

ONLINE SUPPLEMENT

Supplementary materials for “Functional outlier detection for density-valued data with application to robustify distribution-to-distribution regression”

Xinyi Lei^{a,b,c,l}, Zhicheng Chen^{a,b,c,l,*} and Hui Li^{a,b,c}

^aKey Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin, 150090, China;

^bKey Lab of Structures Dynamic Behavior and Control of the Ministry of Education, Harbin Institute of Technology, Harbin, 150090, China;

^cSchool of Civil Engineering, Harbin Institute of Technology, Harbin, 150090, China.

*Corresponding author: Zhicheng Chen; e-mail address: zhichengchen@hit.edu.cn; Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin, 150090, China.

^lThe first two authors contributed equally to this work.

Contents

S.1. Theoretical background	1
S.1.1. Bayes space.....	1
S.1.2. Reproducing kernel Hilbert space (RKHS)	2
S.2. PDF preprocessings for LQD and CLR transformations	2
S.2.1. PDF preprocessing for LQD transformation.....	2
S.2.2. PDF preprocessing for CLR transformation	4
S.3. Supplemental materials for distributional outlier detection	5
S.3.1. Modified boxplot-based detectors for scalar outlier detection.....	5
S.3.2. Supplemental materials for the single-dataset outlier detection method.....	5
S.3.2.1. Illustrations of the two basic transformations: derivative and centralization	5
S.3.2.2. Discussion, justification and comparison for the central function selection.....	7
S.3.2.3. Discussion, justification and comparison for distance (metric) selection.....	11
S.3.2.4. Default settings for the Tree-Distance outlier detection method	16
S.3.3. Supplemental materials for the abnormal association detection method	17
S.3.3.1. Illustrations of representative abnormal PDF-pairs	17
S.3.3.2. Residual calculation using the CLR transformation	18
S.3.3.3. Residual calculation using the LQD transformation.....	21
S.4. Supplemental materials for robust distributional regression.....	21
S.4.1. Weight design for robust regression operator estimation.....	21
S.4.1.1. Weight design for Type I outliers.....	22
S.4.1.2. Weight design for Type II outliers	23
S.4.1.3. Final weight	24
S.4.2. Proof of proposition 1	24
S.4.3. Proof of proposition 2	25
S.5. Simulation studies.....	26
S.5.1. Simulation study I.....	26
S.5.2. Simulation study II	31
S.6. Additional simulation studies.....	34
S.6.1. Additional simulation study I.....	34
S.6.2. Additional simulation study II	36
S.7. Supplemental materials for the real data study	37
S.7.1. The two investigated strain sensors	37
S.7.2. Data preprocessing in the real data analysis	37
S.7.3. Tables of argument settings involved in the real data study	38
S.7.4. Additional discussion on the detected Type II outliers	39
S.7.5. Sensitivity analysis	39
S.7.6. Validity validation for the reconstructed distributions	43
S.7.7. Non-random missing case study	47
S.7.8. Discussion on the practical utility of distribution reconstruction in SHM applications	49

Appendix 1: Adaptive regularization parameter selection for the LQD-RKHS distributional regression model	50
Appendix 2: Basic outlier generation algorithm	51
Appendix 3: Abnormal association-generating process for simulation study II	51
Appendix 4: Two considered competitors originally for ordinary functional outlier detection	53
Appendix 5: Demonstration of nonparametric regression-based dependence quantification	59
References	61

S.1. Theoretical background

S.1.1. Bayes space

The Bayes space, denoted as $\mathfrak{B}^2(I)$, consists of positive functions defined on the common compact interval $I = [a, b]$ with square-integrable logarithms (Van den Boogaart et al. 2014; Hron et al. 2016; Petersen et al. 2022), i.e.,

$$\mathfrak{B}^2(I) = \{f: I \rightarrow \mathbb{R} \mid f(x) > 0, \forall x \in I \text{ and } \int_I |\log f(\tau)|^2 d\tau < +\infty\} \quad (\text{S-1})$$

The Bayes space $\mathfrak{B}^2(I)$ is a separable Hilbert space under the following linear operations and inner product (Egozcue et al. 2006; Van den Boogaart et al. 2014; Hron et al. 2016):

(1) Linear operation

$$\text{Perturbation: } (f \oplus g)(x) = \frac{f(x)g(x)}{\int_I f(\tau)g(\tau)d\tau}, \quad x \in I, \text{ and } f, g \in \mathfrak{B}^2(I) \quad (\text{S-2a})$$

$$\text{Powering: } (\beta \odot f)(x) = \frac{f(x)^\beta}{\int_I f(\tau)^\beta d\tau}, \quad x \in I \text{ and } \beta \in \mathbb{R}, f \in \mathfrak{B}^2(I) \quad (\text{S-2b})$$

where \mathbb{R} stands for the set of real numbers. The perturbation and powering are analogous to the point-wise addition and scalar multiplication of the $L^2(I)$ space.

(2) Inner product

$$\langle f, g \rangle_{\mathfrak{B}} = \frac{1}{2\eta} \int_I \int_I \log \frac{f(t)}{f(s)} \log \frac{g(t)}{g(s)} dt ds, \quad f, g \in \mathfrak{B}^2(I) \quad (\text{S-3})$$

where $\eta = b - a$ is the Lebesgue measure of the compact interval $I = [a, b]$.

Obviously, the univariate continuous PDF supported on $I = [a, b]$ is an element of the Bayes space $\mathfrak{B}^2(I)$. One can easily verify that the space of such PDFs is closed under the linear operations defined in Eq.(S-2). Consequently, the univariate continuous PDFs with common finite support can naturally embedded into a Bayes space.

The Hilbert structure implies that the Bayes space $\mathfrak{B}^2(I)$ is a metric space endowed with the Bayes metric (Talská et al. 2018), i.e.,

$$d_{\mathfrak{B}}(f, g) = \|f \ominus g\|_{\mathfrak{B}} = \|f \oplus (-1 \odot g)\|_{\mathfrak{B}}, \quad \forall f, g \in \mathfrak{B}^2(I) \quad (\text{S-4})$$

where $\|\cdot\|_{\mathfrak{B}} = (\langle \cdot, \cdot \rangle_{\mathfrak{B}})^{1/2}$ stands for the norm induced by the inner product defined in Eq. (S-3).

The Bayes space $\mathfrak{B}^2(I)$ is isometrically isomorphic to the $L^2(I)$ space with the following centered log-ratio (CLR) transformation as the isomorphic mapping (Egozcue et al. 2006; Talská et al. 2018):

$$\text{CLR}[f](x) = \log f(x) - \frac{1}{\eta} \int_I \log f(\tau) d\tau, \quad x \in I, \quad f \in \mathfrak{B}^2(I) \quad (\text{S-5})$$

Then, it follows that

$$d_{\mathfrak{B}}(f, g) = d_{L^2}(\text{CLR}[f], \text{CLR}[g]), \quad \forall f, g \in \mathfrak{B}^2(I) \quad (\text{S-6})$$

where d_{L^2} stands for the L^2 distance defined as $d_{L^2}(\phi_1, \phi_2) = (\int (\phi_1(\tau) - \phi_2(\tau))^2 d\tau)^{1/2}$, $\forall \phi_1, \phi_2 \in L^2(I)$.

S.1.2. Reproducing kernel Hilbert space (RKHS)

(1) Real reproducing kernel

Let D be an arbitrary set, and let $H(k_r)$ be the real RKHS associated with the real reproducing kernel k_r . According to the real RKHS theory (Wahba 1990; Berlinet and Thomas-Agnan 2004; Lian 2007a), $H(k_r)$ is a subspace of $\{f: D \rightarrow \mathbb{R}\}$ consisting of functions defined on D . The reproducing kernel k_r is a symmetric semi-definite function defined on $D \times D$, i.e., $k_r: D \times D \rightarrow \mathbb{R}$, satisfying the following properties (Wahba 1990; Berlinet and Thomas-Agnan 2004; Lian 2007a)

$$k_r(x, \cdot) \in H(k_r), \forall x \in D \quad (\text{S-7})$$

and

$$f(x) = \langle k_r(x, \cdot), f \rangle_{H(k_r)}, \forall x \in D, f \in H(k_r) \quad (\text{S-8})$$

where, $\langle \cdot, \cdot \rangle_{H(k_r)}$ is the inner-product endowed to the RKHS $H(k_r)$.

(2) Operator-valued reproducing kernel

Let G and \tilde{G} be two arbitrary Hilbert spaces, let $\mathcal{H}(K_r)$ denote the RKHS associated with operator-valued reproducing kernel K_r . According to the theory of operator-valued kernels (Kadri et al. 2016; Lian 2007a), $\mathcal{H}(K_r)$ is a subspace of the operator space $\{F: G \rightarrow \tilde{G}\}$ consisting of mappings from G to \tilde{G} . The operator-valued kernel K_r is a symmetric semi-definite mapping from the product Hilbert space $G \times G$ to the other Hilbert space \tilde{G} , i.e., $K_r: G \times G \rightarrow \tilde{G}$, satisfying (Kadri et al. 2016; Lian 2007a)

$$K_r(\psi, \cdot) \in \mathcal{H}(K_r), \forall \psi \in G \quad (\text{S-9})$$

and

$$\langle F(\psi), \alpha \rangle_{\tilde{G}} = \langle K_r(\psi, \cdot) \alpha, F \rangle_{\mathcal{H}(K_r)}, \forall \psi \in G, \alpha \in \tilde{G}, F \in \mathcal{H}(K_r) \quad (\text{S-10})$$

where $\langle \cdot, \cdot \rangle_{\tilde{G}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}(K_r)}$ are the inner-products endowed to the Hilbert space \tilde{G} and the RKHS $\mathcal{H}_{rep}(K)$, respectively.

S.2. PDF preprocessings for LQD and CLR transformations

S.2.1. PDF preprocessing for LQD transformation

Given a PDF $f(x)$ finitely supported on the compact interval $[0,1]$, its log quantile density (LQD) transformation (Petersen and Müller 2016) is defined as

$$\psi(t) = \log\left(\frac{dQ(t)}{dt}\right) = -\log\{f(Q(t))\} \quad (\text{S-11})$$

where $Q(t)$ is the quantile function associated with the PDF $f(x)$.

As pointed out in Subsection 3.1 of the manuscript, the functional data corresponding to the LQD node in the transformation tree are independent of the horizontal translations of curves in the PDF space. In other words, the LQD transformation is “blind” to the position shift of the PDF. Such

a property makes the LQD transformation to be a powerful tool for revealing the shape outliers masked by the “curve net” formed by the variability in horizontal positions of PDFs. However, for other applications such as the LQD-RKHS distributional regression (Chen et al. 2019a) involved in the regression outlier detection (Subsection 3.2) and the robust distributional regression (Section 4), effective measures should be taken to cure such a “blindness” issue.

On the other hand, if the PDF $f(x)$ takes near-zero values, computing the inverse function of the CDF $F(x)$ to obtain the quantile function (involved in the LQD transformation) might also suffer from a numerical issue in the interpolation process.

Fortunately, both the issues raised above can be easily addressed by performing the following preprocessing to the PDF $f(x)$ (Chen et al. 2019a)

$$f^*(x) = (1 - \alpha)f(x) + \alpha, \quad x \in [0, 1] \quad (\text{S-12})$$

where α is a prescribed small positive constant referred to as the PDF preprocessing parameter throughout this study. Note that the PDF $f(x)$ is finitely supported on the compact interval $[0, 1]$, thus the above processing is equivalent to mixing the original distribution by a proportion of uniform distribution $U(0, 1)$. The magnitude of the constant α depends on the specific applications: (1) for outlier detection, α should be smaller (e.g., $\alpha \in [10^{-10}, 10^{-2}]$), otherwise it might increase the variability of the functional data in the LQD node; (2) however, for the LQD-RKHS distributional regression (Chen et al. 2019a), the constant α should be larger (e.g., $\alpha \in [0.2, 0.5]$) as recommended by Chen et al. (2019a), otherwise the LQD transformation might be “blind” to the horizontal translation of PDFs.

For illustration purposes, we consider two PDFs (denoted as $f_1(x)$ and $f_2(x)$) respectively obtained by truncating the densities of norm distributions $N(0.4, 0.05^2)$ and $N(0.6, 0.05^2)$ within the domain of $[0, 1]$, the calculated LQD transformations associated with four different values of α (i.e., $\alpha = 10^{-10}, 10^{-2}, 0.2$ and 0.5) are displayed in Figure S-1. Comparing the LQD transformations shown in Figure S-1, one can see that the “blindness” phenomenon happens to the scenarios of $\alpha = 10^{-10}$ and $\alpha = 10^{-2}$, but disappears in the scenarios of $\alpha = 0.2$ and $\alpha = 0.5$ as the corresponding LQD transformations of the two PDFs can be clearly distinguished from each other.

For the distributional regression application, the regression is performed to the preprocessed PDF. Therefore, to recover the desired regression prediction associated with the target PDF $f(x)$, one should remember to clear the added uniform distribution through performing the following post-processing (Chen et al. 2019a):

$$\hat{f}(x) = \frac{1}{W} \left| \frac{\hat{f}^*(x) - \alpha}{1 - \alpha} \right| \quad \text{with} \quad W = \int_0^1 \left| \frac{\hat{f}^*(\tau) - \alpha}{1 - \alpha} \right| d\tau \quad (\text{S-13})$$

where $\hat{f}^*(x)$ stands for the prediction of $f^*(x)$ obtained by the regression model. For more detailed discussion, readers are referred to Chen et al. (2019a).

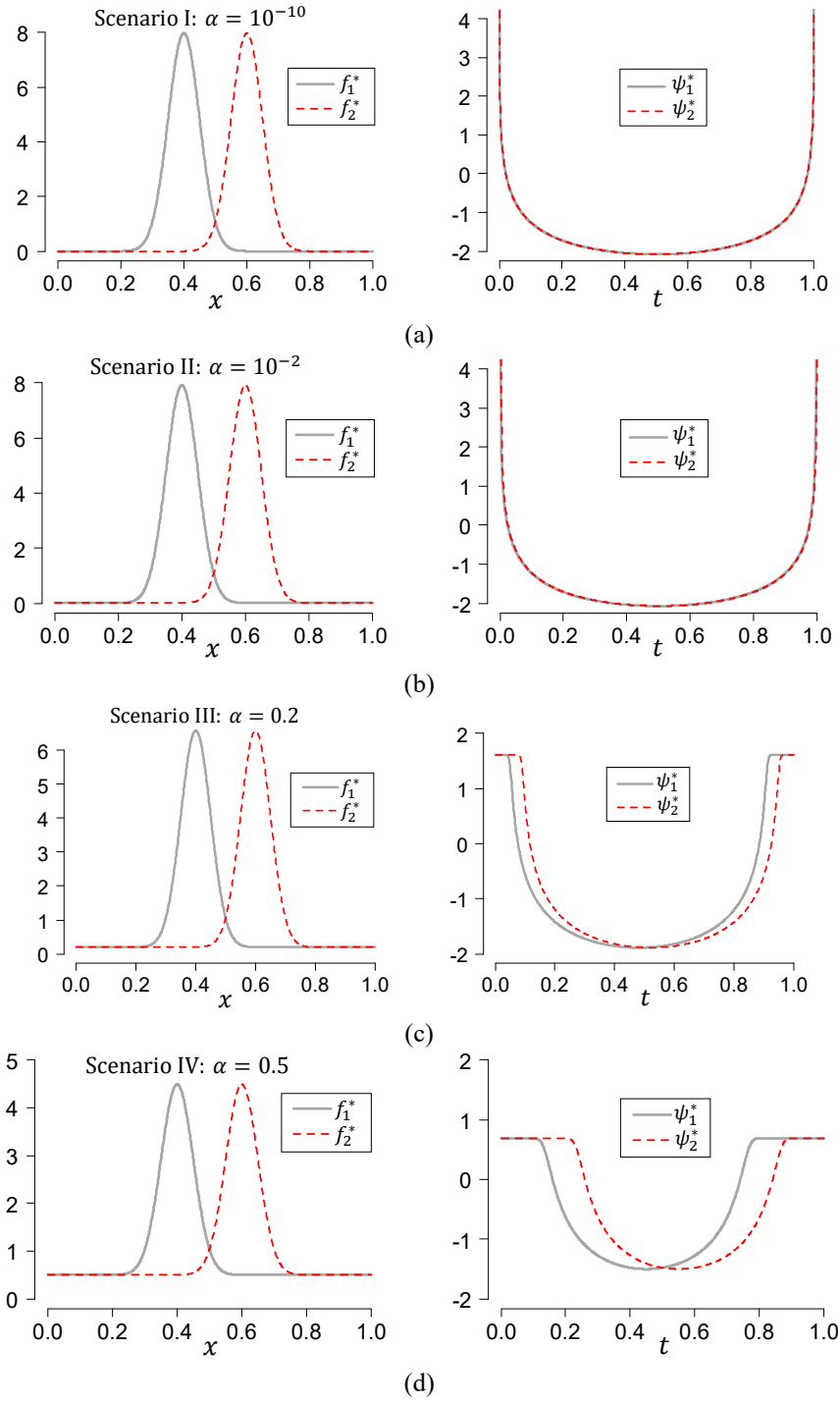


Figure S-1. Comparisons of the LQD transformations associated with two truncated normal densities after they are preprocessed by using Eq.(S-12) with α taking four different values. (a) Scenario I: $\alpha = 10^{-10}$, (b) Scenario II: $\alpha = 10^{-2}$, (c) Scenario III: $\alpha = 0.2$, and (d) Scenario IV: $\alpha = 0.5$. The left column corresponds to the PDFs, while the right column corresponds to the LQD transformations.

S.2.2. PDF preprocessing for CLR transformation

The CLR transformation given in Eq.(S-5) for a PDF taking near-zero values might suffer from a numerical issue in the logarithmic computation or have significant boundary effects, the latter is mainly attributed to the sharp change of the logarithmic function $\log(u)$ near $u = 0$. Therefore,

when calculating the CLR transformations associate with the investigated PDF-valued datasets $\{f_i(x)\}_{i=1}^n$, all the functional samples contained in $\{f_i(x)\}_{i=1}^n$ will be preprocessed in a unified manner as $f_i(x) = (1 - \alpha)f_i(x) + \alpha, i = 1, 2, \dots, n$, if the minimum value of the PDFs (i.e., $\min_{0 \leq i \leq 1} \inf_{x \in [0,1]} \{f_i(x)\}$) is less than 0.1. Unless otherwise stated, such a PDF-prepressing will be performed by default for the CLR transformations involved in this study, and the default value of the PDF preprocessing parameter α is 0.1.

S.3. Supplemental materials for distributional outlier detection

S.3.1. Modified boxplot-based detectors for scalar outlier detection

Let $\{\theta_i\}_{i=1}^n$ be a scalar dataset, the following two detectors modified from standard boxplot will be used to identify the potential outliers contained in $\{\theta_i\}_{i=1}^n$ according to specific situations:

(i) Detector I (two-sided detector):

$$\text{OUT}_{\text{ID}} = \{i \in \{1, 2, \dots, n\} | \theta_i < q_{0.25}(\theta) - r_1 \cdot \text{IQR} \text{ or } \theta_i > q_{0.75}(\theta) + r_1 \cdot \text{IQR}\} \quad (\text{S-14})$$

where OUT_{ID} stands for the index set of the detected outliers, $q_{0.25}(\theta)$ and $q_{0.75}(\theta)$ denote the 25th and 75th percentiles of the dataset $\{\theta_i\}_{i=1}^n$, respectively, IQR is the interquartile range defined as $\text{IQR} = q_{0.75}(\theta) - q_{0.25}(\theta)$, and r_1 is a user prescribed parameter.

(ii) Detector II (one-sided detector):

$$\text{OUT}_{\text{ID}} = \{i \in \{1, 2, \dots, n\} | \theta_i > q_{0.75}(\theta) + r_2 \cdot \text{IQR}\} \quad (\text{S-15})$$

Such a one-sided detector is designed specifically for the scenario that only the one taking abnormally large value can be regarded as the outlier, such as the case with the distance-based detection approach discussed in this study.

S.3.2. Supplemental materials for the single-dataset outlier detection method

S.3.2.1. Illustrations of the two basic transformations: derivative and centralization

As pointed out in Subsection 3.1 of the manuscript, there are generally two basic transformations that have good potential in exposing the abnormal curve patterns of functional data. The first one is performing a derivative to the functions, which helps to expose the curve with abnormal slope (Dai et al. 2020); see Figure S-2 for an illustration. The other one is centralization (i.e., shifting the curves along a direction to make them coincide with each other at a pre-specified feature point), which helps to peel away the masking effects caused by the position variability of the bulk of the curves (Dai et al. 2020); see Figure S-3 for an illustration. However, these two basic operations are not the “panacea” for all situations. For instance, performing the derivative operation to the simulated PDF-valued dataset shown in Figure S-4 cannot reveal the shape outliers, performing the centralization operation to a real PDF-valued dataset (consisting of 150 PDFs of strain measurements investigated in Chen et al. (2019a)) shown in Figure S-5 appears to reveal no

outliers. However, a latter investigation using the Tree-Distance detection method proposed in this study finds that this real dataset contains several shape outliers as shown in Figure S-6. In practical situations, the shape outliers contained in a real distributional dataset usually exhibit various patterns, which are generally difficult to be fully exposed by a single transformation. Moreover, according to our experience, conducting the transformations only in the PDF space also has limited effects in revealing the complicated shape outliers. Thus, it motivates us to propose a more sophisticated transformation system consisting of a collection of transformations (on the basis of these two basic operations) for exposing the complicated distributional shape outliers in different spaces.

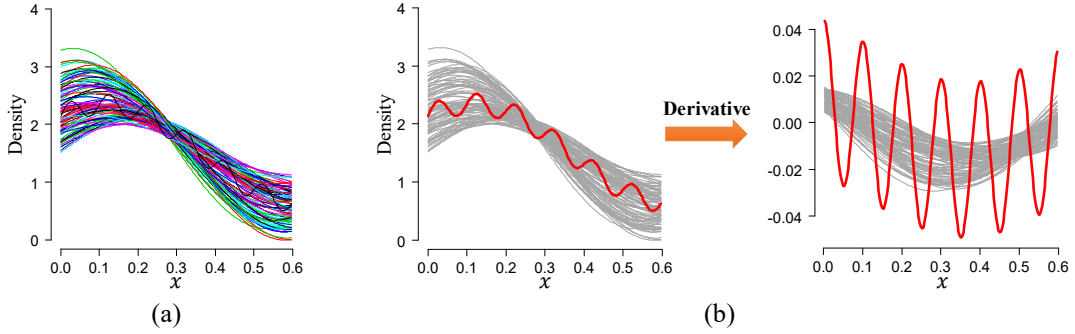


Figure S-2. Illustration of the derivative operation in exposing shape outliers. (a) The original PDF-valued dataset and (b) visualization of the PDFs before and after the derivative operation.

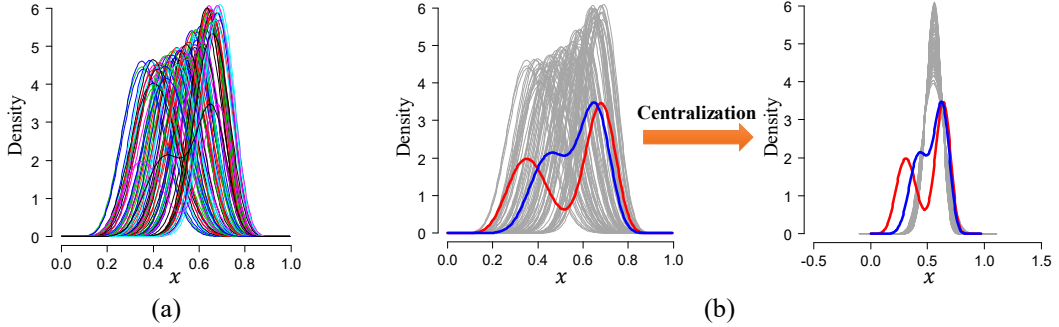


Figure S-3. Illustration of the centralization operation in exposing shape outliers. (a) The original PDF-valued dataset and (b) visualization of the PDFs before and after the centralization operation. The median is selected as the feature point for horizontal alignment.

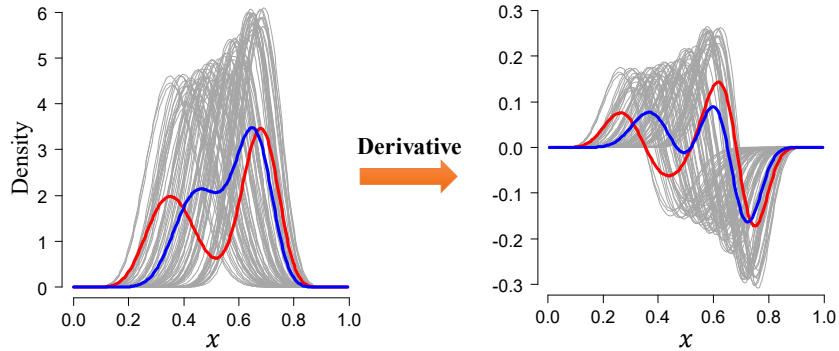


Figure S-4. Visualization of the PDFs of a simulated dataset before and after the derivative operation.

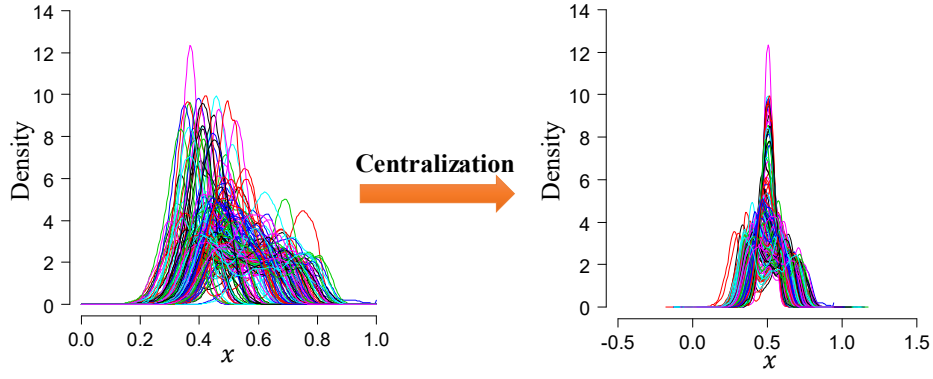


Figure S-5. Visualization of the PDFs of a real dataset before and after the centralization operation.

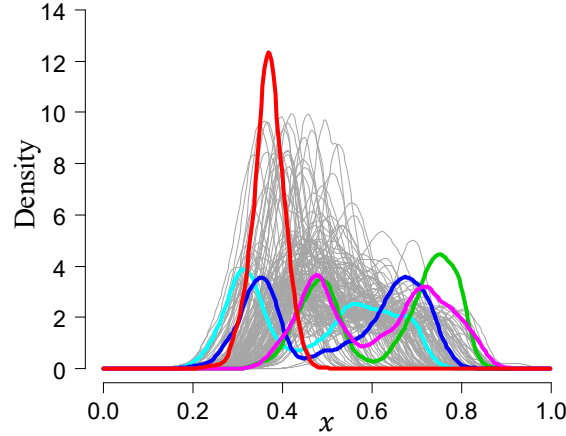


Figure S-6. Detected outlying PDFs (represented by bold colored curves) in the real dataset (shown in the left panel of Figure S-5) using our proposed method described in Subsection 3.1 in the manuscript (i.e., the Tree-Distance method). In the outlier detection, the default argument settings listed later in Table S-1 are adopted.

S.3.2.2. Discussion, justification and comparison for the central function selection

This subsection provides relevant illustrations and justifications for the selected central function (i.e., $m(t)$ in the Eq. (1) of the manuscript) involved in the distance-based functional outlier detection procedure described in Subsection 3.1 of the manuscript. Such a central function serves as the reference function in quantifying the degrees of outlyingness for the functional samples using a user-specified distance. Generally, the reference function is expected to satisfy the following basic requirements: (1) it should locate at the central region of the majority of the functional data; (2) it should be insensitive to the functional outliers presented in the dataset and (3) if the functional data has inherent constraints, the selected central function is also expected to satisfy the constraints.

