated.

In addition to the comparative study conducted above, more comparative studies can be found in the online supplement. For instance, a sensitivity analysis conducted in Section S.7.5 shows that the robustness of our method is not sensitive to the number of data segments. Section S.7.6 provides statistical analysis tests to verify the validity of the reconstructed distributions using inter-sensor correlations. Section S.7.7 provides additional distribution reconstruction tests to consider the case that a large portion of PDFs are consecutively missing, and our method still shows good performance. Moreover, in Section S.7.8 of the supplementary materials for an additional discussion.
supplementary materials, the readers can also find a discussion on the practical utility of the reconstructed distributions in SHM applications.

7. Conclusions and Discussions

This study develops functional outlier detection methods for detecting different types of outlying PDFs by considering the specialities of such data. Based on the proposed functional outlier detection methods, we develop a robust estimator for the LQD-RKHS distribution-to-distribution regression model. The feasibility and usefulness of our proposal are validated by simulation and real data studies.

The Tree-Distance approach leverages various transformations to convert the less detectable shape outliers to more detectable magnitude outliers, and the “exposed” outliers can be easily detected using appropriate distances. Such a strategy is simple and easy to use, and it enables the extraction of different features for outlier detection. Related simulation studies show that the Tree-Distance method can significantly outperform the competitors. Moreover, the simulation studies also show that the transformations have good complementarity in distributional outlier detection.

The distributional regression-based detection method provides an effective tool for detecting abnormal associations of PDF-valued two-tuples. The latter is another anomaly that may hide in the distributional data when PDFs from the two datasets are of concern. Outlier detection methods designed for single dataset inspection are generally ineffective in detecting such anomalies. Related simulation and real data studies demonstrate the effectiveness of the proposed distributional regression-based detection method in abnormal association detections for PDF-valued two-tuples.

The robustness of the proposed robust distributional regression method is achieved by downweighting the impacts of the detected outliers, enabling us to treat the outliers separately by designing different weights for different types of outliers. The real data study shows that the standard LQD-RKHS distributional regression method is highly sensitive to outliers, while the proposed robust version exhibits good performance even when the raw data are seriously contaminated by outliers.

At present, our proposal is only applicable to univariate PDFs, as the quantile function is unavailable for a multivariate distribution; extending to multivariate cases will be a future work. The distributional regression-based data reconstruction approach can only use the spatial correlation of the sensor data to reconstruct the distributions of missing data at its present stage, and tailoring it to consider spatiotemporal correlations will also be a future work.

Supplementary Materials

Supplementary file: Supplementary materials for “Functional outlier detection for density-valued data with application to robustify distribution-to-distribution regression”. This file includes theoretical background, additional technical details, simulation studies, details of proofs, additional illustrative results, data preprocessing procedures, additional comparative studies, additional discussions, etc.

Acknowledgments

We gratefully acknowledge the invaluable comments from the Editor, Associate Editor and two Referees that greatly improved the manuscript.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was financially supported by the National Natural Science Foundation of China (Grant No. 51908166), China Postdoctoral Science Foundation (Grant No. 2019M661287) and Postdoctoral Science Foundation of Heilong Jiang province.

ORCID
Zhicheng Chen https://orcid.org/0000-0002-1587-5892

References

Agarwal, G., and Sun, Y. (2021), “Bivariate Functional Quantile Envelopes with Application to Radiosonde Wind Data,” Technometrics, 63, 199–211. [352]
Alavi, A. H., Hasni, H., Jiao, P., Borchani, W., and Lajnef, N. (2017), “Fatigue Cracking Detection in Steel Bridge Girders through a Self-Powered Sensing Concept,” Journal of Constructional Steel Research, 128, 19–38. [352]
Arribas-Gil, A., and Romo, J. (2014), “Shape Outlier Detection and Visualization for Functional Data: The Outliergram,” Biostatistics (Oxford, England), 15, 603–619. [352]
Azimi, M., and Pekcan, G. (2020), “Structural Health Monitoring Using Extremely Compressed Data Through Deep Learning,” Computer-Aided Civil and Infrastructure Engineering, 35, 597–614. [352]