The resulting transformed functional data associated with the transformation tree can be divided into two categories, namely, the ordinary functional data free from constraints (e.g., the LQD-transformed data) and special functional data with additional constraints (e.g., the CLR-transformed data). Actually, except for the CLR-transformed data, the other functional data associated with the leaf nodes of the tree (i.e., nodes nLQD and DIFF) are all ordinary functional data without constraints.

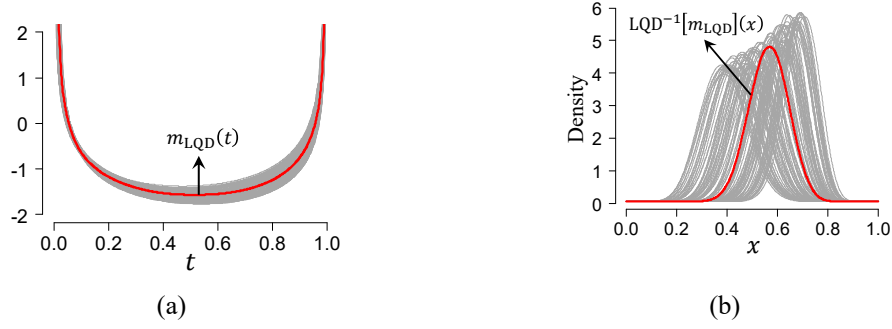


Figure S-7. Visualization of the central function $m_{\text{LQD}}(t)$ computed using the LQD-transformed data (left panel) along with its image $\text{LQD}^{-1}[m_{\text{LQD}}](x)$ in the PDF space (right panel). The light gray curves represent the LQD-transformed data or the PDFs, while the bold red curve represents the computed central function. $\text{LQD}^{-1}[m_{\text{LQD}}](x)$ denotes the image of $m_{\text{LQD}}(t)$ obtained by mapping it into the PDF space using the inverse LQD transformation.

For the ordinary functional data, the central function is selected as the cross-sectional median of the curves. Take the LQD-transformed data (denoted as $\mathcal{D} = \{\psi_i(t)\}_{i=1}^n$) as an example, the central function is computed as

$$m_{\text{LQD}}(t) = \text{median}\{\psi_i(t)\}, t \in [0,1] \quad (\text{S-16})$$

The visualization of the computed $m_{\text{LQD}}(t)$ using a simulated dataset is presented in Figure S-7 (a) as the bold red curve. Such a central function can be mapped into the PDF space by using the inverse LQD transformation (Petersen and Müller 2016) defined as follows:

$$f(x) = \text{LQD}^{-1}[\psi](x) = \theta_\psi \exp\{-\psi(F(x))\} \quad (\text{S-17})$$

with $F^{-1}(t) = \theta_\psi^{-1} \int_0^t e^{\psi(s)} ds$ and $\theta_\psi = \int_0^1 e^{\psi(s)} ds$

Consequently, the inverse LQD transformation of the computed central function, denoted as $\text{LQD}^{-1}[m_{\text{LQD}}](x)$, is a density function as shown in Figure S-7 (b). Obviously, the resulting density $\text{LQD}^{-1}[m_{\text{LQD}}](x)$ locates near the center of the majority of the PDFs.

For the CLR-transformed data, denoted as $\mathcal{D} = \{\phi_i(x)\}_{i=1}^n$ with $\phi_i = \text{CLR}[f_i]$ being the CLR transformation of the density f_i , they are special functions subject to a constraint of integrating to zero (Hron et al. 2016), i.e.,

$$\int \phi_i(\tau) d\tau = 0, \quad i = 1, 2, \dots, n \quad (\text{S-18})$$

To ensure the selected central function can satisfy this constraint, the one in the CLR-transformed dataset $\mathcal{D} = \{\phi_i(x)\}_{i=1}^n$ that is closest to the cross-sectional median function is chosen as the central function, i.e.,

$$m_{\text{CLR}}(x) = \underset{\phi \in \mathcal{D}}{\text{argmin}} d_{L^2}(\phi, \phi_m) \quad (\text{S-19})$$

where $\phi_m(x) = \text{median}\{\phi_i(x)\}_{1 \leq i \leq n}$ is the cross-sectional median function of the functional data in \mathcal{D} , and d_{L^2} stands for the L^2 distance defined as $d_{L^2}(\phi_1, \phi_2) = (\int (\phi_1(\tau) - \phi_2(\tau))^2 d\tau)^{1/2}$. The

computed central function for a given CLR-transformed dataset using this principle is shown in Figure S-8(a) as the bold red curve. Such a central function can also be mapped into the PDF space by using the inverse CLR transformation (Machalova et al. 2016) defined as follows:

$$f(x) = \text{CLR}^{-1}[\phi](x) = \frac{\exp\{\phi(x)\}}{\int \exp\{\phi(\tau)\}d\tau} \quad (\text{S-20})$$

Consequently, the inverse CLR transformation of the computed central function, denoted as $\text{CLR}^{-1}[m_{\text{CLR}}](x)$, is also a density function as shown in Figure S-8 (b). One can see that the density $\text{CLR}^{-1}[m_{\text{CLR}}](x)$ also locates near the center of the majority of the PDFs. For comparison, $\text{CLR}^{-1}[m_{\text{CLR}}](x)$ and $\text{LQD}^{-1}[m_{\text{LQD}}](x)$ (computed earlier using the LQD-transformed data) are plotted in the same plot as shown in Figure S-9, showing that they are close with each other in the PDF space.

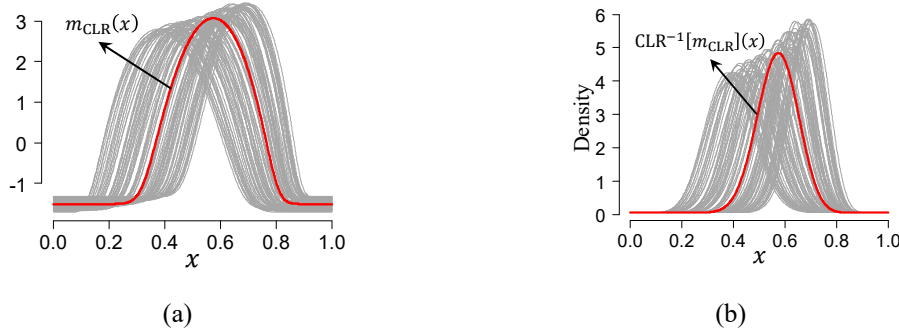


Figure S-8. Visualization of the central function $m_{\text{CLR}}(x)$ computed using the CLR-transformed data (left panel) along with its image $\text{CLR}^{-1}[m_{\text{CLR}}](x)$ in the PDF space (right panel). The light gray curves represent the CLR-transformed data or the PDFs, while the bold red curve represents the computed central function. $\text{CLR}^{-1}[m_{\text{CLR}}](x)$ denotes the image of $m_{\text{CLR}}(x)$ obtained by mapping it into the PDF space using the inverse CLR transformation.

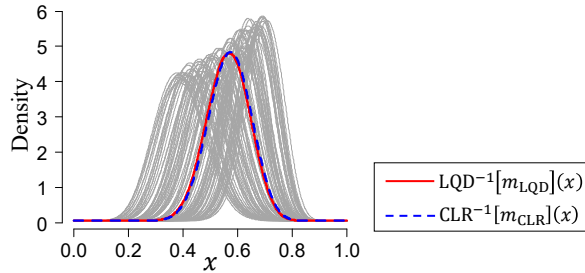


Figure S-9. Comparison of the computed central functions (in the PDF space) using the LQD- and CLR-transformed data, respectively.

In order to investigate the sensitivity of the computed central functions to outlying PDFs, we add ten outlying PDFs to the PDF-valued dataset and recalculate the central functions using the same procedure described above. The contaminated PDF-valued dataset is shown in Figure S-10 (a) with the outlying PDFs represented by bold pink curves. The new computed central functions using the corresponding contaminated LQD- and CLR-transformed data are denoted as $m_{\text{LQD}}^{\#}(t)$ and $m_{\text{CLR}}^{\#}(x)$, respectively. After mapping $m_{\text{LQD}}^{\#}(t)$ (or $m_{\text{CLR}}^{\#}(x)$) into the PDF space using the

inverse LQD (or CLR) transformation, the resulting PDF, denoted as $\text{LQD}^{-1}[m_{\text{LQD}}^{\#}](x)$ (or $\text{CLR}^{-1}[m_{\text{CLR}}^{\#}](x)$), is presented in Figure S-10 (b) (or Figure S-10 (c)) as a blue dashed line. For comparison, the result calculated earlier using the “good” data is also added to the plot as a red solid line. One can see that the computed central functions are insensitive to the outlying PDFs.

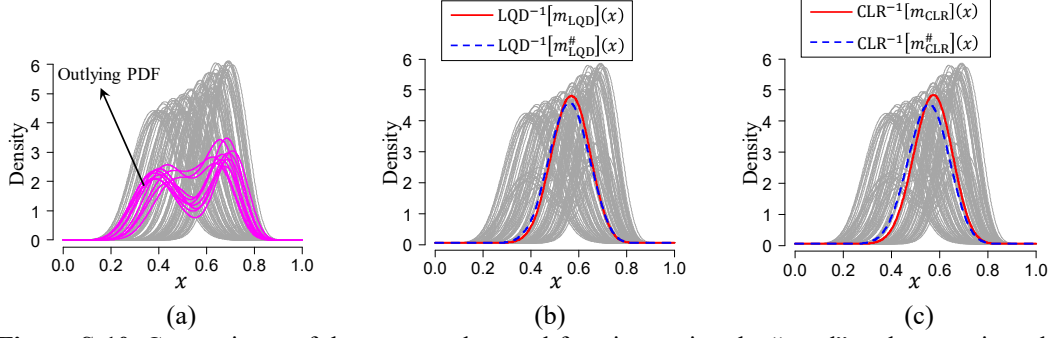


Figure S-10. Comparisons of the computed central functions using the “good” and contaminated data, respectively. (a) Visualization of the contaminated PDF-valued dataset with the pink curves representing the outlying PDFs; (b) Comparison of the computed central functions (in the PDF space) obtained using the LQD-transformed data with and without outliers; (c) Comparison of the computed central functions (in the PDF space) obtained using the CLR-transformed data with and without outliers. The red solid curves in (b) and (c) represent the computed central functions using the “good” data, while the blue dashed curves represent the computed central functions using the contaminated data.

The central function selection procedure described above has advantages in the following aspects: (1) simple and highly efficient; (2) the result can satisfy the constraints possessed by the associated functional data; (3) insensitive to functional outliers. To further demonstrate its advantages, here we consider an alternative central function selection method using the band depth. The band depth is one of the most popular functional depth widely used for ranking functions from center to outward (López-Pintado and Romo 2009). The functional sample possessing the largest computed depth value is defined as the deepest curve in the functional dataset, which can be regarded as the central function. Figure S-11 presents the found central functions of the “good” (left panel) and contaminated (right panel) PDF-valued datasets based on the ranked PDFs using their computed band depths. Unfortunately, in the contaminated dataset, the found central function is an outlying PDF; obviously, such outlying PDF cannot represent the center of the PDFs. Such a failure case provides further evidence that indiscriminately applying the statistical tools developed for ordinary functional data to the PDFs with special constraints can usually lead to misleading results. On the other hand, such a band depth-based approach is also computationally intensive; thus, it is not used in our proposal.

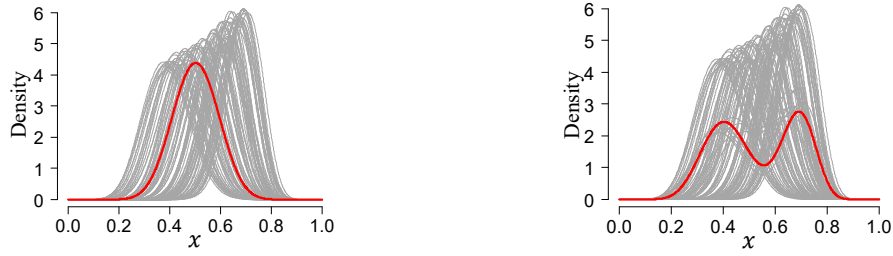


Figure S-11. Visualization of the central function (bold red curve) obtained by ranking the PDFs using the band depth. (a) Computed using the “good” data and (b) computed using the contaminated data.

S.3.2.3. Discussion, justification and comparison for distance (metric) selection

We recommend to perform the distance-based outlier detections for the functional data associated with the nLQD, CLR and DIFF nodes on the transformation tree shown in Figure. 2(b) of the manuscript. Moreover, the outlier detection is also required to be conducted to the medians residing in the MED node, which plays the key role in identifying the horizontal shift outliers. In contrast to the functional data in nodes nLQD, CLR and DIFF, the data in the MED node are scalar data; thus, the corresponding outliers can be directly detected by using the two-sided boxplot detector given in Eq. (S-14). However, for the functional data associated with the nLQD, CLR and DIFF nodes, appropriate distances are required to be selected for converting the functional outlier detection problem into the scalar outlier detection problem by using Eq. (1) in the manuscript. The main idea of the distance-based outlier detection strategy is embedding the functional data into a specific metric space, then using the associated metric (i.e., distance) to perform outlier detection. Thus, the choice of the distance for outlier detection is mainly determined by the metric space into which the investigated functional data can be embedded. In the following, we first discuss how to select the distances for the functional data associated with nodes nLQD and DIFF, as the corresponding data in both nodes are ordinary functional data. For the CLR node, the corresponding data are special functional data and we consider the isometric isomorphism (between the Bayes space and the space where the CLR-transformed data reside) to perform outlier detection; thus, the distance selection for this node will be provided later on.

In this study, the processed functional data in nodes nLQD and DIFF for outlier detection are both ordinary functional data that can be embedded into the $L^2(A)$ space, where A stands for the detection interval given in Eq. (1) of the manuscript. The $L^2(A)$ space is a functional space formed by square integrable real functions on A , and it is a metric space endowed with the L^2 distance defined as

$$d_{L^2}(\phi_1, \phi_2) = \left(\int_A (\phi_1(\tau) - \phi_2(\tau))^2 d\tau \right)^{1/2}, \forall \phi_1, \phi_2 \in L^2(A)$$

On the other hand, since the detection interval A has a finite Lebesgue measure, we have $L^2(A) \subseteq L^1(A)$ (Royden and Fitzpatrick 2010, p.142), where $L^1(A)$ denotes another metric space formed by integrable real functions on A and endowed with the L^1 distance defined as

$$d_{L^1}(\phi_1, \phi_2) = \int_A |\phi_1(\tau) - \phi_2(\tau)| d\tau, \forall \phi_1, \phi_2 \in L^1(A)$$

Consequently, the functional data in nodes nLQD and DIFF can also be regarded as elements of the $L^1(A)$ space. Moreover, the considered PDFs in this study are all smooth continuous functions; thus, the resulting transformed functional data in nodes nLQD and DIFF are continuous functions. Consequently, the data can also be regarded as elements of the metric space $C(A)$ (formed by continuous real functions on A) endowed with the sup distance defined as

$$d_{sup}(\phi_1, \phi_2) = \sup_{\tau \in A} |\phi_1(\tau) - \phi_2(\tau)|, \forall \phi_1, \phi_2 \in C(A)$$

In this sense, the L^2 distance, L^1 distance and sup distance are all valid for performing outlier detection for the functional data in nodes nLQD and DIFF. Generally, the L^2 and L^1 distances are mainly for quantifying the global dissimilarities of the functional data, while the sup distance is mainly for quantifying the local dissimilarities of the functional data (i.e., the deviation of the outermost point (with respect to the reference curve) of a curve).

In practical applications, both the global and local dissimilarity measures have their own advantages and shortcomings in functional outlier detection. For instance, when the outlying curve only significantly deviates from the majority of the data within a local region (such as the case shown in the left panel of Figure S-12 (a)), performing outlier detection using the L^2 (or L^1) distance may yield a disappointed result, namely, the outlying curve cannot be manifested as an outlier in the calculated distances as show in Figure S-12 (a) (or Figure S-12 (b)); however, if we use the sup distance, a much more satisfactory result can be obtained (see Figure S-12 (c)), and the outlying curve can be successfully detected. But, in some situations, the sup distance may also become less powerful than the L^2 or L^1 distance, one representative example is illustrated in Figure S-13, where the sup distance performs poorly and only the most outlying curve (i.e., curve OC_3) has been detected. One can see from Figure S-13 (c) that the calculated sup distances associated with the other three outlying curves (i.e., OC_1 , OC_2 , OC_4) cannot be isolated from several “good” data. To gain insight for such phenomenon, Figure S-14 provides a schematic illustration of the sup distances (from the reference function) for one “good” curve and one outlying curve. By comparing the results in Figure S-14, one can easily understand why the sup distance fails to isolate the three outlying curves from some nonoutlying data. However, if we use the L^2 (or L^1) distance to perform outlier detection, the four outlying curves can be successfully detected as shown in Figure S-13 (a) (or Figure S-13 (b)). We can therefore, conclude that the global and local dissimilarity measures have complementarity in functional outlier detection and they should be used together to enhance the performance of outlier detection. Generally, for global dissimilarity quantification, the L^2 distance and L^1 distance have similar performances. The main simulation studies conducted later in Subsection S.5.1 show that the L^1 distance performs slightly better than the L^2 distance. 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.

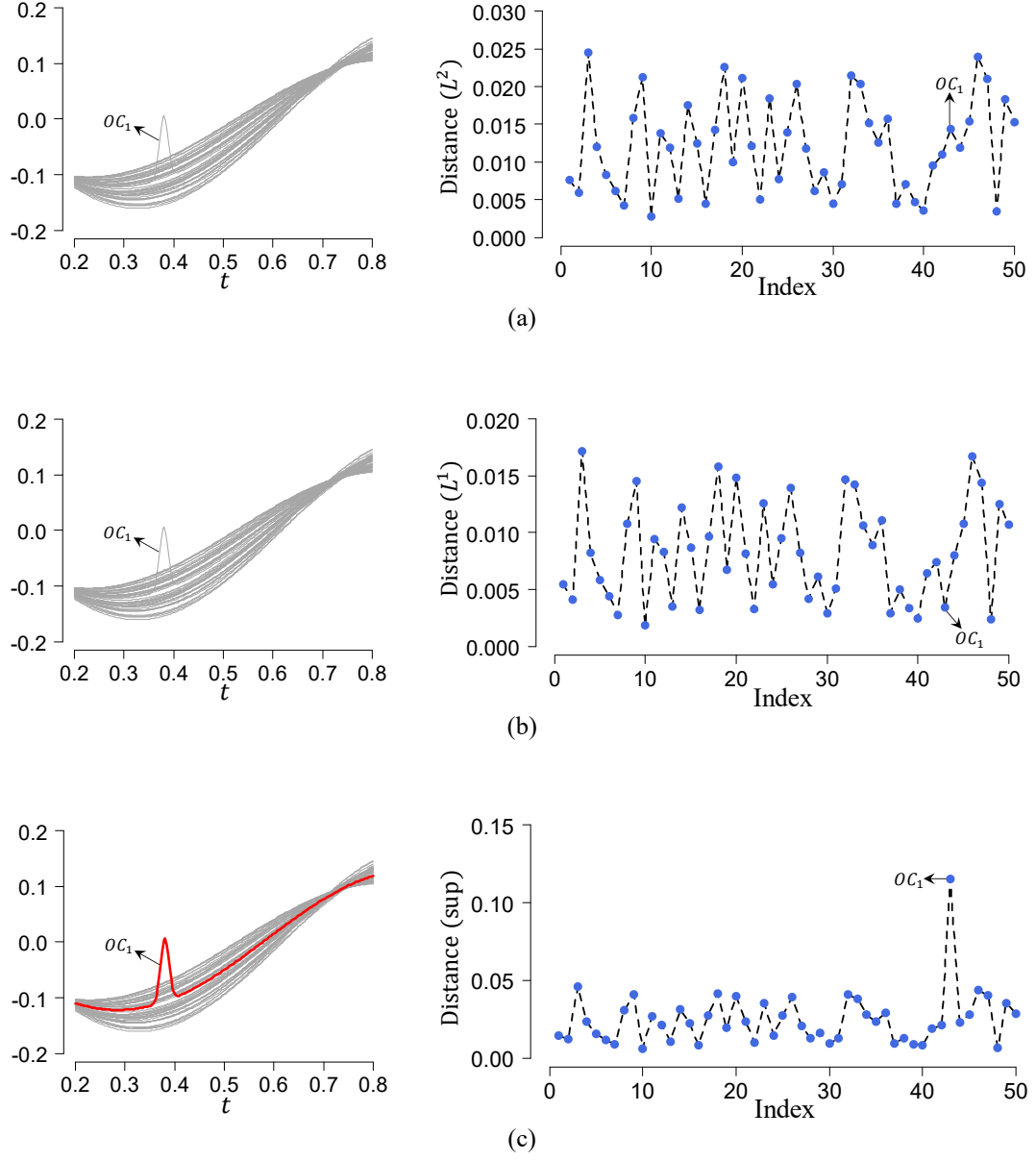


Figure S-12. Comparisons of the functional outlier detection using the (a) L^2 distance, (b) L^1 distance and (c) sup distance for a dataset with one outlying curve OC_1 . The left column corresponds to the detection result (the bold colored curve represents the detected outlier), while the right column corresponds to the calculated distances. The one-sided detector given in Eq.(S-15) is used for detecting the outliers in the dataset of the calculated distances (real-valued data), and the whisker parameter r_2 is set to 2.5.

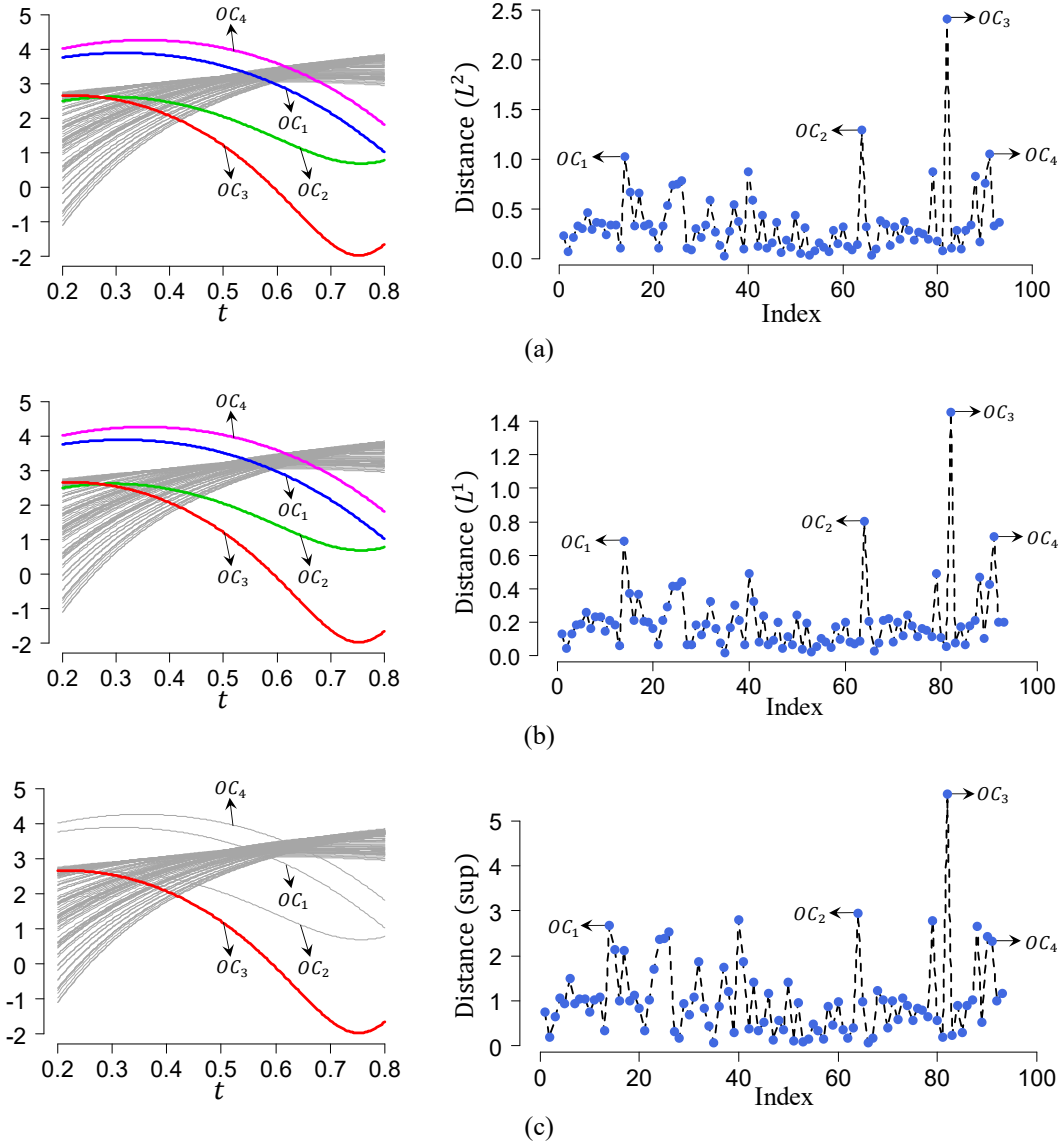


Figure S-13. Comparisons of the functional outlier detection using the (a) L^2 distance, (b) L^1 distance and (c) sup distance for a dataset with four outlying curves OC_1 , OC_2 , OC_3 and OC_4 . The left column corresponds to the detection result (the bold colored curves represent the detected outliers), while the right column corresponds to the calculated distances. The one-sided detector given in Eq.(S-15) is used for detecting the outliers in the dataset of the calculated distances (real-valued data), and the whisker parameter r_2 is set to 2.5.

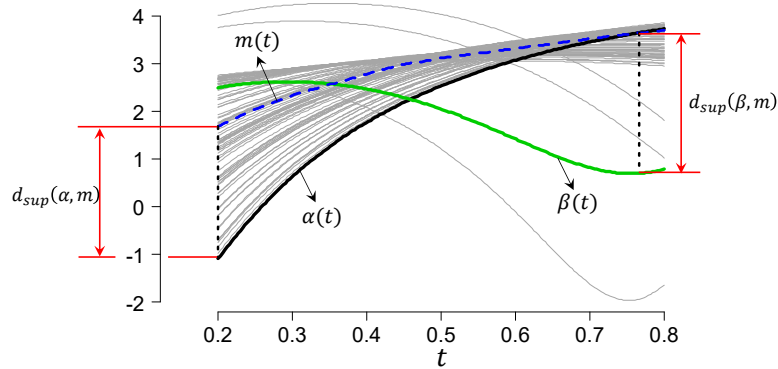


Figure S-14. Schematic illustration of the sup distances (from the reference function $m(t)$) for one "good" curve $\alpha(t)$ and one outlying curve $\beta(t)$ (corresponding to curve OC_2 in Figure S-13).

Remark 1. One may argue that the three outlying curves that failed to be detected by using the sup distance shown in Figure S-13 (c) can be easily detected by a pointwise detection strategy as follows: (1) calculate the following pointwise distances (with the reference function) over a given discretized grid for each curve:

$$\Delta_i(t) = |\beta_i(t) - m(t)|, t \in \{t_1, t_2, \dots, t_T\}, i = 1, 2, \dots, n$$

where $m(t)$ is the pointwise median function served as the reference function, $\{t_1, t_2, \dots, t_T\}$ is the user-specified discretized grid, $\Delta_i(t)$ is called the pointwise distance from the curve $\beta_i(t)$ to the reference function $m(t)$ at the point t ; (2) independently implement an outlier detection to the dataset $\{\Delta_i(t)\}_{i=1}^n$ at each grid point using the one-sided boxplot detector given in Eq.(S-15); (3) if a curve is detected as an outlier at least one grid point, it is regarded as an outlying curve. Of course, such a pointwise detection approach can effectively identify all of the four outlying curves shown in Figure S-13 (c); however, it may also bring another undesirable issue, i.e., high risk of false detection. To illustrate this, a toy detection example using such a detection strategy is presented in Figure S-15, where the two bold colored curves (ought to be regarded as “good” curves) are detected as outliers. We can therefore, conclude that such a pointwise detection approach is too sensitive in functional outlier detection, leading to much higher risk of false detection. Hence, for local anomaly, performing outlier detection using the recommended sup distance is a preferable compromise between the highly sensitive pointwise detection approach and the less sensitive detection approach using the L^2 or L^1 distance.

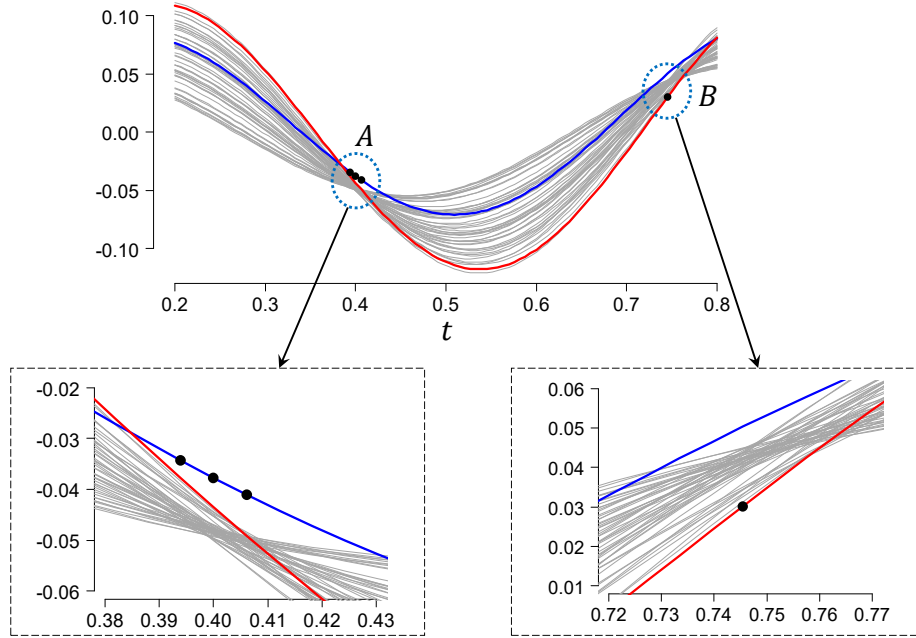


Figure S-15. A toy detection example using the pointwise detection strategy described in Remark 1. The bold colored curves represent the detected outlying curves, while the solid circles represent the detected outliers at the grid points t_j s. The one-sided detector given in Eq.(S-15) is used for detecting the outliers in the dataset of the calculated pointwise distances, and the whisker parameter r_2 is set to 2.5.

As pointed out in Section 3.1 of the manuscript that the role of Branch II in the transformation tree is for embedding the PDFs into the Bayes space for outlier detection. In other words, the PDFs

are treated as elements of the Bayes space; thus, the outlier detection is performed in their own space of the PDFs. The Bayes space itself is a metric space endowed with the distance $d_{\mathfrak{B}}(\tilde{f}, \tilde{g})$ defined in Eq. (S-4). Moreover, the Bayes space is isometrically isomorphic to the L^2 space with the CLR transformation as the isometric isomorphism. Consequently, it has

$$d_{\mathfrak{B}}(\tilde{f}, \tilde{g}) = d_{L^2}(\text{CLR}[\tilde{f}], \text{CLR}[\tilde{g}])$$

Therefore, for the functional data associated with the CLR node, we use L^2 distance to perform outlier detection, so as to be consistent with the original intention that using the $d_{\mathfrak{B}}$ distance to perform outlier detection in the Bayes space.

S.3.2.4. Default settings for the Tree-Distance outlier detection method

With the above recommended distances, the default argument settings of outlier detections for the resulting transformed data associated with nodes nLQD, CLR, DIFF and MED are summarized in Table S-1. The functional data in the nLQD node are computed from the quantile functions, all the functional data after the QF node (in the Branch I of the transformation tree) have been naturally aligned according to the quantiles; thus, one can choose appropriate detection intervals to conduct the curve truncation as shown in Figure S-16 (c). The main reason for performing such a truncation is twofold: (1) reduce the boundary effects in disturbing the outlier detection; (2) restrict the outlier detection within the region of interested. According to our experience, we recommend to independently perform two rounds outlier detections for the functional data in the nLQD node respectively using the detection regions $[0.2, 0.8]$ and $[0.4, 0.6]$ for both the L^1 and sup distances, and then the detected outliers are merged to form the final detection result of the nLQD node.

Table S-1

Default settings of outlier detections using the distance-based method for the transformed data associated with nodes nLQD, CLR, DIFF, MED on the transformation tree. The parameter α in the second column is the PDF preprocessing parameter described in Section S.2. The detection region $[u, v]$ associated with the CLR node is the common support of the translated PDFs in the horizontal centralization processing (performed in the H-CENTR node) described in Subsection 3.1 of the manuscript. Detector I and Detector II stand for the two- and one-sided boxplot-based detectors given in Eq. (S-14) and Eq. (S-15), respectively.

Node	α	Distance	Detector	Whisker	Detection region
nLQD	10^{-10}	L^1 and sup	Detector II	L^1 : 2.5IQR sup: 3.5IQR	$[0.2, 0.8]$ and $[0.4, 0.6]$
CLR	0.1	L^2	Detector II	L^2 : 2.5IQR	$[u, v]$
DIFF	—	L^1 and sup	Detector II	L^1 : 2.5IQR sup: 3.5IQR	$[0, 1]$
MED	—	—	Detector I	1.5IQR	—

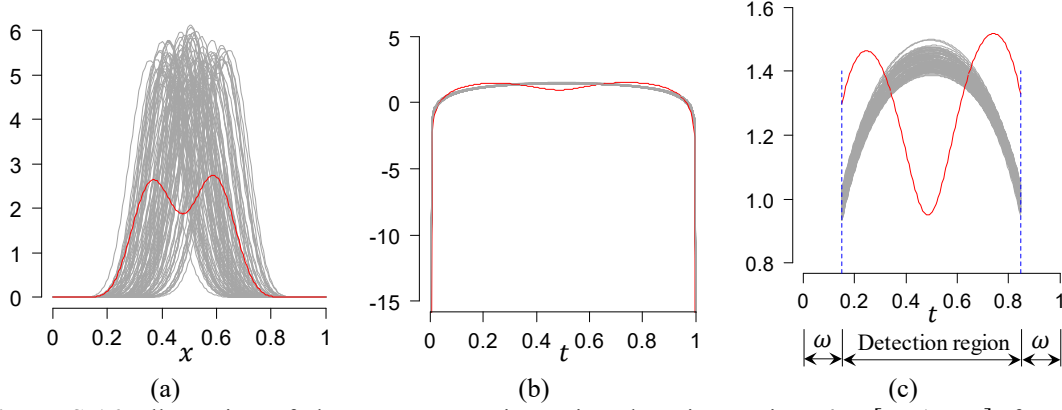


Figure S-16. Illustration of the curve truncation using detection region $A = [\omega, 1 - \omega]$ for the normalized LQD transformations associated with a PDF-valued dataset. (a) The raw PDF-valued dataset, (b) the corresponding normalized LQD transformations before curve truncation, and (c) the corresponding normalized LQD transformations after curve truncation. The gray lines and the red line represent the “good” data and the outlying curve, respectively.

S.3.3. Supplemental materials for the abnormal association detection method

S.3.3.1. Illustrations of representative abnormal PDF-pairs

This subsection provides illustrations for some representative outlying PDF-valued two-tuples. Consider n PDF-valued two-tuples, denoted as $\mathcal{T} = \{g_i(x), f_i(x)\}_{i=1}^n$, formed by elements from two correlated PDF-valued datasets $\{g_i(x)\}_{i=1}^n$ and $\{f_i(x)\}_{i=1}^n$. Representative examples of such correlated PDF-valued datasets are shown in Figure S-17 (or Figure S-18), where the PDFs from the same two-tuple are represented by the curves in the same color. We say that the association of a PDF-valued two-tuple is abnormal if it significantly violates the dependence pattern followed by the majority of the data. A PDF pair with an abnormal association can behave as either abnormal or normal in their respective datasets. Figure S-17 illustrates a representative abnormal PDF-valued two-tuple denoted as $\{g_\beta, f_\beta\}$, where g_β is a “good” curve in the dataset of $\{g_i\}_{i=1}^n$ while f_β is an outlying curve in the dataset of $\{f_i\}_{i=1}^n$. Figure S-18 illustrates another representative abnormal PDF-valued two-tuple denoted as $\{g_\alpha, f_\alpha\}$, where both g_α and f_α are “good” curves in their respective datasets.

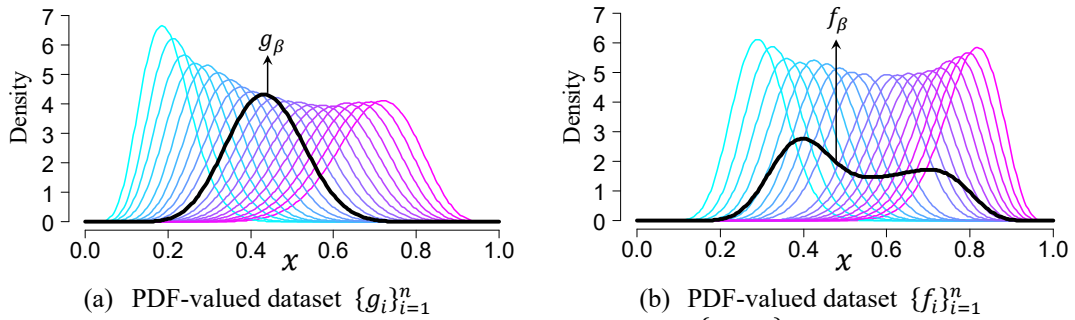


Figure S-17. Visualization of an outlying PDF-valued two-tuple $\{g_\beta, f_\beta\}$ (represented by bold black curves) with f_β being a functional outlier in its own dataset. The PDFs from the same two-tuple are in the same color.

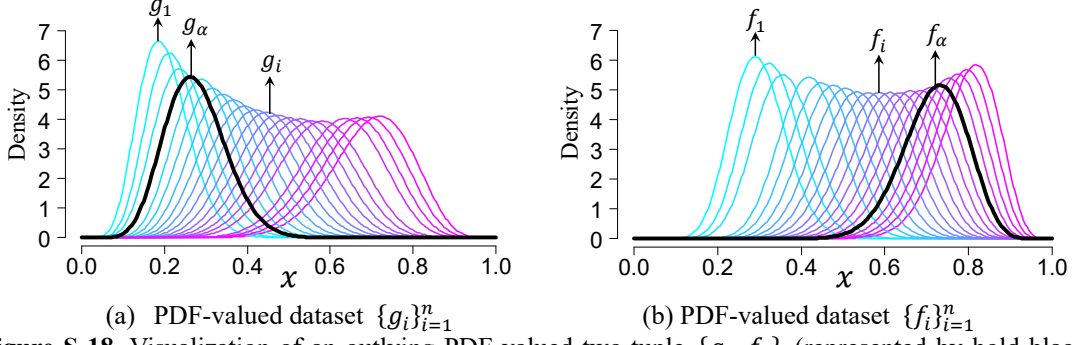


Figure S-18. Visualization of an outlying PDF-valued two-tuple $\{g_\alpha, f_\alpha\}$ (represented by bold black curves) with g_α and f_α both normally behaved in their own datasets. The PDFs from the same two-tuple are in the same color.

For the case shown in Figure S-17, the outlying PDF-pair $\{g_\beta, f_\beta\}$ can be detected by performing an outlier detection to the dataset $\{f_i(x)\}_{i=1}^n$ using an appropriate single-dataset distributional outlier detection method; however, for the case shown in Figure S-18, single-dataset outlier detection methods are no longer suitable for detecting the outlying PDF-pair $\{g_\alpha, f_\alpha\}$.

S.3.3.2. Residual calculation using the CLR transformation

The regression error is inevitable, which means the fitted PDF (obtained by the regression model) may deviate from the target PDF. Note that the regression outlier detection is based on residual outlier detection. The horizontal deviation of the predicted PDFs usually will lead to large residual for the PDFs, especially for the “slim” PDF (will be illustrated later). Such a phenomenon can significantly increase the risk of false detection. In this study, we adopt a median alignment strategy to remedy this issue. Specifically, let f_i denote the target PDF defined on the compact interval $[0,1]$, and let $\text{med}(f_i)$ denote the median defined as

$$\text{med}(f_i) = \inf \left(\left\{ x \in [0,1]: \int_0^x f_i(t) dt \geq \frac{1}{2} \right\} \right) \quad (\text{S-21})$$

Moreover, let \hat{f}_i denote the fitted result of f_i obtained by the distributional regression model. If f_i and \hat{f}_i are close enough in the horizontal direction (i.e., $|\text{med}(f_i) - \text{med}(\hat{f}_i)| \leq \theta_h$ with θ_h being a pre-specified threshold), the residual of f_i (with respect to \hat{f}_i) will be calculated after horizontally translating f_i to \hat{f}_i to make their median points coincide with each other as illustrated in Figure S-19, where $f_i^\#$ stands for the result of f_i after movement. Based on such a median alignment strategy, the implementation of PDF-residual calculation using the Bayes distance is outlined in Algorithm S.1.

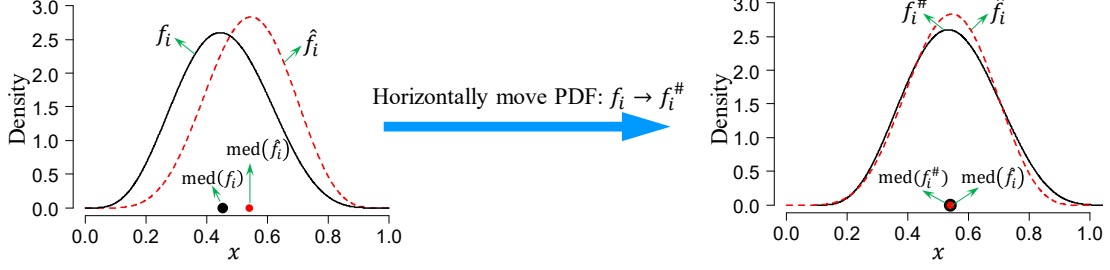


Figure S-19. Illustration of median alignment for two PDFs.

Algorithm S.1: Calculation for $\varepsilon_i^{\mathfrak{B}} = d_{\mathfrak{B}}(\hat{f}_i, f_i | \theta_h, \alpha_{mix}^{\mathfrak{B}})$

Input: PDF pairs (\hat{f}_i, f_i) , threshold θ_h , PDF preprocessing parameter $\alpha_{mix}^{\mathfrak{B}}$

Output: Residual $\varepsilon_i^{\mathfrak{B}}$

1: Calculate the medians $\text{med}(f_i)$ and $\text{med}(\hat{f}_i)$

2: **If** $|\text{med}(f_i) - \text{med}(\hat{f}_i)| \leq \theta_h$ **then**

a: Horizontally translate f_i to \hat{f}_i to make their median points coincide with each other, denote the translated PDF as $f_i^{\#}$

b: Find the common support of $f_i^{\#}$ and \hat{f}_i , denote the common support as $[c_1, c_2]$

c: Set $f_i^{\#*}(x) = (1 - \alpha_{mix}^{\mathfrak{B}})f_i^{\#}(x)\chi_{[c_1, c_2]}(x) + \alpha_{mix}^{\mathfrak{B}}$

$\hat{f}_i^*(x) = (1 - \alpha_{mix}^{\mathfrak{B}})\hat{f}_i(x)\chi_{[c_1, c_2]}(x) + \alpha_{mix}^{\mathfrak{B}}$

where $\chi_{[c_1, c_2]}(\cdot)$ denote the indicator function

d: Compute $\varepsilon_i^{\mathfrak{B}} = d_{\mathfrak{B}}(f_i^{\#*}, \hat{f}_i^*) = d_{L^2}(\text{CLR}[f_i^{\#*}], \text{CLR}[\hat{f}_i^*])$

else

a: Set $f_i^*(x) = (1 - \alpha_{mix}^{\mathfrak{B}})f_i(x) + \alpha_{mix}^{\mathfrak{B}}$, $\hat{f}_i^*(x) = (1 - \alpha_{mix}^{\mathfrak{B}})\hat{f}_i(x) + \alpha_{mix}^{\mathfrak{B}}$

b: Compute $\varepsilon_i^{\mathfrak{B}} = d_{\mathfrak{B}}(f_i^*, \hat{f}_i^*) = d_{L^2}(\text{CLR}[f_i^*], \text{CLR}[\hat{f}_i^*])$

end if

3: Output $\varepsilon_i^{\mathfrak{B}}$

In the following, we use an example to illustrate the negative effects (in regression outlier detection) caused by the horizontal-shift error, as well as to validate the effectiveness of the recommended remedy. We select 50 PDF-valued samples (corresponding to the response variable) in a regression outlier detection test for demonstration, of which 3 PDFs indexed by $i = 17, 39$ and 42 are synthetic abnormal PDFs with their shapes significantly differing from the majority of the data to serve as the regression outliers. After fitting the regression model, we first calculate the residuals for the PDFs without considering the treatment of the median alignment. This can be easily achieved by setting $\theta_h = 0$ in Algorithm S.1 (the PDF preprocessing parameter $\alpha_{mix}^{\mathfrak{B}}$ is set to 0.1) and the results are shown in Figure S-20. The comparison of the fitted results for eight selected PDF-valued samples (marked in Figure S-20 by $i = 6, 17, 18, 22, 27, 39, 42$ and 49) are visualized in Figure S-21. Except the three outlying PDFs (i.e., f_{17} , f_{39} and f_{42}), two non-outlying PDFs (i.e., f_6 and f_{18}) also have high residuals which behave like outliers in the residual plot shown in Figure S-20. It can be seen from Figure S-21 that the shapes of the fitted PDFs of f_6 and f_{18} are highly similar to the target PDFs. Obviously, the high residuals of f_6 and f_{18} are

attributed to the horizontal shift, which will adversely affect the outlier detection in terms of increasing the risk of false detection. Then, we set $\theta_h = 0.2$ and rerun Algorithm S.1, the resulting residuals are shown in Figure S-22. As expected, the negative effects induced by the horizontal-shift error have disappeared, only the three outlying PDFs still hold high residuals.

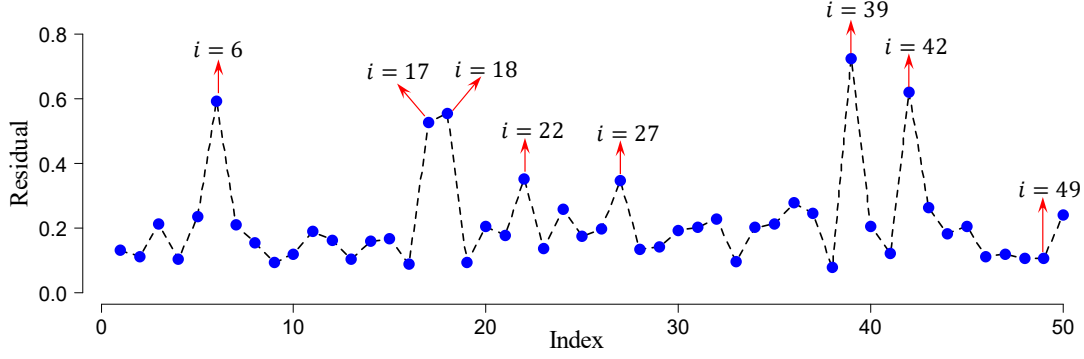


Figure S-20. Residual series $\{\varepsilon_i^g\}_{i=1}^{50}$ calculated by Algorithm S.1 with $\theta_h = 0$ and $\alpha_{mix}^g = 0.1$.

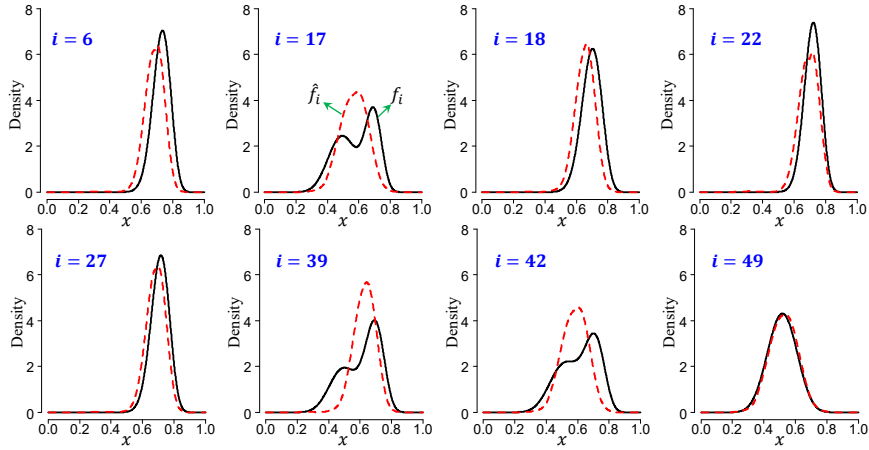


Figure S-21. Comparisons for eight selected PDF samples, the solid line represents the observed PDF to be detected, while the dashed line represents the fitted PDF obtained by the regression model.

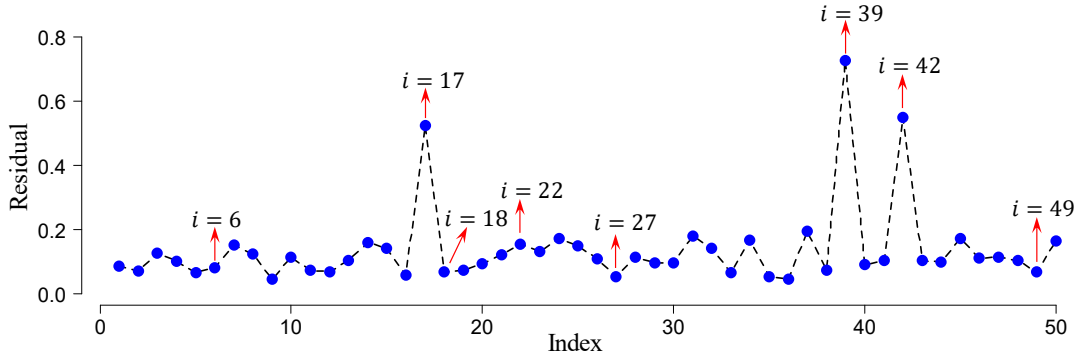


Figure S-22. Residual series $\{\varepsilon_i^g\}_{i=1}^{50}$ calculated by Algorithm S.1 with $\theta_h = 0.2$ and $\alpha_{mix}^g = 0.1$.

S.3.3.3. Residual calculation using the LQD transformation

The residuals calculated using the LQD-transformed PDFs are denoted as $\{\varepsilon_i^{\text{LQD}}\}_{i=1}^n$, and the computational procedure is outlined in Algorithm S.2. As pointed out in Subsection 3.1 of the manuscript, the LQD transformation is insensitive to the horizontal translation of PDFs; thus, we do not consider the aforementioned median alignment in the residual calculation when using the LQD transformation.

Algorithm S.2: Calculation for $\varepsilon_i^{\text{LQD}} = d_{L1}(\text{LQD}[\hat{f}_i|\alpha_{\text{mix}}^{\text{LQD}}], \text{LQD}[f_i|\alpha_{\text{mix}}^{\text{LQD}}])$

Input: PDF pairs (\hat{f}_i, f_i) , PDF preprocessing parameter $\alpha_{\text{mix}}^{\text{LQD}}$

Output: Residual $\varepsilon_i^{\text{LQD}}$

- 1: Set $f_i^*(x) = (1 - \alpha_{\text{mix}}^{\text{LQD}})f_i(x) + \alpha_{\text{mix}}^{\text{LQD}}$, $\hat{f}_i^*(x) = (1 - \alpha_{\text{mix}}^{\text{LQD}})\hat{f}_i(x) + \alpha_{\text{mix}}^{\text{LQD}}$
 - 2: Compute $\psi_i^*(t) = -\log\{f_i^*(Q_i^*(t))\}$, $\hat{\psi}_i^*(t) = -\log\{\hat{f}_i^*(\hat{Q}_i^*(t))\}$, where Q_i^* and \hat{Q}_i^* are the quantile functions associated with f_i^* and \hat{f}_i^* , respectively
 - 3: Compute $\varepsilon_i^{\text{LQD}} = \int_0^1 |\psi_i^*(\tau) - \hat{\psi}_i^*(\tau)| d\tau$
 - 4: Output $\varepsilon_i^{\text{LQD}}$
-

S.4. Supplemental materials for robust distributional regression

S.4.1. Weight design for robust regression operator estimation

This subsection discusses how to design the weights used in Eq. (6) (of the manuscript) for dampening the impacts of functional outliers on the distributional regression operator estimation.

Let $\mathcal{G}_{tr} = \{g_1, g_2, \dots, g_n\}$ and $\mathcal{F}_{tr} = \{f_1, f_2, \dots, f_n\}$ denote the PDF-valued training samples (corresponding to the predictor and response variables, respectively) used for fitting the distribution-to-distribution regression model. As described in the manuscript, the robustness of the regression model is achieved by downweighting the detected outliers. For this purpose, we perform a two-stage outlier detection to the training samples:

- (i) Single dataset outlier detection: outlier detections for the datasets \mathcal{G}_{tr} and \mathcal{F}_{tr} are conducted independently by using the proposed Tree-Distance method described in Subsection 3.1 of the manuscript;
- (ii) Regression outlier detection: the outlier detection for the datasets \mathcal{G}_{tr} and \mathcal{F}_{tr} is conducted jointly by using the distributional regression-based approach described in Subsection 3.2 of the manuscript after the outliers detected in the first stage have been removed.

For convenience, the outliers detected in the first and second stages are called Type I and Type II outliers, respectively. The weights associated with these two types of outliers are designed independently, then combine them to form the final weights, to which we now turn.

S.4.1.1. Weight design for Type I outliers

We consider using the degrees of anomalies computed based on the LQD and CLR transformations to design the desired weights for Type I outliers.

We select dataset $\mathcal{G}_{tr} = \{g_1, g_2, \dots, g_n\}$ to illustrate the weight design procedure, the weights associated with the other dataset (i.e., $\mathcal{F}_{tr} = \{f_1, f_2, \dots, f_n\}$) can be designed in a similar way.

Let $\Psi_{\mathcal{G}} = \{\psi_1^g, \psi_2^g, \dots, \psi_n^g\}$ be the functional dataset composed of the LQD transformations (computed using Eq. (S-11), and the PDF preprocessing parameter α described in Subsection S.2.1 is set to 10^{-10}) of the elements in \mathcal{G}_{tr} . As discussed in Subsection S.3.2.2, the LQD-transformed data are ordinary functional data without constraints, thus the central function of the functional dataset $\Psi_{\mathcal{G}}$ can be selected as the cross-sectional median function denoted as $\psi_c^g(t) = \text{median}_{1 \leq k \leq n} \{\psi_k^g(t)\}$, $\forall t \in [0, 1]$. Then, the dissimilarity of the curve ψ_i^g w.r.t. the central function ψ_c^g is quantified by the L^1 distance, and the result is denoted as $\delta_i^g = d_{L^1}(\psi_i^g, \psi_c^g)$.

Similarly, let $\mathcal{G}_{tr}^{clr} = \{g_1^{clr}, g_2^{clr}, \dots, g_n^{clr}\}$ be the functional dataset composed of the CLR transformations (computed using Eq.(S-5) with the default PDF preprocessing described in Subsection S.2.2) of the elements in \mathcal{G}_{tr} , i.e., $g_i^{clr} = \text{CLR}[g_i]$, $i = 1, 2, \dots, n$. As discussed in Subsection S.3.2.2, the CLR-transformed data are special functional data with the inherent constraint of integrating to zero. To ensure that the selected central function of the functional dataset \mathcal{G}_{tr}^{clr} can satisfy this constraint, the central function is selected as $g_c^{clr}(x) = \text{argmin}_{g^{clr} \in \mathcal{G}_{tr}^{clr}} d_{L^2}(g^{clr}, g_m^{clr})$, where L^2 stands for the L^2 distance and $g_m^{clr}(x) = \text{median}_{1 \leq k \leq n} \{g_k^{clr}(x)\}$ is the cross-sectional median function of the functional data in \mathcal{G}_{tr} . Then, the dissimilarity of the curve g_i^{clr} w.r.t. the corresponding central function g_c^{clr} is quantified by the L^2 distance, and the result is denoted as $\sigma_i^g = d_{L^2}(g_i^{clr}, g_c^{clr})$.

Let $\mathcal{O}_I(\mathcal{G}) \subset \{1, 2, \dots, n\}$ stand for the index set of the detected Type I outliers contained in \mathcal{G}_{tr} . Then, the weight associated with the PDF g_i can be designed as

$$w_I(g_i) = \begin{cases} \left(1 + \frac{|\delta_i^g - m(\delta^g)|}{\text{MAD}(\delta^g)}\right)^{-\rho_1} \left(1 + \frac{|\sigma_i^g - m(\sigma^g)|}{\text{MAD}(\sigma^g)}\right)^{-\rho_1}, & i \in \mathcal{O}_I(\mathcal{G}) \\ 1, & \text{otherwise} \end{cases} \quad (\text{S-22})$$

where $m(\delta^g) = \text{median}_{1 \leq i \leq n} \{\delta_i^g\}$ is the median of the scalar dataset $\{\delta_i^g\}_{i=1}^n$, and $\text{MAD}(\delta^g)$ stands for the median absolute deviation (MAD) calculated by $\text{MAD}(\delta^g) = c \cdot \text{median}_{1 \leq k \leq n} \{|\delta_k^g - m(\delta^g)|\}$ (c is a constant, and we set it to be its default value 1.4826 throughout this study), ρ_1 is a user prescribed tuning factor, $m(\sigma^g)$ and $\text{MAD}(\sigma^g)$ have the similar meanings with $m(\delta^g)$ and $\text{MAD}(\delta^g)$, respectively. The tuning parameter ρ_1 controls the decay rate of the weight function as illustrated in Figure S-23.

Similarly, we can design the weight associated with the PDF f_i based on the Type I outliers detected in \mathcal{F}_{tr} , and the result is denoted as $w_I(f_i)$. The final weight associated with the i th

training sample $\{g_i, f_i\}$ for downweighting the effect of Type I outliers can be obtained by fusing $w_I(g_i)$ and $w_I(f_i)$ as follows (similar to that in Martínez-Hernández et al. (2019)):

$$w_I^i = w_I(g_i) \cdot w_I(f_i), \quad i = 1, 2, \dots, n \quad (\text{S-23})$$

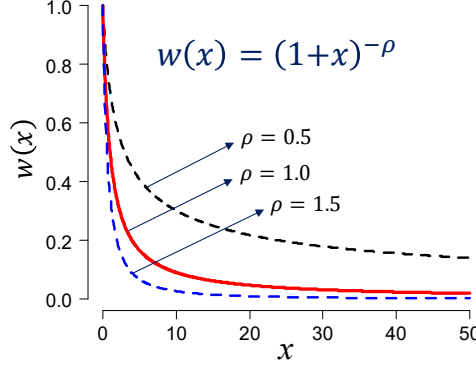


Figure S-23. Schematic illustration of the function $w(x) = (1+x)^{-\rho}$ with different decay parameters.

S.4.1.2. Weight design for Type II outliers

We select the dataset $\mathcal{F}_{tr} = \{f_1, f_2, \dots, f_n\}$ to illustrate the weight design procedure, the weights associated with the other dataset (i.e., $\mathcal{G}_{tr} = \{g_1, g_2, \dots, g_n\}$) can be designed in a similar way.

Let $\varepsilon_i^{\mathfrak{B}}$ and ε_i^{LQD} be the calculated residuals associated with the PDF $f_i \in \mathcal{F}_{tr}$ by using the Algorithm S.1 and Algorithm S.2, respectively. It is worth noting that the results of $\varepsilon_i^{\mathfrak{B}}$ and ε_i^{LQD} actually have been calculated in the residual diagnosis of the regression outlier detection stage. Before using the residuals to design the desired weights, we perform a normalization processing to them as follows:

$$\begin{aligned} \tilde{\varepsilon}_i^{\mathfrak{B}} &= \frac{\varepsilon_i^{\mathfrak{B}} - \min_{1 \leq k \leq n} \varepsilon_k^{\mathfrak{B}}}{\max_{1 \leq k \leq n} \varepsilon_k^{\mathfrak{B}} - \min_{1 \leq k \leq n} \varepsilon_k^{\mathfrak{B}}}, \quad i = 1, 2, \dots, n \\ \tilde{\varepsilon}_i^{LQD} &= \frac{\varepsilon_i^{LQD} - \min_{1 \leq k \leq n} \varepsilon_k^{LQD}}{\max_{1 \leq k \leq n} \varepsilon_k^{LQD} - \min_{1 \leq k \leq n} \varepsilon_k^{LQD}}, \quad i = 1, 2, \dots, n \end{aligned} \quad (\text{S-24})$$

where $\min_{1 \leq k \leq n} a_k$ and $\max_{1 \leq k \leq n} a_k$ stand for the minimum and maximum values of the dataset $\{a_1, a_2, \dots, a_n\}$, respectively.

Let $\mathcal{O}_{II}(\mathcal{F}) \subset \{1, 2, \dots, n\}$ stand for the index set of the type II outliers detected in the PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n, g_i \in \mathcal{G}_{tr}, f_i \in \mathcal{F}_{tr}$. Then, the weight associated with the PDF f_i can be designed as follows:

$$w_{II}(f_i) = \begin{cases} \left(1 + \frac{|\tilde{\varepsilon}_i - m(\tilde{\varepsilon})|}{\text{MAD}(\tilde{\varepsilon})}\right)^{-\rho_2}, & i \in \mathcal{O}_{II}(\mathcal{F}) \\ 1, & \text{otherwise} \end{cases} \quad (\text{S-25})$$

where $\tilde{\varepsilon}_i = (\varepsilon_i^{\mathfrak{B}} + \varepsilon_i^{LQD})/2$, $m(\tilde{\varepsilon}) = \text{median}\{\tilde{\varepsilon}_i\}_{1 \leq i \leq n}$, and $\text{MAD}(\tilde{\varepsilon})$ is the associated median absolute deviation calculated similarly with its counterpart in Eq.(S-22).

Similarly, we can design the weight associated with the PDF $g_i \in \mathcal{G}_{tr}$ based on the Type II outliers, and the result is denoted as $w_{II}(g_i)$. Then, the final weight associated with the i th training sample $\{g_i, f_i\}$ for downweighting the effect of Type II outliers can be also obtained by fusing $w_{II}(g_i)$ and $w_{II}(f_i)$ as follows:

$$w_{II}^i = w_{II}(g_i) \cdot w_{II}(f_i), \quad i = 1, 2, \dots, n \quad (\text{S-26})$$

S.4.1.3. Final weight

On the basis of the weights associated with the Type I and Type II outliers designed above, the final weights used in Eq. (6) (of the manuscript) for downweighting the impacts of detected distributional outliers can be obtained as follows:

$$w_i = w_I^i \cdot w_{II}^i, \quad i = 1, 2, \dots, n \quad (\text{S-27})$$

S.4.2. Proof of proposition 1

Proposition 1 can be proofed in a similar way with Theorem 1 in Lian (2007b).

Proof: recall that the regression operator F_{reg} is assumed to reside in the RKHS $\mathcal{H}(K_r)$, i.e., $F_{reg} \in \mathcal{H}(K_r)$. According to the property of operator-valued reproducing kernel given in Eq.(S-9), it follows that $K_r(\cdot, \psi_j^{g*}) \in \mathcal{H}(K_r), j = 1, 2, \dots, n$. Thus, $\{K_r(\cdot, \psi_j^{g*}) : j = 1, 2, \dots, n\}$ can span a subspace of $\mathcal{H}(K_r)$ as follows:

$$\mathcal{H}_0(K_r) = \left\{ \sum_{j=1}^n K_r(\cdot, \psi_j^{g*}) \alpha_j, \quad \alpha_j \in H \right\} \quad (\text{S-28})$$

where H stands for another Hilbert space. Let $\mathcal{H}_0^\perp(K_r) \subset \mathcal{H}(K_r)$ be the orthogonal complement of $\mathcal{H}_0(K_r)$, then $\forall G \in \mathcal{H}_0^\perp(K_r)$, we have for any $\alpha_j \in H$ that

$$\langle K_r(\cdot, \psi_j^{g*}) \alpha_j, G \rangle_{\mathcal{H}(K_r)} = 0, \quad j \in \{1, 2, \dots, n\} \quad (\text{S-29})$$

Moreover, the regression operator $F_{reg} \in \mathcal{H}(K_r)$ can be decomposed as

$$F_{reg} = F_0 + G, \quad F_0 \in \mathcal{H}_0(K_r), G \in \mathcal{H}_0^\perp(K_r) \quad (\text{S-30})$$

According to the reproducing property given in Eq.(S-10), it follows that

$$\langle G(\psi_j^{g*}), \alpha_j \rangle_H = \langle K_r(\cdot, \psi_j^{g*}) \alpha_j, G \rangle_{\mathcal{H}(K_r)} = 0 \quad (\text{S-31})$$

In view of the arbitrariness of α_j , it follows that $G(\psi_j^{g*}) = 0$, thus $F_{reg}(\psi_j^{g*}) = F_0(\psi_j^{g*}) + G(\psi_j^{g*}) = F_0(\psi_j^{g*})$. Further note that G is orthogonal to F_0 , it has $\|F_{reg}\|_{\mathcal{H}(K_r)} = \|F_0\|_{\mathcal{H}(K_r)} + \|G\|_{\mathcal{H}(K_r)} > \|F_0\|_{\mathcal{H}(K_r)}$ for $G \neq 0$. Consequently, it holds the following inequality for the objective function in Eq. (6) of the manuscript:

$$J(F_{reg}) = \sum_{i=1}^n w_i \left\| \xi_i^{f*} - F_{reg}(\psi_i^{g*}) \right\|_2^2 + \lambda_s \|F_{reg}\|_{\mathcal{H}(K_r)}^2 \quad (\text{S-32})$$

$$> \sum_{i=1}^n w_i \left\| \mathbf{f}_i^{f*} - F_0(\psi_i^{g*}) \right\|_2^2 + \lambda_s \|F_0\|_{\mathcal{H}(K_r)}^2$$

which suggests the optimal solution of the regression operator F_{reg} takes the following general form:

$$F_{reg} = \sum_{j=1}^n K_r(\cdot, \psi_j^{g*}) \alpha_j \quad (\text{S-33})$$

This completes the proof.

S.4.3. Proof of proposition 2

Proof. According to the reproducing properties given in Eqs. (S-8) and (S-10), it follows that

$$\langle b_1 k_r(s, \cdot), b_2 k_r(l, \cdot) \rangle_{H(k_r)} = b_1 b_2 k_r(s, l), \quad \forall b_1, b_2 \in \mathbb{R}, s, l \in \{1, 2, \dots, m\} \quad (\text{S-34})$$

and

$$\begin{aligned} \langle K_r(\psi_1, \cdot) \alpha_1, K_r(\psi_2, \cdot) \alpha_2 \rangle_{\mathcal{H}(K_r)} &= \langle K_r(\psi_1, \psi_2) \alpha_1, \alpha_2 \rangle_{H(k_r)}, \\ \forall \psi_1, \psi_2 \in L^2[0, 1], \alpha_1, \alpha_2 \in H(k_r) \end{aligned} \quad (\text{S-35})$$

It is worth noting that $\mathcal{H}(K_r)$ and $H(k_r)$ are two different Hilbert spaces, $L^2[0, 1]$ and $H(k_r)$ in Eq. (S-35) corresponds to G and \tilde{G} in Eq. (S-10), respectively.

Using the properties given in Eqs. (S-34) and (S-35), the objective function given in Eq.(6) of the manuscript can be transformed into the following matrix form by substituting Eqs. (7) and (9) of the manuscript into Eq. (6).

$$J(\mathbf{B}) = \|\mathbf{W}(\mathbf{Y} - \mathbf{ABK})\|_F^2 + \lambda_s \text{trace}(\mathbf{ABKB}^T) \quad (\text{S-36})$$

where \mathbf{A} , \mathbf{B} , \mathbf{W} , and \mathbf{Y} are the matrixes defined in Eq. (11) of the manuscript, \mathbf{K} is the Gram matrix of the real kernel (see Eq. (10) of the manuscript), and $\|\cdot\|_F$ is the Frobenius norm defined as $\|\mathbf{U}\|_F = (\sum_{i=1}^n \sum_{j=1}^m u_{ij}^2)^{\frac{1}{2}}$. Let $\frac{\partial J(\mathbf{B})}{\partial \mathbf{B}} = 0$, which yields the following matrix equation:

$$\mathbf{AW}^2 \mathbf{ABK}^2 + \lambda_s \mathbf{ABK} = \mathbf{AW}^2 \mathbf{YK} \quad (\text{S-37})$$

This matrix equation can be solved using vectorization. Let $\text{vec}(\mathbf{U})$ denote the vectorization of the matrix $\mathbf{U} = \{u_{ij}\}_{i=1}^n \{j=1}^m$ defined as

$$\text{vec}(\mathbf{U}) = (u_{11} \ u_{21} \ \dots \ u_{n1} \ u_{12} \ u_{22} \ \dots \ u_{n2} \ \dots \ u_{1m} \ u_{2m} \ \dots \ u_{nm})^T \quad (\text{S-38})$$

Noting further the properties of $\text{vec}(\mathbf{AXC}) = (\mathbf{C}^T \otimes \mathbf{A}) \text{vec}(\mathbf{X})$ and $(\mathbf{AC}) \otimes (\mathbf{BD}) = (\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D})$, then Eq.(S-37) can be converted to

$$\left((\mathbf{K} \otimes (\mathbf{AW})) (\mathbf{K} \otimes (\mathbf{WA})) + \lambda_s (\mathbf{K} \otimes \mathbf{A}) \right) \text{vec}(\mathbf{B}) = (\mathbf{K} \otimes (\mathbf{AW})) \text{vec}(\mathbf{WY}) \quad (\text{S-39})$$

Let $\mathbf{C}_1 = \mathbf{K} \otimes (\mathbf{AW})$ and $\mathbf{C}_2 = \mathbf{K} \otimes (\mathbf{WA})$ complete the proof.

S.5. Simulation studies

S.5.1. Simulation study I

This simulation study is conducted to validate the effectiveness of the proposed Tree-Distance outlier detection method, as well as to assess its performance by comparing it to its competitors. We employ a mixture distribution model, formed by taking a convex combination of beta distribution and truncated generalized Pareto distribution (tGPD) to generate the PDF-valued data. Specifically, we consider the following two scenarios:

$$\text{Scenario I: } f_i(x) = (1 - \eta)f_{Beta}(x; a_i, b_i) + \eta f_{tGPD}(x; 2.0, 4.0), i = 1, 2, \dots, n \quad (\text{S-40})$$

$$\text{Scenario II: } f_i(x) = (1 - \eta)f_{Beta}(x; a_i, b_i) + \eta f_{tGPD}(x; 0.5, 0.5), i = 1, 2, \dots, n$$

where $\eta \in [0, 1]$ is the combination coefficient, f_{Beta} is the density of the beta distribution, f_{tGPD} is the density of the tGPD obtained as follows

$$f_{tGPD}(x; \kappa, \sigma) = \frac{f_{GPD}(x; \kappa, \sigma)}{\int_0^1 f_{GPD}(\tau; \kappa, \sigma) d\tau}, x \in [0, 1] \quad (\text{S-41})$$

where $f_{GPD}(x; \kappa, \sigma)$ is the density of the generalized Pareto distribution (GPD) defined as

$$f_{GPD}(x; \kappa, \sigma) = \left(\frac{1}{\sigma}\right) \left(1 + \kappa \frac{x}{\sigma}\right)^{-1-\frac{1}{\kappa}} \quad (\text{S-42})$$

The parameters a_i and b_i associated with the beta distribution in both scenarios are i.i.d. realizations of uniform distributions, i.e., $a_i \sim U[10, 35]$ and $b_i \sim U[14, 20]$. In each scenario, we consider four different values for the combination parameter η , i.e., $\eta = 0, 0.15, 0.30$ and 0.45 , to yield four different models referred to as Models I, II, III, and IV (listed in Table S-2) throughout this simulation study.

In each scenario, we independently use the four models to generate four different PDF-valued datasets, with each dataset consisting of $n = 100$ curves. We then employ Algorithm S.7 in Appendix 2 to introduce 10 outlying PDFs into each simulated dataset (i.e., the contamination ratio is 10%). The parameter ζ_{hs} in Algorithm S.7 is set to 0, meaning that only the shape outliers are generated while the horizontal-shift outliers are not considered because the latter is much easier to detect. The parameter ϖ in Algorithm S.7 is set to 0.2. Representative simulated data using the eight models (listed in Table S-2) are visualized in Figure S-24.

Table S-2

Considered models for distributional data generation.

	Model I	Model II	Model III	Model IV
Scenario I	$\eta = 0$	$\eta = 0.15$	$\eta = 0.30$	$\eta = 0.45$
Scenario II	$\eta = 0$	$\eta = 0.15$	$\eta = 0.30$	$\eta = 0.45$

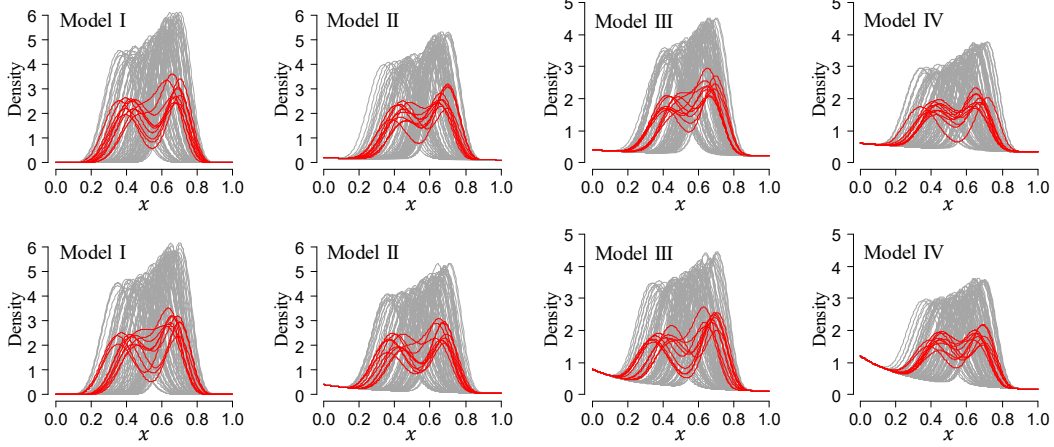


Figure S-24. Representative simulated data associated with the eight models, the gray lines and red lines stand for the “good” data and outlying data, respectively. The first row corresponds to Scenario I, while the second row corresponds to Scenario II.

We consider three different detection schemes and compare their performances.

(i) Tree-Distance method

In this detection scheme, we perform outlier detections on the transformed data associated with the nLQD, CLR, DIFF, and MED nodes using the default argument settings listed in Table S-1. The outliers detected from the four nodes are merged to form the final detection results. Note that before transforming the PDFs to their CLR-representations (corresponding to the CLR node), they should be aligned horizontally in the H-CENTR node, and the feature point for such an alignment is calculated by the average of the median and mode.

(ii) QF-FDO method

In this detection scheme, we consider using the tool of functional directional outlyingness (FDO) defined by Dai and Genton (2019) to perform outlier detection in the quantile function (QF)-space for the simulated distributional data. For convenience, such a detection method is referred to as QF-FDO method throughout the rest of this study. The relevant computational details are provided in Appendix 4. The detection region is set as $[0.2, 0.8]$. The outliers in the MO- and VO-directions are detected by the two- and one-sided boxplot-based detectors (see Eq. (S-14) and Eq. (S-15)), respectively. We consider three different whisker parameters, valued at 1.5, 2.0, and 2.5, for VO-outlier detection, while the whisker parameter for MO-outlier detection is fixed at 1.5. Such settings yield three different detection cases for the QF-FDO method.

(iii) Warping function-based detection method using the phase distance

The warping function-based detection strategy considered here essentially belongs to the elastic depth-based approach proposed by Harris et al. (2021). The outliers are detected based on the phase information of the CDFs captured by the warping functions. The relevant computational details and the argument settings are provided in Appendix 4.

In the simulated PDF dataset consisting of 100 curves, let n_{out} and n_{norm} denote the number of outlying PDFs and non-outlying PDFs, respectively, and n_c^{DT} and n_f^{DT} denote the number of correctly and falsely detected outlying PDFs, respectively. The performance of the

detection methods can be assessed based on the correct and false detection rates, defined as follows:

$$\text{Correct detection rate: } p_c = \frac{n_c^{DT}}{n_{out}} \times 100\%$$

$$\text{False detection rate: } p_f = \frac{n_f^{DT}}{n_{norm}} \times 100\%$$

We repeat the above outlier detection experiment for 1000 times using re-simulated functional data to calculate the average correct and false detection rates for each detection scheme, and the results are reported in Tables S-3~5. For comparison, we also calculate the average correct (false) detection rates associated with the outliers detected in the four considered nodes (i.e., MED, nLQD, CLR, and DIFF) on the tree of the first detection scheme, and the corresponding results are listed in columns 3~6 of Table S-3. It is worth noting that the data-generating processes of Model I are the same in Scenarios I and II; thus, the detection results of Model I are only reported for Scenario I.

Table S-3

The calculated average correct detection rates and false detection rates (in brackets) associated with the Tree-Distance detection scheme. The data in the third, fourth, fifth, and sixth columns correspond to the results detected in the nodes of MED, nLQD, CLR, DIFF, respectively, while the last column corresponds to the merged results detected in the four considered nodes on the tree. Both p_c and p_f (in brackets) are presented in percentage terms.

Scenario	Model	MED	nLQD	CLR	DIFF	TREE
		$p_c(\%)$ (p_f)(%)	$p_c(\%)$ (p_f)(%)	$p_c(\%)$ (p_f)(%)	$p_c(\%)$ (p_f)(%)	$p_c(\%)$ (p_f)(%)
Scenario I	Model I	0.00 (0.12)	98.49 (0.00)	98.77 (0.00)	0.00 (0.00)	99.42 (0.12)
	Model II	0.00 (0.14)	95.87 (0.00)	97.76 (0.00)	0.00 (0.00)	98.60 (0.14)
	Model III	0.00 (0.09)	93.74 (0.00)	97.64 (0.00)	0.00 (0.00)	98.33 (0.09)
	Model IV	0.00 (0.06)	87.68 (0.00)	97.14 (0.00)	0.00 (0.00)	97.74 (0.06)
Scenario II	Model I	—	—	—	—	—
	Model II	0.00 (0.10)	97.47 (0.05)	94.97 (0.00)	0.00 (0.00)	98.60 (0.14)
	Model III	0.00 (0.03)	97.10 (0.23)	85.00 (0.00)	0.00 (0.00)	97.61 (0.25)
	Model IV	0.00 (0.01)	82.49 (0.01)	70.95 (0.00)	0.00 (0.00)	88.02 (0.02)

For the first detection scheme, it can be seen from Table S-3 that the detected outliers are mainly from the nodes of nLQD and CLR, and rarely from the nodes of MED and DIFF. In Scenario I, the CLR node slightly outperforms the nLQD node, whereas, in Scenario II, the opposite is the case. Recall that the outliers detected by the tree are obtained by merging the outliers identified in the four considered nodes, the calculated average correct detection rates (false detection rates) associated with the tree (listed in the last column of Table S-3) are the final results of the Tree-Distance detection scheme. It is evident from the results that the average correct detection rates of the tree are higher than those associated with the nLQD and CLR nodes, especially for Model IV of Scenario II, indicating that the outliers detected in the nLQD and CLR nodes are different. Thus, the two transformations have complementarity in outlier detection.

Note that the role of the MED node is mainly to detect the horizontal-shift outliers; thus, the outliers detected from this node are mainly false positives because the simulated data only contain shape outliers with no significant horizontal deviations with respect to the bulk of the data. The DIFF node shows almost no contribution to the outlier detection in this experiment; however, we cannot conclude that the DIFF node is useless and should be removed from the transformation tree considering that it may have its merits in other situations. To demonstrate this, we provide an additional simulation study using another simulated PDF-valued dataset in Subsection S.6.1, where the DIFF node plays the dominant role in uncovering the outliers.

Table S-4

The calculated average correct detection rates and false detection rates (in brackets) associated with the QF-FDO detection scheme. The data in the third, fourth, and fifth columns correspond to the results detected by setting the whisker parameters associated with the VO-outliers to be 1.5, 2.0, and 2.5, respectively, while the whisker parameter associated with the MO-outliers is fixed at 1.5. Both p_c and p_f (in brackets) are presented in percentage terms.

Scenario	Model	1.5IQR (VO)	2.0IQR (VO)	2.5IQR (VO)
		$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)
Scenario I	Model I	61.12 (2.85)	54.39 (1.46)	48.98 (0.73)
	Model II	65.38 (2.76)	58.82 (1.34)	52.65 (0.68)
	Model III	68.04 (2.13)	61.55 (0.97)	55.26 (0.44)
	Model IV	71.00 (6.21)	63.06 (4.56)	55.67 (3.40)
	Model I	—	—	—
Scenario II	Model II	63.99 (2.90)	57.06 (1.46)	51.03 (0.74)
	Model III	71.59 (2.41)	65.10 (1.13)	58.66 (0.52)
	Model IV	59.25 (16.02)	55.99 (15.32)	53.47 (14.78)

For the second detection scheme (see Table S-4), the calculated average correct detection rates are significantly lower than those of the first detection scheme. Comparing the average false detection rates between Table S-3 and Table S-4, it appears that the second detection scheme also has a higher risk of false detection, especially for Model IV in Scenario II. Reducing the whisker parameter of the boxplot associated with the VO-outliers can increase its outlier detection power; however, only a slight improvement is observed in this experiment, which also leads to a higher risk of false detection.

The results associated with the third detection scheme, listed in Table S-5, show that the average correct detection rates are lower than 50% for the considered cases, meaning that fewer than half (on average) of the outliers have been successfully identified by the warping function-based method. Moreover, the calculated average false detection rates associated with the third detection scheme are significantly higher than those of the other two detection schemes for most cases.

Table S-5

The calculated average correct detection rates and false detection rates (in brackets) associated with the warping function-based detection scheme. Both p_c and p_f (in brackets) are presented in percentage terms.

Scenario	Model I $p_c(\%)$ ($p_f(\%)$)	Model II $p_c(\%)$ ($p_f(\%)$)	Model III $p_c(\%)$ ($p_f(\%)$)	Model IV $p_c(\%)$ ($p_f(\%)$)
Scenario I	43.08 (8.11)	43.15 (8.35)	38.09 (8.48)	32.33 (8.79)
Scenario II	—	42.18 (8.73)	38.74 (9.03)	35.78 (9.04)

Comparing Tables S-3~5, we see that the first detection scheme (i.e., the proposed Tree-Distance method) performs excellently with high accuracy and low risk of false detection, while the second detection scheme significantly underperforms in comparison, and the third detection scheme is the worst performer.

As mentioned in Subsection 3.1 of the manuscript (detailed in Subsection S.3.2.3 of this document), both the L^1 and L^2 distances are valid for performing outlier detection for the functional data associated with nodes nLQD and DIFF (involved in the Tree-Distance method). From the demonstrations presented in Subsection S.3.2.3 of this document, we get that the L^1 and L^2 distances are mainly suited for measuring the global dissimilarity of the functional samples. In the performance comparative study conducted above, the Tree-Distance method is executed using the default argument settings listed in Table S-1, where the L^1 distance is chosen for global dissimilarity quantification for the functional data associated with nodes nLQD and DIFF. If we replace the L^1 distance by the L^2 distance in the default argument settings for nodes nLQD and DIFF, we are curious to see how the detection results of the Tree-Distance method would change. For this purpose, we re-run the Tree-Distance method to the same 1000 PDF-valued datasets generated earlier using the new argument settings (i.e., the L^1 distance associated with nodes nLQD and DIFF in Table S-1 is replaced by the L^2 distance, the other argument settings remain unchanged), and the re-calculated average correct/false detection rates are reported in Table S-6. Comparing the results in Tables S-3 and S-6, one can see that the L^1 distance and L^2 distance have similar performances in most situations except for the case of Model IV of Scenario II. In the latter case, the L^1 distance performs slightly better than the L^2 distance. This is also the main reason why we select the L^1 distance as the default distance for global dissimilarity quantification for the functional data associated with nodes nLQD and DIFF in this study. Such a default setting does not mean that the L^1 distance would perform better than the L^2 distance in all situations, one can also use the L^2 distance to perform outlier detection for nodes nLQD or DIFF depending on the specific situation.

Table S-6

The re-calculated average correct detection rates and false detection rates (in brackets) associated with the Tree-Distance detection scheme using the new argument settings (i.e., the L^1 distance associated with nodes nLQD and DIFF in Table S-1 is replaced by the L^2 distance, the other argument settings remain unchanged). Both p_c and p_f (in brackets) are presented in percentage terms.

Scenario	Model	MED	nLQD	CLR	DIFF	TREE
		$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)	$p_c(\%)$ ($p_f(\%)$)
Scenario I	Model I	0.00 (0.12)	98.51 (0.00)	98.77 (0.00)	0.00 (0.01)	99.42 (0.13)
		0.00 (0.14)	96.30 (0.00)	97.76 (0.00)	0.00 (0.00)	98.67 (0.15)
	Model II	0.00 (0.09)	93.93 (0.00)	97.64 (0.00)	0.00 (0.00)	98.36 (0.09)
	Model III	0.00 (0.06)	85.85 (0.00)	97.14 (0.00)	0.00 (0.01)	97.59 (0.06)
	Model IV	—	—	—	—	—
	Model I	0.00 (0.10)	97.72 (0.06)	94.97 (0.00)	0.00 (0.01)	98.79 0.15
	Model II	0.00 (0.03)	97.09 (0.29)	85.00 (0.00)	0.00 (0.00)	97.54 (0.31)
	Model III	0.00 (0.01)	75.22 (0.00)	70.95 (0.00)	0.00 (0.00)	84.71 (0.02)
Scenario II	Model IV	—	—	—	—	—
	Model I	0.00 (0.10)	97.72 (0.06)	94.97 (0.00)	0.00 (0.01)	98.79 0.15
	Model II	0.00 (0.03)	97.09 (0.29)	85.00 (0.00)	0.00 (0.00)	97.54 (0.31)
	Model III	0.00 (0.01)	75.22 (0.00)	70.95 (0.00)	0.00 (0.00)	84.71 (0.02)

S.5.2. Simulation study II

This simulation study aims to validate the effectiveness of the distributional regression-based approach in abnormal association detection (i.e., regression outlier detection).

Algorithm S.3: Generating PDF-valued two-tuples

Input: Number of two-tuples n

Output: PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$

1: **for** $i = 1$ **to** n **do**

 a: Generate parameters for $g_i(x)$

$$a_i \sim U(10, 40), b_i \sim U(14, 40), q_i \sim U(0, 0.5)$$

 b: Generate PDF $g_i(x) = (1 - q_i)\text{BetaPdf}(x; a_i, b_i) + q_i\text{BetaPdf}(x; 2a_i, b_i)$

 c: Generate parameters for $f_i(x)$

$$e_i \sim N(0, 5^2)$$

$$c_i = 2.5a_i + \sqrt{a_i} - 15 + e_i, \quad d_i = 0.5\sqrt{a_i b_i} + 45 - 0.8a_i + e_i$$

$$z_i = (c_i + d_i)/2$$

 d: Generate PDF $f_i(x) = (1 - q_i)\text{BetaPdf}(x; c_i, d_i) + q_i\text{BetaPdf}(x; z_i, z_i)$

end for

2: Output $\{g_i, f_i\}_{i=1}^n$

First, we use Algorithm S.3 to simulate $n = 100$ groups of correlated PDF-valued two-tuples denoted as $\{g_i, f_i\}_{i=1}^n$. The parameters of the distributions corresponding to $f_i, i = 1, 2, \dots, n$ are nonlinearly dependent on those corresponding to $g_i, i = 1, 2, \dots, n$. Representative functional samples of the simulated PDF-valued data $\{g_i\}_{i=1}^{100}$ and $\{f_i\}_{i=1}^{100}$ are visualized in the left column of Figure S-25, and the right column presents five typical curves selected from each PDF dataset. Both PDF datasets contain unimodal and bimodal curves, and such distributional data are relatively

complex from the angle of the curve shapes. The implementation details for generating abnormal associations in $\{g_i, f_i\}_{i=1}^n$ based on an intra-element exchange strategy are summarized in Algorithm S.9 (in Appendix 3). Note that the intra-element exchange can only rearrange the order of the elements. If viewed independently, the curve plots of $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$ after processing by Algorithm S.9 are the same as those in the upper and lower left panels of Figure S-25, respectively. To visualize the simulated abnormal associations, we plot the curves of the abnormal PDF-valued two-tuple (denoted as $\{g_j, f_j\}$) along with the histogram of random samples generated from the original distribution (the one before performing the element exchange operation), as shown in Figure S-26 (a), where the abnormal PDF can be distinguished as it no longer fits the histogram well. For comparison, Figure S-26(b) also illustrates another representative PDF-valued two-tuple (denoted as $\{g_k, f_k\}$) with a normal association.

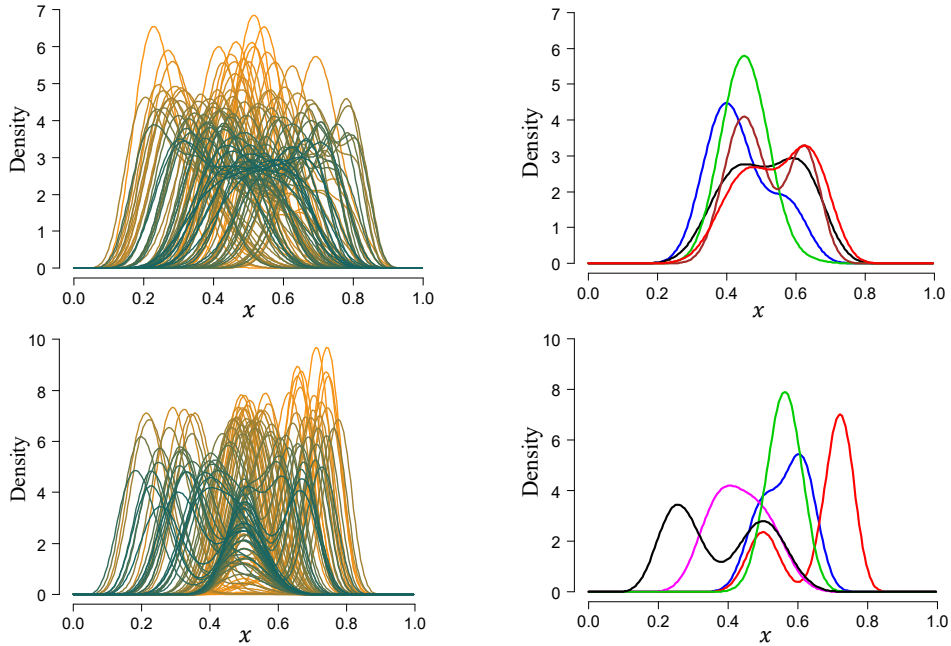


Figure S-25. Representative simulated functional samples of the PDF-valued datasets $\{g_i\}_{i=1}^{100}$ (the upper row) and $\{f_i\}_{i=1}^{100}$ (the lower row). The left column corresponds to the whole curves contained in the functional datasets, while the right column corresponds to five selected representative curves from the PDFs shown in the left panel of the same row.

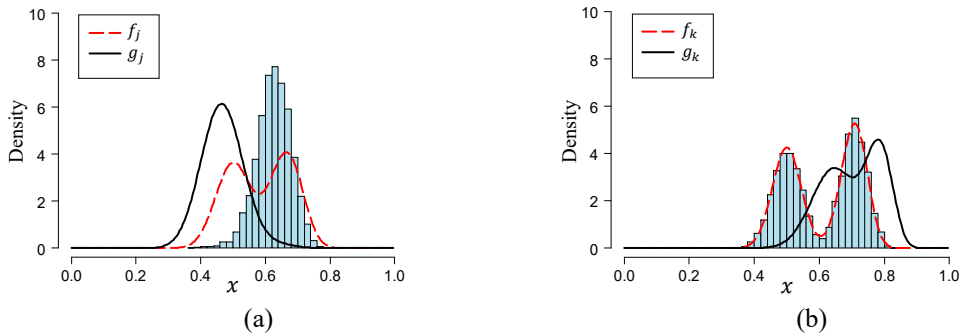


Figure S-26. (a) Visualization of a representative simulated abnormal PDF-valued two-tuple and (b) visualization for a representative simulated “good” PDF-valued two-tuple. The histogram is drawn based on 5000 random samples drawn from the original distribution associated with f_j (or f_k) before the PDF exchange.

In the following, the distributional regression-based approach described in Subsection 3.2 of the manuscript with implementation details summarized in Algorithm 1 in the manuscript, is employed to detect the regression outliers. For the LQD-RKHS distribution-to-distribution regression model, the Gaussian kernel given in Eq. (8) (of the manuscript) is selected as the reproducing operator kernel, and the related parameter σ is determined by the following principle (similar to that in Chen et al. (2019a)):

$$\sigma = \sum_{i \in I_{tr}} \sum_{j \in I_{tr}} \left(\int |\psi_i^{g^*}(\tau) - \psi_j^{g^*}(\tau)|^2 d\tau \right)^{1/2} / |I_{tr}|^2 \quad (\text{S-43})$$

where I_{tr} stands for the index set of the training data, $|I_{tr}|$ represents the number of elements contained in I_{tr} . The other argument settings are listed in Table S-7. Similar to Subsection S.5.1, the performance is assessed based on the correct and false detection rates averaged over a series of repeated detection experiments. In each repetition, a total of 100 groups of PDF-valued two-tuples are simulated using Algorithm S.3, and 10 abnormal associations are generated by Algorithm S.9 to contaminate the original distributional data. We set $(M_g, M_f) = (0, 5), (2, 3), (4, 1)$, and $(5, 0)$ in Algorithm S.9 to consider four different contamination scenarios. Using the results of 500 repeated detection experiments, Table S-8 lists the calculated average correct and false detection rates for the four considered contamination scenarios. The correct detection rates are all greater than 80%, with 92.24% being the best, indicating that the distributional regression-based approach can effectively detect regression outliers. The false detection rates are approximately 6%, implying that out of the 90 “good” curves, approximately six curves will be falsely identified as outliers, on average. At first glance, the false detection rate is relatively high, which might be attributed to the regression error. Recall that a function-to-function regression model can be generally written as $f(x) = \Gamma(g(x)) + \varepsilon(x)$ with Γ and $\varepsilon(x)$ being the regression operator and functional error term (assumed to be zero mean), respectively, and the functional response $f(x)$ is independent of the error term $\varepsilon(x)$. Thus, given a specified predictor g_k , only the conditional mean $E(f_k|g_k) = \Gamma(g_k)$ can be predicted by the fitted regression model, whereas the quantity of the error is unpredictable. If the simulated error ε_k in generating the experimental data is considerably large, the corresponding PDF-valued two-tuple $\{g_k, f_k\}$ may also be an outlier. Although we have leveraged a horizontal threshold θ_h in Algorithm S.1 to reduce the risk of false detection caused by horizontal shift errors of PDFs (see Subsection S.3.3.2 for details), the remaining shape errors may also lead to false detections.

Table S-7

Argument settings for the regression outlier detection.

$(\alpha_{mix}^{\text{LQD}}, \alpha_{mix}^{\text{S}})$	$(r_2^{\text{LQD}}, r_2^{\text{S}})$	θ_h	θ_λ	m	N_{iters}^{reg}
(0.3, 0.1)	(1.5, 1.5)	0.15	0.01	5	4

Table S-8

The calculated correct and false detection rates associated with the regression outlier detection for four different contamination scenarios.

(M_g, M_f)	(0,5)	(2,3)	(4,1)	(5,0)
p_c (%)	92.24	81.76	81.20	84.06
p_f (%)	5.57	7.01	6.81	5.95

In general, with average correct detection rates greater than 80%, the proposed distributional regression-based outlier detection method performs well, as the simulated data are quite complex (see Figure S-25), and the false detection rates are also acceptable. If we reduce the complexity of the simulated data, the detection method can be expected to perform better. For comparison, we also conduct an additional simulation study with less complicated PDF-valued data in the next section. As expected, the detection performance is much better, and the correct detection rates (averaged over 500 replicated detection experiments) are all greater than 98% (see Table S-10 in Subsection S.6.2).

S.6. Additional simulation studies

S.6.1. Additional simulation study I

This subsection provides an additional simulation study, and the outlying PDF detection method described in Subsection S.5.1 is employed to identify the synthetic outlying curves. Similar to that in Subsection S.5.1, the average detection performance of interest is also based on a series of repeated experiments. In each simulation run, we first independently generate a functional dataset consisting of 100 curves using the following model:

$$v_i(x) = a_i \sin(2\pi x) + b_i \cos(2\pi x), x \in [0, 1], i = 1, \dots, 100 \quad (\text{S-44})$$

with $a_i \stackrel{i.i.d.}{\sim} U(0.012, 0.05)$ and $b_i \stackrel{i.i.d.}{\sim} U(0.012, 0.075)$. The simulated functional dataset is referred to as $S_v = \{v_i(x)\}_{i=1}^{100}$. After introducing 10 functional outliers by using Algorithm S.4, all functions in S_v are converted to PDFs through the following principle:

$$f_i(x) = \frac{v_i(x) - b}{\int_0^1 (v_i(\tau) - b) d\tau}, \quad v_i \in S_v \text{ and } b = \min_{v_k \in S_v} \inf_{x \in [0, 1]} \{v_k(x)\} \quad (\text{S-45})$$

The resulting PDF-valued dataset is denoted as $S_f = \{f_i(x)\}_{i=1}^{100}$, representative samples of S_f are displayed in Figure S-27, where the red lines represent the synthetic outliers.

Then, we apply the proposed Tree-Distance detection scheme, with the default argument settings as those in Subsection S.5.1, to the collection of simulated PDFs in S_f . We independently repeat such detection experiments for 1000 times, the calculated average values of correct and false detection rates are reported in Table S-9.

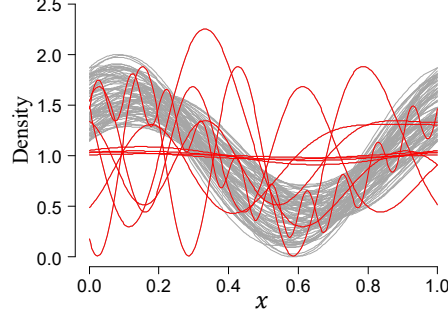


Figure S-27. Representative simulated PDF-valued dataset composed of 90 “good” densities (gray lines) and 10 outlying densities (red lines) using model (S-45) and Algorithm S.4.

Table S-9

The calculated average correct detection rates and false detection rates in Additional simulation study I. The data in the second, third, fourth and fifth columns correspond to the results detected in the nodes of MED, nLQD, CLR, DIFF, respectively, while the last column corresponds to the merged results detected in the four considered nodes on the tree.

	MED	nLQD	CLR	DIFF	TREE
$p_c(\%)$	41.43	45.20	28.17	61.83	99.00
$p_f(\%)$	0.00	1.14	0.01	0.00	1.15

Algorithm S.4: Generate functional outliers

Input: Functional dataset $S_v = \{v_i(x)\}_{i=1}^n$ with $v_i(x)$ defined on the compact interval $[0,1]$, the number of outliers N_o

Output: The contaminated functional dataset

1: Initialize the set of outliers $S_{out} = \emptyset$

2: **for** $i = 1$ **to** $N_o - 1$ **do**

Generate $z \sim U(0,1)$

if $z < 0.6$ **then**

Compute $\gamma(x) = c(x - 4) + x^3$ with $c \sim U(-4.5, -2)$, $x \in [0,1]$

Compute $v_o(x) = a_o \sin(2\pi\gamma(x)) + b_o \cos(2\pi\gamma(x))$

with $a_o \sim U(0.02, 0.05)$ and $b_o \sim U(0, 0.075)$, $x \in [0,1]$

else

Compute $v_o(x) = a_o \sin(2\pi x) + b_o \cos(2\pi x)$

with $a_o \sim U(0, 0.008)$ and $b_o \sim U(0, 0.008)$, $x \in [0,1]$

end if

Put v_o into the outlier set $S_{out} \leftarrow v_o$

end for

3: Calculate the pointwise median function

$$v_{med}(x) = \underset{1 \leq k \leq n}{\text{median}}\{v_k(x)\}, \forall x \in [0,1]$$

4: Use the L^2 distance to find an element in S_v closest to $v_{med}(x)$, and denote the found element as v_m

5: Compute $v_o(x) = v_m(x) + 0.02 \sin(20\pi x)$, $x \in [0,1]$, and set $S_{out} \leftarrow v_o$

6: Randomly select N_o elements in S_v to be replaced by the generated outliers stored in S_{out}

7: Output the outlier contaminated functional dataset S_v

S.6.2. Additional simulation study II

This subsection conducts an additional simulation study for regression outlier detection in parallel with the one conducted in Subsection S.5.2.

Algorithm S.5: Generating PDF-valued two-tuples for Additional simulation study II

Input: Number of two-tuples n

Output: PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$

- 1: Independently generate parameters for $g_i(x), i = 1, 2, \dots, n$
 - $\mathbf{A} = \{a_1, a_2, \dots, a_n\}$ with a_i being i.i.d. samples of $U(14, 30)$
 - $\mathbf{B} = \{b_1, b_2, \dots, b_n\}$ with b_i being i.i.d. samples of $U(14, 20)$
 - 2: Independently generate the errors
 - $\mathbf{E} = \{e_1, e_2, \dots, e_n\}$ with e_i being i.i.d. samples of $N(0, 3^2)$
 - 3: **for** $i = 1$ **to** n **do**
 - a: Generate PDF $g_i(x) = \text{BetaPdf}(x; a_i, b_i)$
 - b: Generate parameters for $f_i(x)$

$$c_i = 40 \frac{a_i - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A})} + 12 + e_i, \quad d_i = \sqrt{a_i b_i + a_i} + e_i$$
 - c: Generate PDF $f_i(x) = \text{BetaPdf}(x; c_i, d_i)$
 - end for**
 - 4: Output $\{g_i, f_i\}_{i=1}^n$
-

Algorithm S.6: Generating abnormal PDF associations by inserting outliers

Input: PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$, the bivariate parameter (N_g, N_f) for controlling the number of outliers introduced to $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$, respectively, and coefficients ζ_{hs} and ϖ

Output: The contaminated PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$

- 1: Denote the PDF-valued datasets $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$ as S_{PDF}^g and S_{PDF}^f , respectively
 - 2: **repeat**
 - a: Randomly insert N_g outlying PDFs into S_{PDF}^g using Algorithm S.7
$$(S_{PDF}^g, \text{IDE}_g) = \text{PDFoutlier_Insert}(S_{PDF}^g, N_g, \zeta_{hs}, \varpi)$$
 - b: Randomly insert N_f outlying PDFs into S_{PDF}^f using Algorithm S.7
$$(S_{PDF}^f, \text{IDE}_f) = \text{PDFoutlier_Insert}(S_{PDF}^f, N_f, \zeta_{hs}, \varpi)$$
 - until** $\text{IDE}_g \cap \text{IDE}_f = \emptyset$ (\emptyset denotes the empty set)
 - 3: Output the contaminated PDF-valued two-tuples, i.e.,
$$\{g_i, f_i\}, g_i \in S_{PDF}^g, f_i \in S_{PDF}^f, i = 1, 2, \dots, n$$
-

In each run, 100 groups of correlated PDF-valued two-tuples denoted as $\{g_i, f_i\}_{i=1}^{100}$ are generated by using Algorithm S.5, and the abnormal associations are generated by using Algorithm S.6 with $\zeta_{hs} = 0$ and $\varpi = 0.25$. We consider seven different contamination scenarios with (N_g, N_f) (denoting the associated numbers of outliers introduced to $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$) valued at (10, 0), (8, 2), (6, 4), (5, 5), (4, 6), (2, 8) and (0, 10), respectively. In each contamination scenario, the distributional regression-based detection method with the same argument settings of those in Subsection S.5.2 is employed to detect the abnormal associations (i.e., the regression outliers). Based on 500 repeated detection tests, the calculated average correct and false detection rates are

listed in Table S-10 for the seven considered contamination scenarios.

Table S-10

The calculated average correct and false detection rates associated with the regression outlier detection conducted in Additional simulation study II for the seven considered contamination scenarios.

(N_g, N_f)	(10, 0)	(8, 2)	(6, 4)	(5, 5)	(4, 6)	(2, 8)	(0, 10)
p_c (%)	99.82	99.22	99.18	98.54	98.90	98.98	99.70
p_f (%)	4.08	4.79	5.22	5.48	5.30	5.00	3.97

S.7. Supplemental materials for the real data study

S.7.1. The two investigated strain sensors

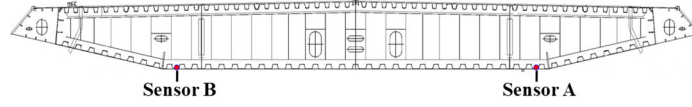


Figure S-28. Positions of the two strain sensors investigated in the real data study. The cross-section of the bridge is the same as that in Chen et al. (2019a).

S.7.2. Data preprocessing in the real data analysis

The sample frequency of the strain sensor is 4Hz, thus the 8 days of measurements of each sensor include a total of 2764800 data points. Before PDF estimation, a rough pretreatment is conducted to the raw data since the data quality of Sensor B is too poor. Specifically, let \mathbf{x} and \mathbf{y} be the vectors of the selected 8 days of time-ordered measurements collected by Sensor A and Sensor B (see Figure S-29 (a)), respectively, then the two-tuple $(\mathbf{x}(k), \mathbf{y}(k))$ would be removed from (\mathbf{x}, \mathbf{y}) if the following condition is satisfied:

$$\mathbf{x}(k) = \text{NaN} \text{ or } \mathbf{y}(k) = \text{NaN} \text{ or } \mathbf{y}(k) > 50 (\mu\epsilon)$$

where NaN stands for the missing data. Finally, the remaining data are merged together in the original time order and then scaled to $[0, 1]$ (similar to that in Chen et al. (2019a)) for each sensor, the resulting data are visualized in Figure S-29 (b).

For density estimation, the post-processed measurements are divided into 120 segments for each sensor. Each segment consists of 11604 data points (equal to around 48 minutes measurement amount). The PDFs are estimated by kernel density estimator using the segment data as samples (similar to that in Chen et al. (2019a)). The estimated PDFs are denoted as $\{\hat{g}_i\}_{i=1}^{120}$ and $\{\hat{f}_i\}_{i=1}^{120}$ for Sensor A and Sensor B, and visualized in Figures S-30 (a) and (b), respectively.

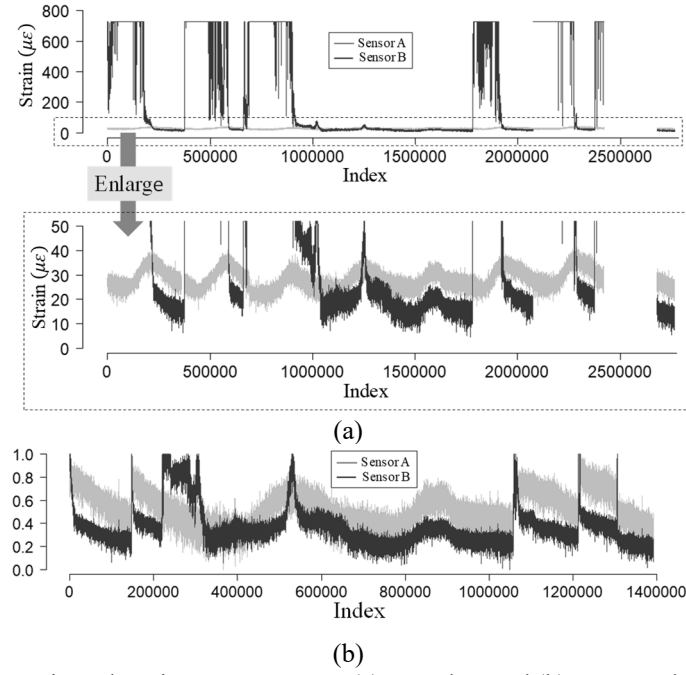


Figure S-29. The investigated strain measurements. (a) Raw data and (b) processed data (eliminating the missing or large values ($>50\mu\epsilon$), and then scaled to $[0,1]$).

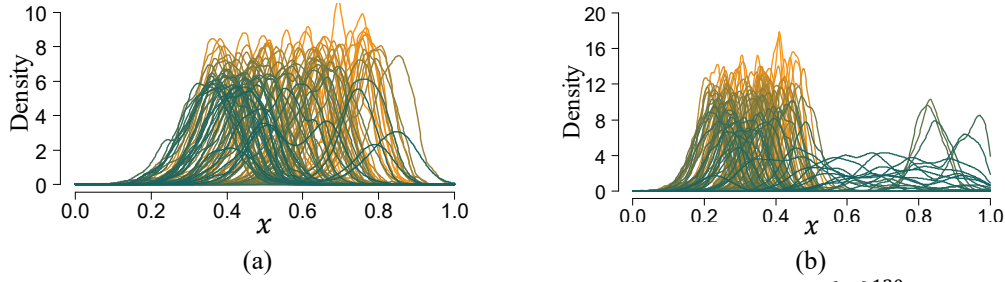


Figure S-30. The estimated PDFs of post-processed strain measurements: (a) $\{\hat{g}_i\}_{i=1}^{120}$ associated with Sensor A and (b) $\{\hat{f}_i\}_{i=1}^{120}$ associated with Sensor B.

S.7.3. Tables of argument settings involved in the real data study

Table S-11

Argument settings for the two-stage initial outlier detection conducted in the real data study in Section 6 of the manuscript

Stage I: Single dataset outlier detection

Detection method	The Tree-Distance method described in Subsection 3.1 of the manuscript
Type of outliers	Type I
Argument setting	The default argument settings as used in Simulation study I (presented in Subsection S.5.1 of this document)

Stage II: Regression outlier detection

Detection method	The abnormal association detection method described in Subsection 3.2 of the manuscript
Type of outliers	Type II
Argument setting	Same as those listed in Table S-7 except that the whisker parameters $(r_2^{\text{LQD}}, r_2^{\text{B}})$ are set to $(3.0, 3.0)$

Table S-12

Argument settings for the standard and robust LQD-RKHS distributional regression methods

Common arguments (for both the standard and robust LQD-RKHS methods)

Operator kernel K_r in Eq. (4) (or Eq. (6))	The Gaussian kernel given in Eq. (8)
σ in Eq. (8)	Determined according to Eq.(S-43) of this document
α_{mix}^{LQD} in Eq. (2)	$\alpha_{mix}^{LQD} = 0.3$
Truncation order of FPCA	$m = 5$
Regularization parameter λ_s	Adaptively selected using the generalized cross-validation procedure described in Appendix 1 of this document

Additional arguments for the robust LQD-RKHS method

Argument settings for Type I and Type II outlier detections	The same as those in Table S-11
Weights w_i s in Eq. (6)	Determined according to the principles described in Subsection S.4.1 of this document, and the relevant parameters ρ_1 and ρ_2 in the weight functions given in Eqs. (S-22) and (S-25) are set to 1

S.7.4. Additional discussion on the detected Type II outliers

The outlying PDFs detected in the second stage (i.e., the Type II outliers) are visualized in Figure 5 (b) of the manuscript as bold colored curves. Recall that the Type II outliers belong to the regression outliers, which correspond to the abnormal associations of the PDF-valued two-tuples. The bold curves in the same color shown in Figure 5 (b) of the manuscript belong to the same two-tuple. The anomaly (with respect to the majority of the data) of the detected Type II outlier can only be stand out when viewed in pair. For comparison purposes, Figure S-31 also displays the curves w.r.t. the bulk of the curves for six selected PDF pairs with normal associations. Comparing the normal PDF pairs in Figure S-31, one can see that their horizontal positions are correlated with each other. Obviously, the three detected Type II outlying PDFs shown in Figure 5 (b) of the manuscript violate the correlation pattern exhibited in Figure S-31.

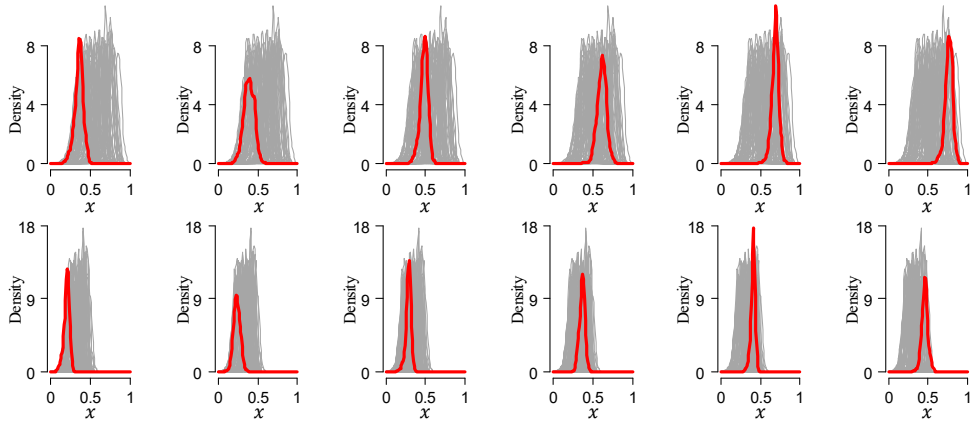


Figure S-31. Comparisons for six pairs of PDFs (represented by bold red curves) with normal associations. The first row corresponds to Sensor A, and the second row corresponds to Sensor B.

S.7.5. Sensitivity analysis

This subsection provides an in-depth comparative study to check whether the results obtained by our proposed robust LQD-RKHS distributional regression method are still relatively robust

(compared to the standard LQD-RKHS method) when using different data segmentation schemes.

In the former study, the post-processed strain monitoring data from the two sensors are divided into 120 segments to obtain 120 pairs of PDF-valued samples. If we choose a different data segmentation scheme, i.e., the number of segments is not 120, it would lead to a new PDF-valued dataset with different sample sizes and different curve shapes. We are curious to see how the proposed robust LQD-RKHS method would perform under the new data segmentation scheme and whether it is still relatively robust (compared to the standard method). Recall that the robustness of the proposed robust regression method is achieved by downweighting the detected outlying PDFs contained in the training samples; thus, the main influential factor of the robustness of our method is whether the outlying PDFs (affecting the regression model adversely) can be successfully detected by the proposed distributional outlier detection methods. The latter (i.e., outlier detection methods) are not sensitive to the sample size of the distributional data; thus, the proposed robust distributional regression method is also expected to be insensitive to the number of data segments. In the following, we will conduct a comparative study to check this.

Table S-13

The six considered data segmentation schemes

Segmentation scheme	Scheme 1	Scheme 2	Scheme 3	Scheme 4	Scheme 5	Scheme 6
Number of segments	60	90	120	150	200	600

We consider six different data segmentation schemes (listed in Table S-13) with 60, 90, 120, 150, 200 and 600 segments. Then, we can obtain six different PDF-valued datasets consisting of PDF two-tuples, denoted as $\{\hat{g}_{1,i}, \hat{f}_{1,i}\}_{i=1}^{60}$, $\{\hat{g}_{2,i}, \hat{f}_{2,i}\}_{i=1}^{90}$, $\{\hat{g}_{3,i}, \hat{f}_{3,i}\}_{i=1}^{120}$, $\{\hat{g}_{4,i}, \hat{f}_{4,i}\}_{i=1}^{150}$, $\{\hat{g}_{5,i}, \hat{f}_{5,i}\}_{i=1}^{200}$ and $\{\hat{g}_{6,i}, \hat{f}_{6,i}\}_{i=1}^{600}$, associated with the six data segmentation schemes. Next, we take the first data segmentation scheme (i.e., Scheme 1) as an example to detail the remaining test/training data selection as well as the associated distributional regression analysis. The data processing and analysis for other schemes are conducted in a similar way. For convenience, let N_1 denote the sample size of the resulting PDF-valued dataset associated with Scheme 1. After implementing the initial distributional outlier detection, a total of $[0.3N_1]$ (the factor 0.3 is fixed for Schemes 1~6) pairs of PDFs are randomly selected from the “good” dataset (i.e., the PDF-valued dataset after removing the outlying PDFs detected in the initial outlier detection stage) to serve as the test functional samples, and the remaining data (including the detected outliers) are used as the training functional samples. This test/training data selection is independently repeated five times, yielding five different groups of test/training datasets (which are referred to as Groups 1~5). Then, we separately use the standard and robust LQD-RKHS distributional regression methods to predict the PDFs (associated with Sensor B) in the test datasets. The regularization parameters involved in the regression models are adaptively selected by using the generalized cross-validation method described in Appendix 1. The prediction error for each test PDF is also quantified by the integrated

absolute error (IAE) similar to that in the manuscript. The boxplots of the prediction errors associated with each test/training group are displayed in Figure S-32 for all six considered data segmentation schemes. Comparing the results in Figure S-32, one can see that the proposed method significantly outperforms the standard method in all the considered test/training groups associated with the six segmentation schemes. Since the proposed method can maintain its advantage of robustness in a wide range of data segmentation schemes, we can conclude that the robustness of our method is insensitive to the number of data segments.

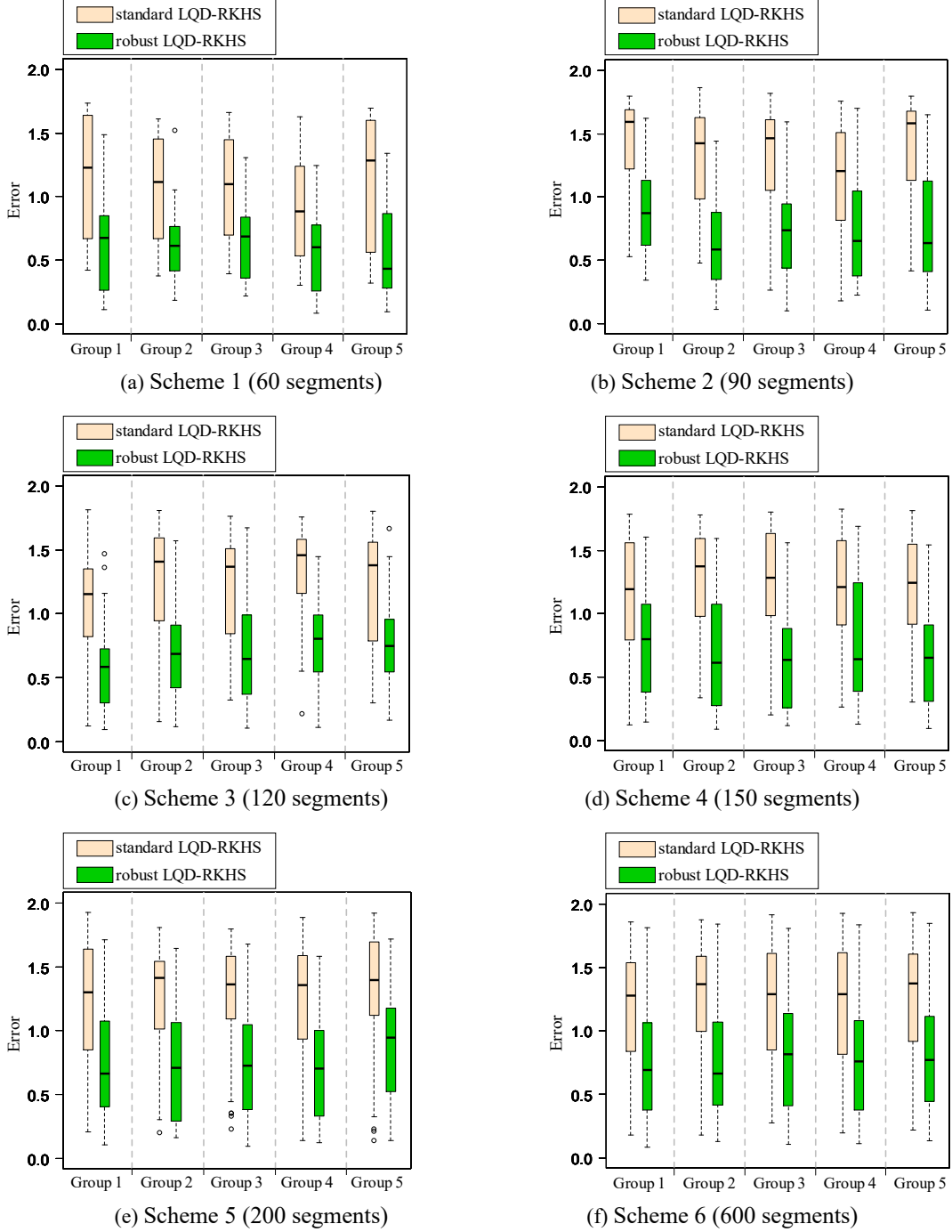
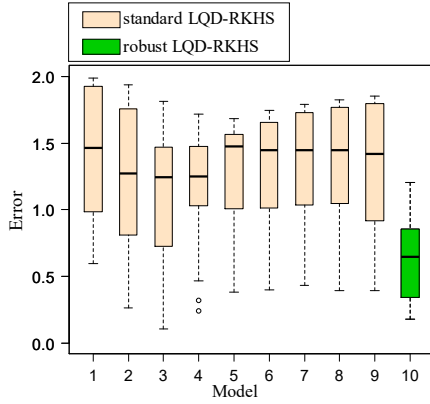
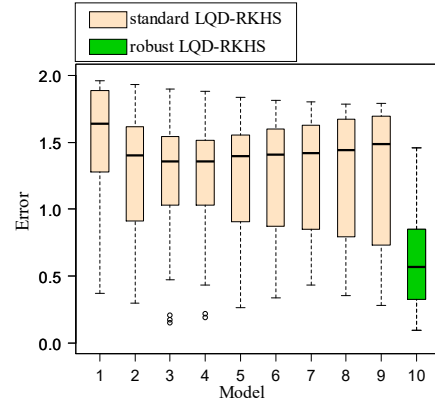


Figure S-32. Comparisons of the prediction errors of the standard and robust LQD-RKHS distributional regression models for different data segmentation schemes. The regularization parameters are adaptively selected by the generalized cross-validation method.

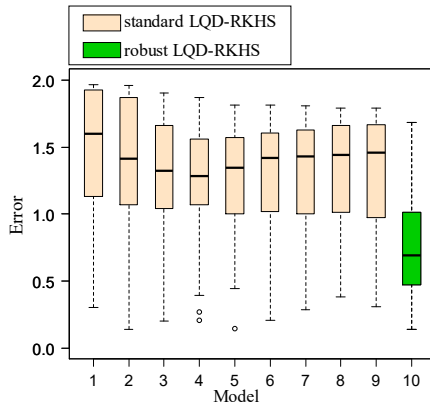
Remark. It is worth noting that the regularization parameter λ_s has a significant impact on RKHS-based regression. In the comparative study conducted above, such regularization parameters are adaptively selected by the generalized cross-validation method, and the resulting parameters can be different for the standard and robust LQD-RKHS methods (as the GCV procedure for the robust method considers removing the detected outlying PDFs). Consequently, one may argue that the better performance of the robust method is probably attributed to a better regularization parameter. To settle this doubt, we provide an in-depth comparative study. To avoid the strength of the standard version being suppressed by a suboptimal regularization parameter, we consider nine different regularization parameters, valued at 0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 4.0, and 8.0 (the value of the regularization parameter selected using the GCV procedure falls into the range of these considered parameters), for the standard LQD-RKHS method. The regularization parameter for the robust version is fixed at 0.1. Consequently, 10 different regression models can be constructed. We also consider six different data segmentation schemes, as listed in Table S-13, and for each scheme, we consider one group of test/training data (independently reselected in a similar way as described above). The prediction errors associated with the ten regression models are summarized as boxplots as shown in Figure S-33. The proposed robust method also significantly outperforms the standard method under all considered regularization parameters.



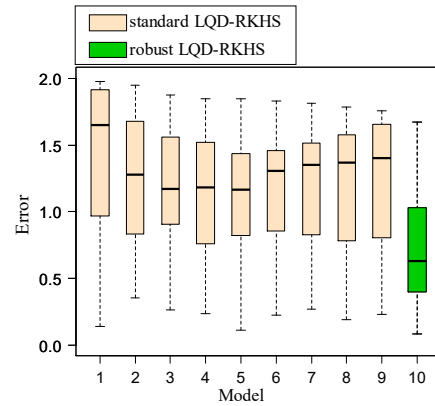
(a) Scheme 1 (60 segments)



(b) Scheme 2 (90 segments)



(c) Scheme 3 (120 segments)



(d) Scheme 4 (150 segments)

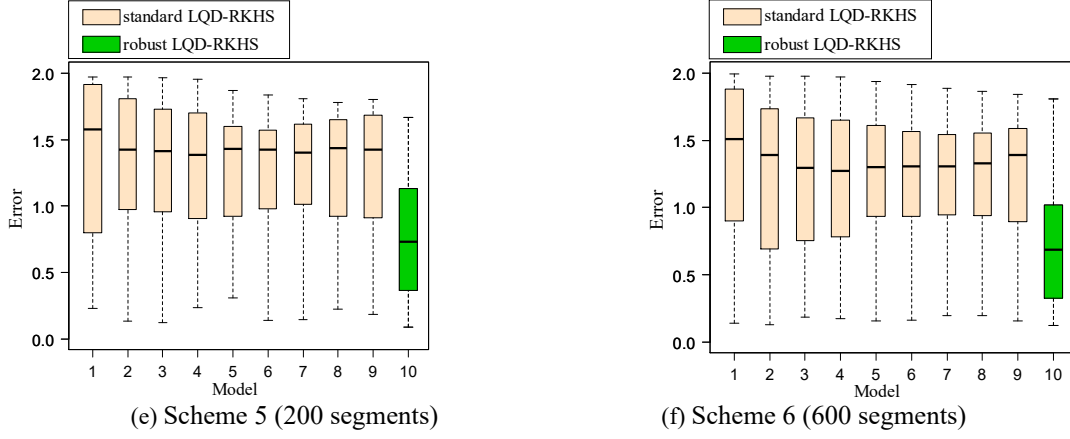


Figure S-33. Comparisons of the prediction errors of the standard and robust LQD-RKHS distributional regression models for different data segmentation schemes. Models 1~9 correspond to the standard LQD-RKHS method with $\lambda_s = 0.001, 0.01, 0.05, 0.1, 0.5, 1.0, 2.0, 4.0$ and 8.0 , respectively, and Model 10 corresponds to the robust LQD-RKHS method with $\lambda_s = 0.1$.

S.7.6. Validity validation for the reconstructed distributions

This subsection provides statistical analysis tests to examine the validities of the reconstructed distributions. The distributional regression-based data reconstruction method uses the information of the cross-correlation (of the PDF-valued data) between the two sensors to reconstruct the distributions of the missing data. Therefore, the validity of the reconstructed distributions depends on whether the two PDF-valued datasets associated with the two sensors are correlated or not. Generally, the higher degree of correlation between the two distributional datasets associated with the two sensors is, the higher reliability of the reconstructed distributions that can be achieved. If the two distributional datasets are independent or weakly correlated, the reconstructed distributions obtained by the distributional regression method can be regarded as invalid.

For scalar data, the correlation between two datasets $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ can be easily discovered by observing the pattern of the scatter plot of the data points $\{(x_i, y_i)\}_{i=1}^n$. However, checking whether two PDF-valued datasets are correlated is not straightforward. One effective strategy is to check whether the FPC scores of the curves in one dataset are dependent on the curves of the other dataset. Recently, Chen et al. (2020) proposed an approach for scalar-on-function dependence analysis. Moreover, they also defined a coefficient to quantify the strength of such a dependence. Here, we employ a similar strategy as that in Chen et al. (2020) to examine the distributional correlation.

Recall that in the case study conducted in Section 6 of the manuscript, we have obtained 120 pairs of PDFs denoted as $\{\hat{g}_i(x), \hat{f}_i(x)\}_{i=1}^{120}$, and 97 pairs are classified into the “good” dataset (after the initial outlier detection). Here, only the “good” data are used to examine the distributional correlation to avoid distorting the analysis results by the outlying PDFs. For convenience, the 97 pairs of “good” PDFs are denoted as $\{\hat{g}_k(x), \hat{f}_k(x)\}_{k=1}^N$ ($N = 97$), and their corresponding LQD

representations are denoted as $\{\psi_k^{g*}(t), \psi_k^{f*}(t)\}_{k=1}^N$ (computed using Eq. (2) in the manuscript with $\alpha_{mix}^{LQD} = 0.3$). Let $\{\xi_{k,j}^{f*}\}_{k=1}^N$ be the FPC scores of the functional data $\{\psi_k^{f*}(t)\}_{k=1}^N$ associated with the j th FPC. The issue of concern is checking whether the real-valued data $\{\xi_{k,j}^{f*}\}_{k=1}^N$ (associated with Sensor B) is dependent on the functional data $\{\psi_k^{g*}(t)\}_{k=1}^N$ (associated with Sensor A) or not; if the dependence exists, how strong is it? Such a scalar-on-function dependence can be analyzed by using the regression-based procedure described in Subsection 2.3 of Chen et al. (2020). The main idea of such an approach is first to build an RKHS-based nonlinear function-to-real regression model to capture the dependent relationship between the two datasets, and then the strength of the dependence can be quantified based on the amount of variability of the data explained by the regression model. Specifically, let $\xi_{k,j}^{f*} = Q_j(\psi_k^{g*}(t)) + \varepsilon_{k,j}$, $k = 1, 2, \dots, N$ denote the RKHS-based nonlinear regression model for relating $\{\psi_k^{g*}(t)\}_{k=1}^N$ to $\{\xi_{k,j}^{f*}\}_{k=1}^N$ (the FPC scores of $\{\psi_k^{f*}(t)\}_{k=1}^N$ on the j th FPC). The regularization parameter involved in the RKHS-based regression model is determined by using the generalized cross-validation (GCV) procedure described in Appendix 3 of Chen et al. (2020). The regularization parameter selected in such a way not only effectively avoids the regression model overfitting the data but also helps the regression model to reliably capture the underlying dependence of the data (see Section 4 in Chen et al. (2020) for the relevant validations). The fitted results of $\{\xi_{k,j}^{f*}\}_{k=1}^N$ obtained by the estimated regression model are denoted as $\hat{\xi}_{k,j}^{f*} = \hat{Q}_j(\psi_k^{g*}(t))$, $k = 1, 2, \dots, N$; then, the strength of dependence between $\{\xi_{k,j}^{f*}\}_{k=1}^N$ and $\{\psi_k^{g*}(t)\}_{k=1}^N$ can be quantified by using the following coefficient (Chen et al. 2020):

$$r_j = \frac{\text{std}(\{\hat{\xi}_{k,j}^{f*}\})}{\text{std}(\{\xi_{k,j}^{f*}\})} = \frac{\sqrt{\sum_{k=1}^N (\hat{\xi}_{k,j}^{f*} - \text{mean}(\{\hat{\xi}_{k,j}^{f*}\}))^2}}{\sqrt{\sum_{k=1}^N (\xi_{k,j}^{f*} - \text{mean}(\{\xi_{k,j}^{f*}\}))^2}} \quad (\text{S-46})$$

where $\text{std}(\{\hat{\xi}_{k,j}^{f*}\})$ and $\text{mean}(\{\hat{\xi}_{k,j}^{f*}\})$ represent the sample standard deviation and sample mean of the dataset $\{\hat{\xi}_{k,j}^{f*}\}_{k=1}^N$, respectively. To help readers better understand the mechanism of such an approach in scalar-on-function dependence quantification, we also provide a simple example using only scalar data to illustrate how the method works (the case of scalar-on-function dependence encountered in this case study is analogous) in Appendix 5.

Taking $j = 2$ (corresponding to the second FPC) as an example, the FPC scores denoted as $\{\xi_{k,2}^{f*}\}_{k=1}^N$ along with their fitted values (obtained by the regression model and denoted as $\{\hat{\xi}_{k,2}^{f*}\}_{k=1}^N$) are plotted in the same figure as shown in Figure S-34; one can see that the variation of

the data $\{\xi_{k,2}^{f*}\}_{k=1}^N$ can be largely explained by the optimal regression model determined by the GCV procedure, indicating that the real-valued data $\{\xi_{k,2}^{f*}\}_{k=1}^N$ are strongly dependent on the functional data $\{\psi_k^g(t)\}_{i=1}^N$; the strength of dependence computed using Eq. (S-46) is 0.797. If we randomly disorder the data in $\{\xi_{k,2}^{f*}\}_{k=1}^N$, then the dependence disappears. After re-fitting the regression model using the disordered data, the results are displayed in Figure S-35; obviously, the regression model cannot “explain” any variation of the data, and the strength of dependence computed using Eq. (S-46) becomes 0.000.

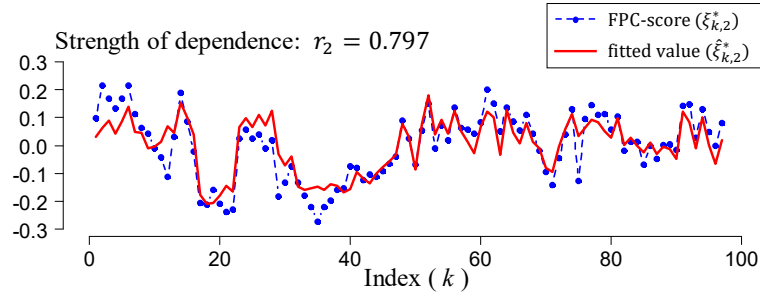


Figure S-34. Visualization of the fitting result for the second FPC.

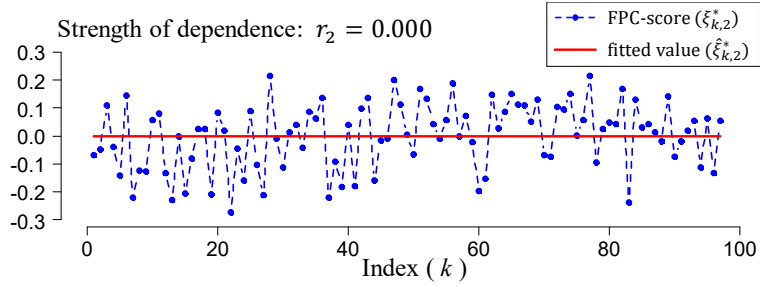


Figure S-35. The fitting result of the regression model estimated using the disordered FPC scores of the second FPC.

For the FPC scores associated with other FPCs, the scalar-on-function dependence can be analyzed in a similar way, and the results are displayed in Figure S-36 for the 5 considered FPCs (including the second FPC investigated earlier). From the results in Figure S-36, one can see that the FPC scores of the LQD representations of the PDFs $\{\hat{f}_k\}_{k=1}^N$ from Sensor B are highly correlated with the LQD-representations of the PDFs $\{\hat{g}_k\}_{k=1}^N$ from Sensor A, indicating that the distributions of the strain monitoring data collected by the two sensors are highly correlated; thus, it verifies the validity of the reconstructed distributions using the inter-sensor distributional correlation information.

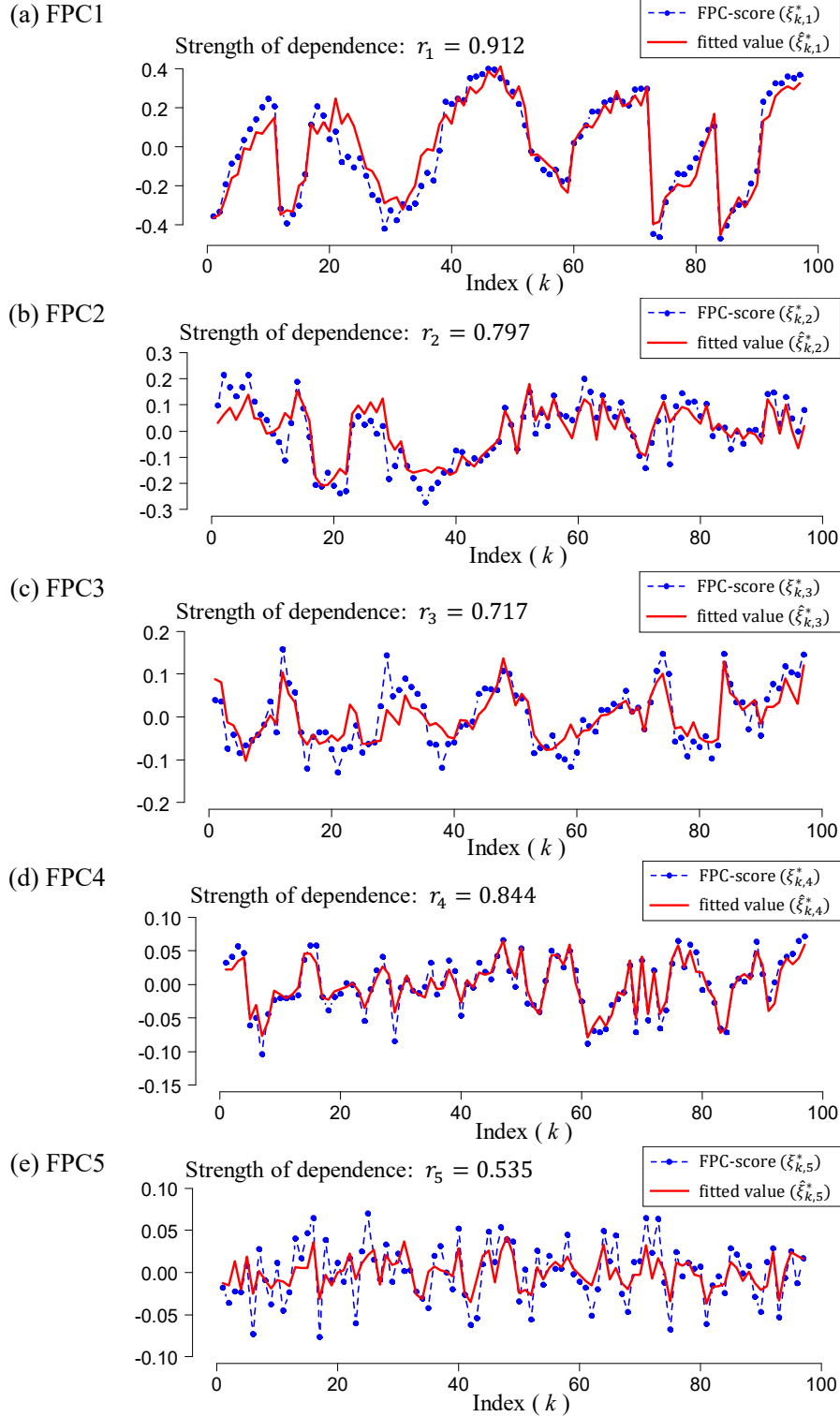


Figure S-36. Visualization of the fitting results for the first five FPCs.

Here, the data-on-function dependence analyses for different FPCs are performed separately, which enables us to directly check the dependence by plotting the real-valued data sequence along with the corresponding fitted sequence (obtained by the regression model) in the same figure. In this setting, the data-on-function dependence can be easily discovered through visual inspection. However, in the proposed robust distribution-to-distribution regression method, the FPC scores

associated with the same functional sample are considered as a whole; otherwise, the integrity of the functional data might be destroyed.

S.7.7. Non-random missing case study

In the real data study conducted in Section 6 of the manuscript, the training/test data are selected using a random split scheme, which corresponds to the case of missing at random. Here, we provide an additional case study to consider the non-random missing case. According to the initial outlying PDF detection conducted in Section 6 of the manuscript, among the 120 pairs of PDFs denoted as $\{\hat{g}_i, \hat{f}_i\}_{i=1}^{120}$, 23 pairs are classified into the abnormal dataset, and the remaining 97 pairs are classified into the “good” dataset. For convenience, we denote the “good” dataset as $\{(\hat{g}_k^{good}, \hat{f}_k^{good}): k = 1, 2, \dots, 97\}$. In this additional case study, a portion of the “good” PDFs from $\{\hat{f}_k^{good}: k = 1, 2, \dots, 97\}$ (i.e., the “good” PDFs associated with Sensor B) are assumed to be consecutively missing, and the following two scenarios are considered:

Scenario I: 20 consecutive “good” PDFs $\{\hat{f}_k^{good}: k = 50, 51, \dots, 69\}$ are assumed to be missing;

Scenario II: 30 consecutive “good” PDFs $\{\hat{f}_k^{good}: k = 50, 51, \dots, 79\}$ are assumed to be missing.

In Scenario I, the remaining 100 pairs of PDFs (including the 23 pairs of abnormal PDFs) are used as the training dataset. Similarly, the training dataset in Scenario II is composed of the remaining 90 pairs of PDFs.

Both the standard and robust LQD-RKHS distributional regression methods are considered for reconstructing the missing PDFs. The argument settings for these two methods are the same with their counterparts in the real data study conducted in Section 6 of the manuscript. The comparisons of the reconstructed PDFs for Scenario I and Scenario II are presented in Figure S-37 and Figure S-38, respectively. The results show that the proposed robust LQD-RKHS distributional regression method works well in the two considered consecutively missing scenarios, and its performance is much better than the standard LQD-RKHS method.

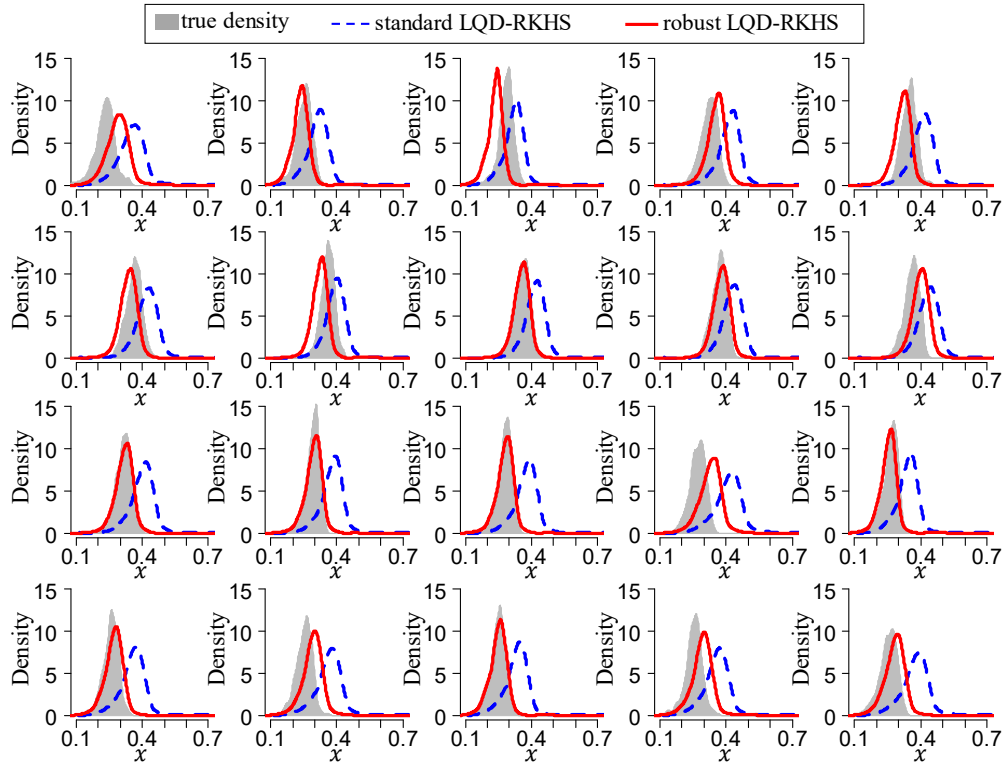


Figure S-37. Comparisons of the reconstruction results for the 20 consecutively missing PDFs in Scenario I.

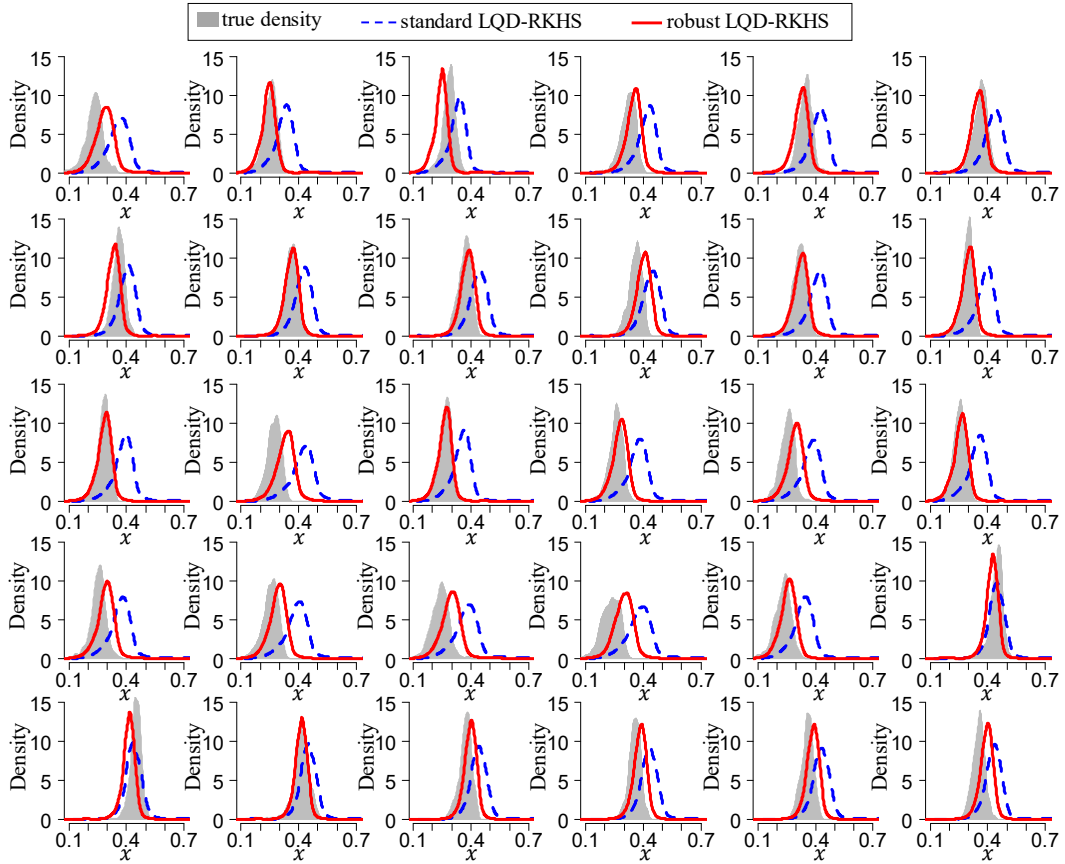


Figure S-38. Comparisons of the reconstruction results for the 30 consecutively missing PDFs in Scenario II.

S.7.8. Discussion on the practical utility of distribution reconstruction in SHM applications

Structural health monitoring (SHM) systems operate in complex environments, data loss and contamination are common and unavoidable under existing technologies. Consequently, data quality is of crucial concern in subsequent data modeling, data analysis and data mining involved in various SHM applications. As an important measure for improving the quality of SHM data, sensor faults correction concerning of correcting the information of distorted or corrupted data (including missing data) (Yi et al. 2017) has become an important research topic in the field of SHM. The robust distributional regression method developed in this study has wide applications in sensor faults correction.

Firstly, the distributional regression-based data reconstruction method has good potential in extreme value analysis of load effects (of bridge structures) under data contamination (Chen et al. 2018). The strain response investigated in the real data study is one of the most important load effects of bridge structures. The extreme value distribution of strain responses is of crucial importance in bridge safety condition evaluation, reliability analysis as well as performance-based design. If a large portion of the strain monitoring data are missing or severely contaminated (such as the case shown in Figure 4 of the manuscript), the extracted extreme value samples (such as the block maxima values, the samples exceeding a high threshold (OBrien et al. 2015; Coles 2001)) for extreme value distribution modeling would be scarce or distorted. Consequently, the built extreme value models would become less reliable and even severely distorted, leading to unreliable or misleading analysis results in subsequent bridge safety evaluation or reliability analysis. If we have reliable statistical methods to reconstruct (or correct) the distributions of the missing (or contaminated) data, we can use the reconstructed (or corrected) distributions to generate the same number of samples to impute the missing (or contaminated) data; then, we can use the post-processed data (after data imputation) to extract the extreme value samples for extreme value distribution modeling and relevant extreme value analysis or prediction, see Chen et al. (2018) for more details. It is worth noting that the extreme value prediction is based on the distributional information of the samples taking large values, therefore the imputed data are required to follow a similar distribution of the true missing data. In this sense, the proposed robust distributional regression method has great potential in recovering the distributional information of missing data.

On the other hand, it has long been recognized in the statistical community that the imputation model for missing data should account for the underlying distribution as well as the uncertainty of missing data (Murray 2018), such requirements are also of great importance for addressing the missing data in the SHM field. Recently, Chen et al. (2019b) proposed a two-stage imputation framework for missing SHM data by combining distributional regression with the copula-based imputation method that can naturally meet these requirements. In this framework, the density functions of missing data are firstly reconstructed by a distributional regression method; then, in

consideration of uncertainty, the missing measurements are imputed by random samples generated from a conditional distribution model (constructed based on the nonparametric copula and the reconstructed distributions). A notable advantage of such a strategy is that the imputed data can approximately follow the underlying distribution of the lost data. The distribution construction technique plays a crucial role in such applications. Moreover, the reconstructed distributions can further be used to correct the contaminated time series of the strain monitoring data using the copula-based imputation approach, see Chen et al. (2019b) for more details.

Appendix 1: Adaptive regularization parameter selection for the LQD-RKHS distributional regression model

The regression operator in the LQD-RKHS distribution-to-distribution regression (DtDR) model is estimated by solving a regularized least-squares problem, and the regularization parameter λ_s can be adaptively selected by the generalized cross-validation (GCV) method (Golub et al. 1979; Wahba 1990; Lian 2007a).

According to the theory of the GCV (Wahba 1990; Lian 2007a), the regularization parameter for the LQD-RKHS DtDR model can be chosen as follows:

$$\lambda_{GCV} = \underset{\lambda_s > 0}{\operatorname{argmin}} \frac{\frac{1}{n} \|(I - A(\lambda_s)) \operatorname{vec}(Y)\|_2^2}{\left[\frac{1}{n} \operatorname{trace}(I - A(\lambda_s)) \right]^2} \quad (\text{S-47})$$

$$\text{with } A(\lambda_s) := (K \otimes A)[(K \otimes A) + \lambda_s I]^{-1}$$

where A , Y and $K = I_{m \times m}$ are matrices corresponding to those in Eq. (24) of Chen et al. (2019a), I is an identity matrix of size $mn \times mn$ (m is the FPC order in the LQD-RKHS DtDR model and n is the number of training PDFs). $\text{GCV}(\lambda_s) = \frac{\frac{1}{n} \|(I - A(\lambda_s)) \operatorname{vec}(Y)\|_2^2}{\left[\frac{1}{n} \operatorname{trace}(I - A(\lambda_s)) \right]^2}$ is also called the GCV statistic.

Similar to Lian (2007a), we can compute the GCV statistic at a pre-specified grid of λ_s denoted as $\mathcal{C}(\lambda) = \{\lambda_{s,1}, \dots, \lambda_{s,N_\lambda}\}$, then the regularization parameter can be effectively estimated as

$$\lambda_{GCV} = \underset{\lambda_s \in \mathcal{C}(\lambda)}{\operatorname{argmin}} \{ \text{GCV}(\lambda_{s,1}), \dots, \text{GCV}(\lambda_{s,N_\lambda}) \} \quad (\text{S-48})$$

For the robust LQD-RKHS distributional regression model developed in Section 4 of the manuscript, the regularization parameter λ_s can be automatically determined by a similar way. The only difference is that we recommend to remove the detected outlying PDFs from the training dataset before implementing the GCV procedure.

In addition to the GCV method, the regularization parameter involved in the RKHS-based regression method can also be adaptively selected by using the ordinary leave-one-out cross-validation procedure as summarized in Appendix 3 of Chen et al. (2019a). Compared to the GCV procedure described above, the leave-one-out cross-validation method is much more computationally intensive.

Appendix 2: Basic outlier generation algorithm

Algorithm S.7: $(S_{PDF}, IDE) = \text{PDFoutlier_Insert}(S_{PDF}, N_o, \zeta_{hs}, \varpi)$

Input: PDF-valued dataset $S_{PDF} = \{f_i\}_{i=1}^n$, number of outliers N_o , and coefficients ζ_{hs} and ϖ

Output: The contaminated PDF-valued dataset S_{PDF} , and the index set of outliers denoted as IDE

1: Calculate the modes of the n PDFs $\{f_i\}_{i=1}^n$, respectively, and denotes the set of modes as

$$\Pi = \{\pi_i\}_{i=1}^n \text{ with } \pi_i = \underset{x}{\operatorname{argmax}} f_i(x)$$

2: Construct two subsets of PDFs

$$U = \{f_i \in S_{PDF} | \pi_i > q_{1-\varpi}(\Pi)\}, L = \{f_i \in S_{PDF} | \pi_i < q_{\varpi}(\Pi)\}$$

where $q_{\varpi}(\Pi)$ stands for the (100ϖ) th percentile of the dataset Π

3: Set $\Xi = \{1, 2, \dots, n\}$ and $IDE = \emptyset$

4: **for** $i = 1$ **to** N_o **do**

a: Generate $z \sim U(0, 1)$

b: **if** $z > \zeta_{hs}$ **then** (*generate shape outlier*)

Randomly select one PDF from U , and denote it as $h_1 \in U$

Randomly select one PDF from L , and denote it as $h_2 \in L$

Compute $h(x) = qh_1(x) + (1 - q)h_2(x)$ with $q \sim U(0.4, 0.6)$

else (*generate horizontal-shift outlier*)

Generate $y \sim U(0, 1)$, $a \sim U(2, 5)$, $b \sim U(13, 16)$, $c \sim U(17, 22)$ and $d \sim U(2, 5)$

Compute $h(x) = \text{BetaPdf}(x; a, b) \cdot \chi_{\{y > 0.5\}}(y) + \text{BetaPdf}(x; c, d) \cdot \chi_{\{y \leq 0.5\}}(y)$

end if

c: Randomly select an element from Ξ denoted as k_i

d: Perform PDF replacement $S_{PDF}[k_i] \leftarrow h$

e: Set $\Xi = \Xi \setminus \{k_i\}$, and put k_i into the index set $IDE \leftarrow k_i$

end for

5: Output S_{PDF} and the index set $IDE = \{k_1, k_2, \dots, k_{N_o}\}$

Appendix 3: Abnormal association-generating process for simulation study II

The PDF-valued two-tuples simulated by Algorithm S.3 in simulation study II can be equally written as the following structured data:

$$\left\{ \begin{matrix} g_1 \\ f_1 \end{matrix} \right\}, \dots, \left\{ \begin{matrix} g_j \\ f_j \end{matrix} \right\}, \dots, \left\{ \begin{matrix} g_k \\ f_k \end{matrix} \right\}, \dots, \left\{ \begin{matrix} g_n \\ f_n \end{matrix} \right\}$$

Clearly, f_i depends on g_i since their corresponding distributional parameters obey the relationship given in Line c of Algorithm S.3. If we disorder the matches of the PDF two-tuples via performing an intra-set element exchange as illustrated in Figure S-39, it can produce two abnormal associations of PDFs. Such a strategy can be used to simulate the abnormal associations for validating the effectiveness of the distributional regression-based outlier detection method. Algorithm S.8 presents the implementation of element exchange for single dataset based on the peak information of PDFs, Algorithm S.9 details the final implementation of the abnormal association generation for contaminating the PDF-valued two-tuples.

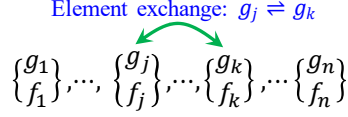


Figure S-39. Illustration of intra-set element exchange

Algorithm S.8: $(S_{PDF}, IDE) = \text{Element_Exchange}(S_{PDF}, M)$

Input: PDF-valued dataset S_{PDF} consisting of n PDFs denoted as $\{g_i\}_{i=1}^n$, and a parameter M for controlling the number of element exchanges in S_{PDF}

Output: S_{PDF} after element exchange, and the index set denoted as IDE for locating the places where the element exchanges occur

1: Construct the peak-value set

$$\Pi^g = \{\pi_1^g, \dots, \pi_n^g\} \text{ with } \pi_i^g = \sup_{x \in [0,1]} g_i(x), i = 1, \dots, n$$

2: Construct PDF subsets

$$S_1 = \{g_i \in S_{PDF} \mid \pi_i^g \leq q_{0.2}(\Pi^g)\}, S_2 = \{g_i \in S_{PDF} \mid \pi_i^g \geq q_{0.8}(\Pi^g)\}$$

where $q_{0.2}(\Pi^g)$ and $q_{0.8}(\Pi^g)$ are the 20th and 80th percentiles of the dataset Π^g

3: Randomly select M elements from S_1 , and denote the selected PDFs as

$$H_L = \{g_{l_1}, g_{l_2}, \dots, g_{l_M}\} \text{ with } l_1, l_2, \dots, l_M \text{ being the curve indices in } S_{PDF}$$

4: Randomly select M elements from S_2 , and denote the selected PDFs as

$$H_U = \{g_{u_1}, g_{u_2}, \dots, g_{u_M}\} \text{ with } u_1, u_2, \dots, u_M \text{ being the curve indices in } S_{PDF}$$

5: Perform element exchange in S_{PDF}

$$g_{l_1} \rightleftharpoons g_{u_1}, g_{l_2} \rightleftharpoons g_{u_2}, \dots, g_{l_M} \rightleftharpoons g_{u_M}$$

6: Output the element-exchanged PDF dataset S_{PDF} along with the index set $IDE = \{l_1, \dots, l_M, u_1, \dots, u_M\}$

Algorithm S.9: Generating abnormal PDF associations by exchanging elements

Input: PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$, the bivariate parameter (M_g, M_f) for controlling the number of element exchanges in $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$

Output: the contaminated PDF-valued two-tuples $\{g_i, f_i\}_{i=1}^n$

1: Denote the PDF-valued datasets $\{g_i\}_{i=1}^n$ and $\{f_i\}_{i=1}^n$ as S_{PDF}^g and S_{PDF}^f , respectively

2: **repeat**

a: Perform element exchange on S_{PDF}^g using Algorithm S.8

$$(S_{PDF}^g, IDE_g) = \text{Element_Exchange}(S_{PDF}^g, M_g)$$

b: Perform element exchange on S_{PDF}^f using Algorithm S.8

$$(S_{PDF}^f, IDE_f) = \text{Element_Exchange}(S_{PDF}^f, M_f)$$

until $IDE_g \cap IDE_f = \emptyset$ (\emptyset denotes the empty set)

3: Output the contaminated PDF-valued two-tuples, i.e.,

$$\{g_i, f_i\}, g_i \in S_{PDF}^g, f_i \in S_{PDF}^f, i = 1, 2, \dots, n$$

Appendix 4: Two considered competitors originally for ordinary functional outlier detection

(1) Competing method I: the functional directional outlyingness (FDO)-based method

Computational details

The functional directional outlyingness (FDO) was defined by Dai and Genton (2019), then it has been successfully applied in outlier detection for ordinary functional data with potential for uncovering magnitude outliers and shape outliers simultaneously. However, according to our test, directly performing outlier detection in the PDF space using the FDO tool usually results in poor performance (related illustrations will be given latter). To remedy this embarrassing, we take a strategy of applying the FDO tool in the quantile function (QF) space (i.e., the QF node in the transformation tree) for outlier detection, which is referred to as the QF-FDO method throughout this study. As illustrated in Figures 3(a) and (b) of the manuscript, after converting the PDFs to quantile functions, the horizontal-shift outlying PDFs have become the magnitude outliers, while the shape outliers are still hidden in the bulk of the data. In contract to the disordered PDFs, the quantile functions are in a much more organized pattern as they have been registered according to quantiles.

Consider a functional dataset denoted as $\{f_i(x)\}_{i=1}^n$ with each element being a PDF defined on the compact interval $[0, 1]$, let $Q_i(t)$ be the quantile function corresponding to $f_i(x)$. Also, let $\text{median}_{1 \leq k \leq n}\{a_k\}$ denote the sample median of the real-valued dataset $\{a_i\}_{i=1}^n, a_i \in \mathbb{R}$. Note that when fixing t at $t = t_0$, the set of $\{Q_i(t_0)\}_{i=1}^n$ is a real-valued dataset. Then, the pointwise outlyingness of $Q_i(t)$ at $t = t_0$ can be calculated by (Dai and Genton 2019)

$$\text{dSDO}(Q_i(t_0)) = \frac{Q_i(t_0) - \text{median}_{1 \leq k \leq n}\{Q_k(t_0)\}}{\text{MAD}(Q(t_0))} \quad (\text{S-49})$$

where $\text{MAD}(Q(t_0))$ is the median absolute deviation (MAD) calculated by $\text{MAD}(Q(t_0)) = c \cdot \text{median}_{1 \leq k \leq n}\left\{\left|Q_k(t_0) - \text{median}_{1 \leq j \leq n}\{Q_j(t_0)\}\right|\right\}$ with c being a constant (we set c to be its default value 1.4826 throughout this study).

Based on the FDO theory (Dai and Genton 2019), the magnitude and shape anomalies of a given quantile function $Q_i(t)$ can be respectively measured by the mean outlyingness (MO) and the variation of outlyingness (VO) calculated as follows:

$$\begin{aligned} \text{MO}(Q_i) &= \int_0^1 \text{dSDO}(Q_i(t)) \omega(t) dt \\ \text{VO}(Q_i) &= \int_0^1 |\text{dSDO}(Q_i(t)) - \text{MO}(Q_i)|^2 \omega(t) dt \end{aligned} \quad (\text{S-50})$$

where $\omega(t)$ is the weight function, which is commonly chosen as $\omega(t) = 1/\lambda(I)$ with $\lambda(I)$

being the Lebesgue measure of the quantile function's domain of definition denoted as I .

Generally, the magnitude outliers and shape outliers in the collection of quantile functions would stand out as MO-outliers (abnormal in the MO direction) and VO-outliers (abnormal in the VO direction), respectively. The MO- and VO-outliers can be efficiently identified by the two- and one-sided boxplot-based detectors given in Eqs. (S-14) and (S-15), respectively.

A small simulation study

We conduct a simulation study to examine the detection performance of the FDO-based method. To begin with, we generate a PDF-valued dataset composed of 100 functions by using Algorithm S.10 with the input arguments setting to $n = 100$, $\delta_1 = 36$ and $\delta_2 = 63$. Then, we use Algorithm S.7 to generate and insert 5 outlying PDFs into the simulated functional dataset by setting $N_o = 5$, $\zeta_{hs} = 0$, and $\varpi = 0.2$. Let $S_{\text{contam}} = \{f_i\}_{i=1}^n$ denote the contaminated PDF dataset, then we use it to produce the following four different PDF datasets:

Model I: $S_{\text{contam}}^{\text{I}} = \{f_i^{\text{I}}\}_{i=1}^n$ with $f_i^{\text{I}} = f_i$, $f_i \in S_{\text{contam}}$

Model II: $S_{\text{contam}}^{\text{II}} = \{f_i^{\text{II}}\}_{i=1}^n$ with $f_i^{\text{II}} = 0.9 * f_i + 0.1$, $f_i \in S_{\text{contam}}$

Model III: $S_{\text{contam}}^{\text{III}} = \{f_i^{\text{III}}\}_{i=1}^n$ with $f_i^{\text{III}} = 0.7 * f_i + 0.3$, $f_i \in S_{\text{contam}}$

Model IV: $S_{\text{contam}}^{\text{IV}} = \{f_i^{\text{IV}}\}_{i=1}^n$ with $f_i^{\text{IV}} = 0.5 * f_i + 0.5$, $f_i \in S_{\text{contam}}$

Figure S-40 displays the four simulated datasets.

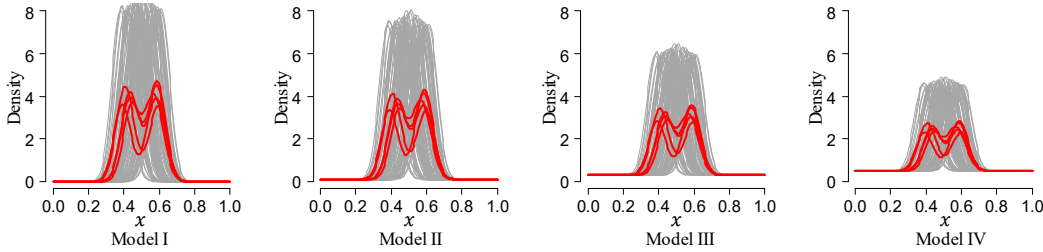


Figure S-40. Simulated PDF-valued datasets, each dataset consists of 100 curves with bold red curves standing for synthetic outlying PDFs.

Algorithm S.10: PDF generation procedure

Input: Number of PDFs n , parameters δ_1 and δ_2

Output: PDF-valued dataset $\{f_i\}_{i=1}^n$

- 1: Independently generate $a_i \sim U(\delta_1, \delta_2)$, $i = 1, 2, \dots, n$
 - 2: Sort $\{a_1, a_2, \dots, a_n\}$ in ascending order and denote the resulting series as $\{b_1, b_2, \dots, b_n\}$
 - 3: Generate PDF $f_i(x) = \text{BetaPdf}(x; a_i, b_i)$, $i = 1, 2, \dots, n$, where BetaPdf stands for the PDF of the beta distribution with parameters a_i and b_i .
 - 4: Output $\{f_i\}_{i=1}^n$
-

The whisker parameters of the boxplot-based detectors for the MO- and VO-outliers are set to $r_1 = 1.5$ and $r_2 = 2.5$, respectively. We first directly apply the FDO-based method to the PDFs (just replace the quantile function $Q_i(t)$ in Eqs. (S-49) and (S-50) by the corresponding PDF

$f_i(x)$), the detection results are shown in Figure S-41. Then, we perform the FDO-based outlier detection in the QF-space with the same parameter settings, the detection results are shown in Figure S-42. Unfortunately, excepting the QF-space detection for Model I exhibits a satisfactory detection result, grossly high false detection phenomena occur to the remaining scenarios. Comparing the MS-plots (the scatter plot of VO versus MO) as shown in the second row of Figure S-41 and Figure S-42 for the eight cases, it is evident that only the QF-space detection for Model I exhibits normal pattern. Further investigation found that such a poor performance is mainly attributed to the curve overlapping occurred at the lower and upper tails of the collection of the functions (PDFs or QFs).

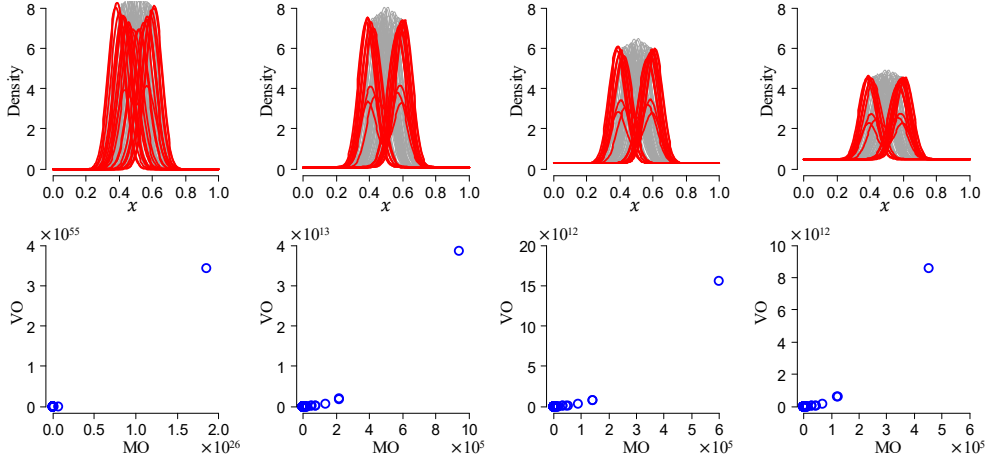


Figure S-41. Outlier detection results for the four datasets presented in Figure S-40 by directly applying the FDO-based method to PDFs. First row corresponds to the PDFs with the detected outliers represented by red curves, second row corresponds to the associated MS-plots (scatter plots of the MO- and VO-values).

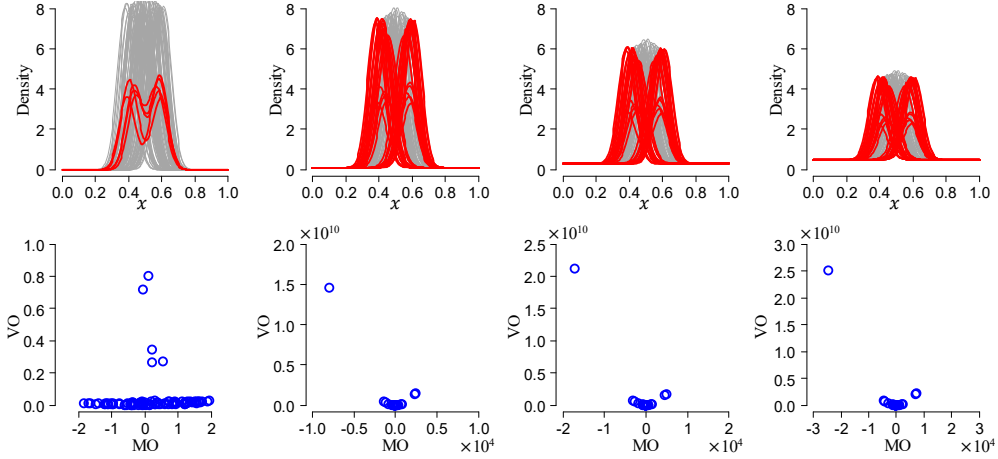


Figure S-42. Same as Figure S-41 except that the outliers are detected by performing the FDO-based method in the QF-space.

Fortunately, the curves of quantile functions have been naturally aligned according to their quantiles, facilitating us to choose a unified detection interval by truncating the lower and upper parts of I (i.e., the domain of definition of the quantile function). Denote the selected detection interval as $A = [\omega, 1 - \omega]$; the curves outside this interval are truncated as illustrated in the left panel of Figure S-43, which is equivalent to cut off the lower and upper tails of the PDFs as shown

in the right panel of Figure S-43. To apply the FDO-based approach to the resulted quantile functions, we only need to replace the $Q_i(t)$ and $\omega(t)$ in Eq. (S-50) by $Q_i(t)\chi_A(t)$ and $\omega(t) = 1/\lambda(A)$, respectively.

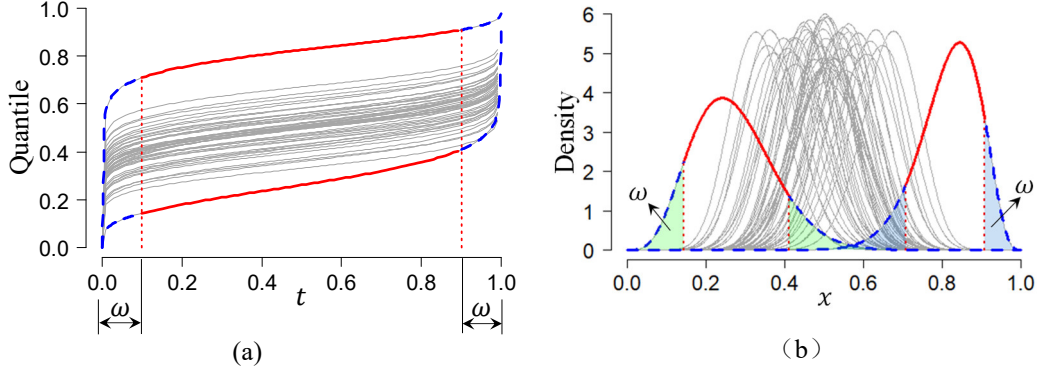


Figure S-43. Illustration of curve truncation using the selected detection region. (a) in quantile function space and (b) in PDF space.

(2) Competing method II: the warping function-based method

The warping function-based detection strategy considered in this study essentially belongs to the elastic depth-based approach proposed by Harris et al. (2021) for ordinary functional data. Analogous to the ordinary functional data setting, the distributional outliers can also be detected based on the phase information of the CDFs captured by the warping functions. In the following, we first present the computational details for the warping functions, followed by the outlier detection method.

Computational details for warping functions

This subsection provides computational details for the warping functions (used in pairwise alignment of cumulative distribution functions (CDFs)) as well as a strategy for dealing with the related numerical issue.

Given two PDFs $f_1(x)$ and $f_2(x)$ defined on the compact interval $[0,1]$, denote their corresponding CDFs as $F_1(x)$ and $F_2(x)$ (i.e., $F_i(x) = \int_{-\infty}^x f_i(\tau)d\tau, i = 1,2$). The problem of interest in this subsection is determining the warping functions, denoted as $\gamma_{12}(x)$ and $\gamma_{21}(x)$, subject to

$$\begin{aligned} (F_1 \circ \gamma_{12})(x) &= F_2(x), x \in [0,1] \\ (F_2 \circ \gamma_{21})(x) &= F_1(x), x \in [0,1] \end{aligned} \tag{S-51}$$

respectively. Theoretically, $\gamma_{12}(x)$ is the inverse element w.r.t. $\gamma_{21}(x)$ and vice versa, thus $\gamma_{12} \circ \gamma_{21} = \gamma_{id}$ with γ_{id} being the identity warping function defined as $\gamma_{id}(x) = x, x \in [0,1]$. The warping function γ_{12} represents the phase information of $F_1(x)$ w.r.t. $F_2(x)$. From the angle of curve alignment (also termed curve registration), γ_{12} plays the role of deforming the shape of

$F_1(x)$ to reach the shape of $F_2(x)$. In the community of functional data analysis, similar curve alignment is widely used in phase-amplitude separation (Srivastava et al. 2011; Srivastava and Klassen 2016), and the extracted phase information captured by warping functions can provide a useful feature in shape outlier detection (Harris et al. 2021).

It is worth noting that ordinary functional data have both amplitude and phase variabilities (Srivastava and Klassen, 2016), the warping functions used in pairwise alignments usually have to be solved by the dynamic programming (DP) algorithm in the square root slope velocity function (SRVF) framework described in Srivastava et al. (2011). In contrast, the CDFs only have phase variability, the warping functions can also be directly computed as:

$$\begin{aligned}\gamma_{12}(x) &= (F_1^{-1} \circ F_2)(x), x \in [0,1] \\ \gamma_{21}(x) &= (F_2^{-1} \circ F_1)(x), x \in [0,1]\end{aligned}\tag{S-52}$$

Compared to the time-consuming DP algorithm, this direct computation approach is much more efficient. Denote the computed result of the warping function γ_{12} (or γ_{21}) as $\hat{\gamma}_{12}$ (or $\hat{\gamma}_{21}$). If the PDFs take values near zero, $\hat{\gamma}_{12}$ and $\hat{\gamma}_{21}$ obtained by direct computation may be problematic, that is, they may fail to satisfy $\hat{\gamma}_{12} \circ \hat{\gamma}_{21} = \gamma_{id}$ as illustrated in Figure S-44(e).

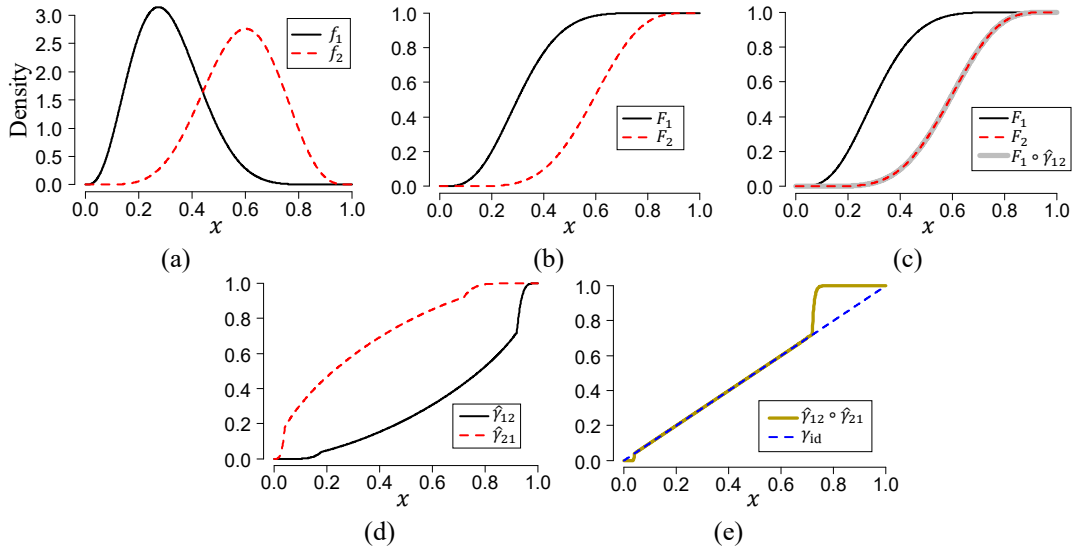


Figure S-44. CDF alignment of the simulated raw distributional data by using direct computed warping function. (a) Original PDFs f_1 and f_2 , (b) corresponding CDFs F_1 and F_2 , (c) comparison of $F_1 \circ \hat{\gamma}_{12}$ and F_2 , (d) calculated warping functions, and (e) comparison of $\hat{\gamma}_{12} \circ \hat{\gamma}_{21}$ and γ_{id} .

According to our experience, such a defect can be remedied by adding a small proportion of uniform distribution to the original distributions, that is, the PDFs are recommended to be preprocessed as follows:

$$f_i(x) = (1 - \alpha)f_i(x) + \alpha, x \in [0,1], i = 1,2\tag{S-53}$$

where α is a positive constant (termed PDF preprocessing parameter). By setting $\alpha = 0.1$, the re-calculated warping functions by direct computation are shown in Figure S-45 (d). Clearly, the line of $\hat{\gamma}_{12} \circ \hat{\gamma}_{21}$ agrees well with the line of γ_{id} (see Figure S-45 (e)), indicating the numerical issue

has been effectively overcome.

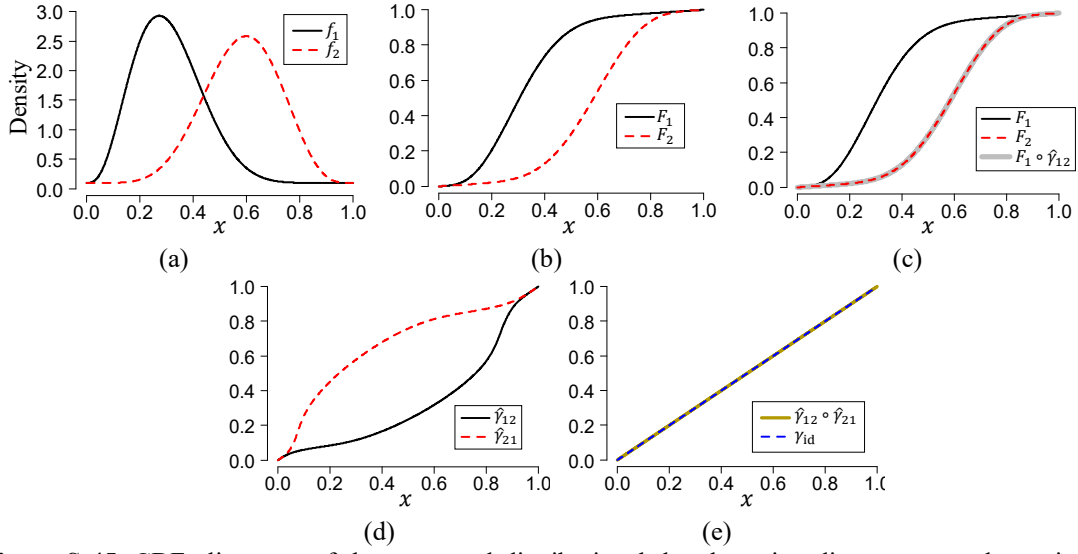


Figure S-45. CDF alignment of the processed distributional data by using direct computed warping function. (a) Processed PDFs f_1 and f_2 (obtained by using Eq. (S-53) with $\alpha = 0.1$), (b) corresponding CDFs F_1 and F_2 , (c) comparison of $F_1 \circ \hat{\gamma}_{12}$ and F_2 , (d) calculated warping functions, and (e) comparison of $\hat{\gamma}_{12} \circ \hat{\gamma}_{21}$ and γ_{id} .

Outlier detection method

As aforementioned, the distributional outliers can be detected based on the phase information of the CDFs captured by the warping functions. Given two CDFs $F_i(x)$ and $F_j(x)$, the warping function for extracting the phase difference of $F_i(x)$ with respect to $F_j(x)$ is calculated using the aforementioned direct computation method. Next, the phase distance (defined by Harris et al. (2021)) for the two CDFs can be calculated as $d_p(F_i, F_j) = \arccos\langle q_e, q_{ij} \rangle$ with $q_e(x) = 1$ and $q_{ij}(x) = \sqrt{d\gamma_{ij}(x)/dx}$ being the square-root slope functions (SRSFs) of the warping functions $\gamma_e(x) = x$ (identity warping function) and $\gamma_{ij}(x)$, respectively. With such phase distance at hand, we can calculate the phase depth defined by Harris et al. (2021). Then, the shape outliers hidden in the CDFs can be identified as phase anomalies by using Algorithm 1 described by Harris et al. (2021), only replacing the amplitude depth by the phase depth. The parameter k (in Algorithm 1 of Harris et al. (2021)) for controlling the whisker is set to 2.0. As pointed out in the previous, if the PDFs take near-zero values, the corresponding warping functions obtained using the direct computation method may be problematic. According to our experience, choosing the PDF preprocessing parameter $\alpha = 0.1$ can effectively alleviate the numerical issue; therefore, if the minimum value of the PDFs in a dataset is less than 0.1, all the PDFs in the same dataset will be processed using Eq. (S-53) with $\alpha_i = 0.1$.

Appendix 5: Demonstration of nonparametric regression-based dependence quantification

In Subsection S.7.6 of this document, we employ the nonparametric regression-based dependence quantification method described in Chen et al. (2020) to quantify the dependence of the FPC scores of the functional data from one sensor to the functional data from the other sensor, which is a scalar-on-function dependence quantification problem. In order to help the readers better understand the mechanism of such an approach in scalar-on-function dependence quantification, this appendix provides a simple example using only scalar data to illustrate how the method work (the case of scalar-on-function dependence is analogous).

We used the following nonlinear model to generate the data:

$$Y = 3 \sin(7\pi X) + 2X^5 + \varepsilon,$$

where $X \sim U(0,1)$, $\varepsilon \sim N(0, 2.65^2)$

Let $Z = 3 \sin(7\pi X) + 2X^5$ represent the noiseless data. We generate 200 samples for investigation, and the resulting real-valued datasets are denoted as $\{x_i\}_{i=1}^{200}$, $\{y_i\}_{i=1}^{200}$ and $\{z_i\}_{i=1}^{200}$. The data sequences of $\{x_i\}_{i=1}^{200}$ and $\{y_i\}_{i=1}^{200}$ are visualized in Figure S-46 and Figure S-47, respectively. Figure S-48 visualizes the scatter plot of $\{x_i, y_i\}_{i=1}^{200}$. The main task of this case study is to quantify the strength of dependence between Y and X . Using the generated samples $\{x_i, y_i\}_{i=1}^{200}$, the estimated linear correlation coefficient between Y and X is $\rho_{XY} = 0.1578$, obviously the dependence has been underestimated by this dependence measure.

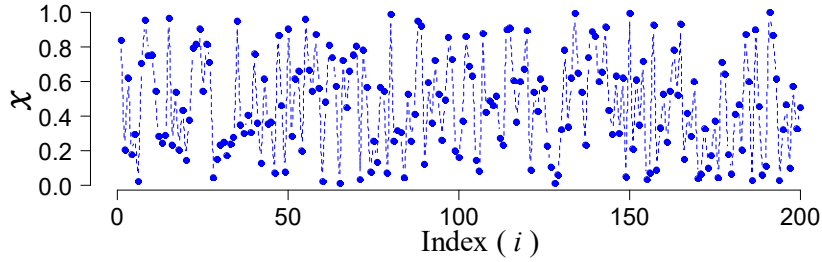


Figure S-46. Visualization of the generated data sequence $\{x_i\}_{i=1}^{200}$

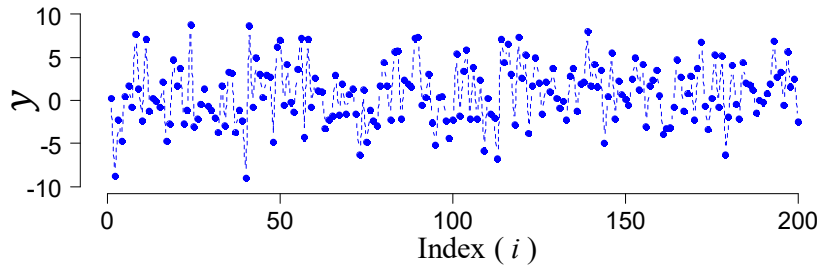


Figure S-47. Visualization of the generated data sequence $\{y_i\}_{i=1}^{200}$

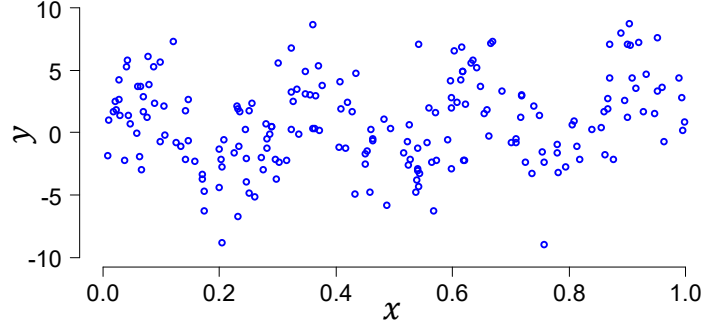


Figure S-48. The scatter plot of $\{x_i, y_i\}_{i=1}^{200}$.

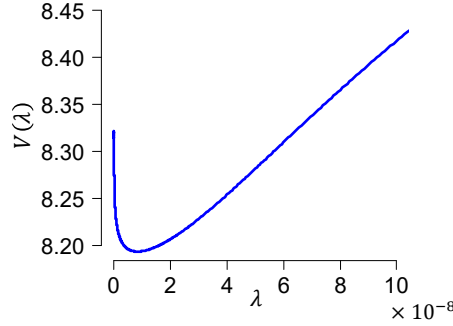


Figure S-49. Visualization of the calculated GCV statistic $V(\lambda)$ on different values of λ .

Then, we will re-quantify the dependence between Y and X using the nonparametric regression-based strategy described in Chen et al. (2020). We first use the scalar version of the RKHS-based nonparametric regression model (which is similar to the version of function-to-real regression presented in Appendix 3 of Chen et al. (2020), the main different is using the scale reproducing kernel $K(x_1, x_2) = \exp\left\{-\frac{1}{2\sigma^2}|x_1(\tau) - x_2(\tau)|^2\right\}$ instead of the functional reproducing kernel $K(u, v) = \exp\left\{-\frac{1}{2\sigma^2}\int |u(\tau) - v(\tau)|^2 d\tau\right\}$ to fit the underlying relationship between X and Y . The regularization parameter λ was determined by a similar generalized cross-validation (GCV) method described in Appendix 3 of Chen et al. (2020), which is selected as $\hat{\lambda}_{opt} = \underset{\lambda > 0}{\operatorname{argmin}} V(\lambda)$ (the calculated function of GCV statistic $V(\lambda)$ is shown in Figure S-49). The fitting result under the optimal regularization parameter is shown in Figure S-50, the fitted values agree well with the true values (i.e., $z = 3 \sin(7\pi x) + 2x^5$), indicating that the RKHS-based nonparametric regression model can well capture the underlying nonlinear relationship between Y and X . The fitting result can also be visualized in the form of data sequence as shown in Figure S-51. The sample standard deviations of the observations $\{y_i\}_{i=1}^{200}$ and the fitted values $\{\hat{y}_i\}_{i=1}^{200}$ are $s_Y = 3.4843$ and $s_{\hat{Y}} = 2.1554$, respectively. Finally, the dependence between Y and X is quantified by the ratio of the sample standard deviations as follows:

$$r = \frac{s_{\hat{Y}}}{s_Y} = \frac{2.1554}{3.4843} = 0.6186$$

Compared to the linear correlation coefficient $\rho_{XY} = 0.1578$, the degree of dependence calculated using nonparametric regression-based strategy is much more reasonable in this nonlinear dependence case.

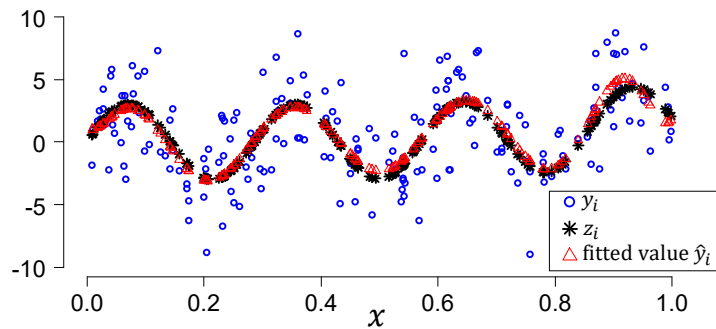


Figure S-50. Visualization of the fitting result in the form of scatter plot (with respect to x).

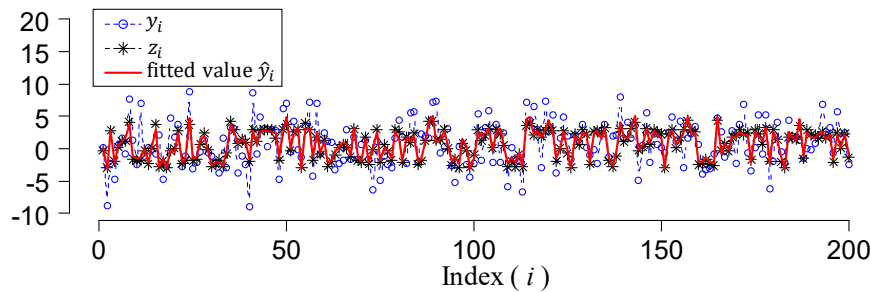


Figure S-51. Visualization of the fitting result in the form of data sequence.

References

- Berlinet, A., and Thomas-Agnan, C. (2004), *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, New York: Springer.
- Chen, Z., Bao, Y., Li, H., and Spencer Jr, B. F. (2018), “A Novel Distribution Regression Approach for Data Loss Compensation in Structural Health Monitoring,” *Structural Health Monitoring*, 17, 1473–1490.
- Chen, Z., Bao, Y., Li, H., and Spencer Jr., B. F. (2019a), “LQD-RKHS-based Distribution-to-Distribution Regression Methodology for Restoring the Probability Distributions of Missing SHM Data,” *Mechanical Systems and Signal Processing*, 121, 655–674.
- Chen, Z., Li, H., and Bao, Y. (2019b), “Analyzing and Modeling Inter-Sensor Relationships for Strain Monitoring Data and Missing Data Imputation: A Copula and Functional Data Analytic Approach,” *Structural Health Monitoring*, 18, 1168–1188.
- Chen, Z., Bao, Y., Tang, Z., Chen, J., and Li, H. (2020), “Clarifying and Quantifying the Geometric Correlation for Probability Distributions of Inter-Sensor Monitoring Data: A Functional Data Analytic Methodology,” *Mechanical Systems and Signal Processing*, 138, 106540.
- Coles, S. (2001), *An Introduction to Statistical Modeling of Extreme Values*, London: Springer.
- Dai, W., and Genton, M. G. (2019), “Directional Outlyingness for Multivariate Functional Data,” *Computational Statistics & Data Analysis*, 131, 50–65.
- Dai, W., Mrkvička, T., Sun, Y., and Genton, M. G. (2020), “Functional Outlier Detection and Taxonomy by Sequential Transformations,” *Computational Statistics & Data Analysis*, 149, 106960.

- Egozcue, J., Díaz-Barrero, J., and Pawłowsky-Glahn, V. (2006), “Hilbert Space of Probability Density Functions Based on Aitchison Geometry,” *Acta Mathematica Sinica, English Series*, 22, 1175–1182.
- Golub, G. H., Heath, M., and Wahba, G. (1979), “Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter,” *Technometrics*, 21(2), 215–223.
- Harris, T., Tucker, J. D., Li, B., and Shand, L. (2021), “Elastic Depths for Detecting Shape Anomalies in Functional Data,” *Technometrics*, 63, 1–11.
- Hron, K., Menafoglio, A., Templ, M., Hrušová, K., and Filzmoser, P. (2016), “Simplicial Principal Component Analysis for Density Functions in Bayes Spaces,” *Computational Statistics & Data Analysis*, 94, 330–350.
- Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., and Audiffren, J. (2016), “Operator-Valued Kernels for Learning from Functional Response Data,” *Journal of Machine Learning Research*, 17, 613–666.
- Lian, H. (2007a), “Nonlinear Functional Models for Functional Responses in Reproducing Kernel Hilbert Spaces,” *The Canadian Journal of Statistics*, 35, 597–606.
- (2007b), “Nonlinear Functional Models for Functional Responses in Reproducing Kernel Hilbert Spaces,” arXiv preprint arXiv:math/0702120v2.
- López-Pintado, S., and Romo, J. (2009), “On the Concept of Depth for Functional Data,” *Journal of the American Statistical Association*, 104, 718–734.
- Machalova, J., Hron, K., and Monti, G.S. (2016), “Preprocessing of Centred Logratio Transformed Density Functions Using Smoothing Splines,” *Journal of Applied Statistics*, 43, 1419–1435.
- Martínez-Hernández, I., Genton, M. G., and González-Farías, G. (2019), “Robust Depth-Based Estimation of the Functional Autoregressive Model,” *Computational Statistics & Data Analysis*, 131, 66–79.
- Murray, J. S. (2018), “Multiple Imputation: A Review of Practical and Theoretical Findings,” *Statistical Science*, 33, 142–159.
- O'Brien, E.J., Schmidt, F., Hajjalizadeh, D., Zhou, X.-Y., Enright, B., Caprani, C.C., Wilson, S. and Sheils, E. (2015), “A Review of Probabilistic Methods of Assessment of Load Effects in Bridges,” *Structural Safety*, 53, 44–56.
- Petersen, A., and Müller, H.-G. (2016), “Functional Data Analysis for Density Functions by Transformation to a Hilbert Space,” *The Annals of Statistics*, 44, 183–218.
- Petersen, A., Zhang, C., and Kokoszka, P. (2022), “Modeling Probability Density Functions as Data Objects,” *Econometrics and Statistics*, 21, 159–178.
- Royden, H. L. and Fitzpatrick, P. M. (2010), *Real Analysis (4th ed.)*, Englewood Cliffs: Prentice Hall.
- Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J.S. (2011), “Registration of Functional Data Using Fisher-Rao Metric,” arXiv preprint arXiv:1103.3817.
- Srivastava, A., and Klassen, E.P. (2016), *Functional and Shape Data Analysis*, New York: Springer.
- Talská, R., Menafoglio, A., Machalová, J., Hron, K., and Fišerová, E. (2018), “Compositional Regression with Functional Response,” *Computational Statistics & Data Analysis*, 123, 66–85.
- Van den Boogaart, K., Egozcue, J., and Pawłowsky-Glahn, V. (2014), “Bayes Hilbert Spaces”, *Australian & New Zealand Journal of Statistics*, 54, 171–194.
- Wahba, G., (1990), *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, URL <http://dx.doi.org/10.1137/1.9781611970128>.
- Yi, T.-H., Huang, H.-B., and Li, H.-N. (2017), “Development of Sensor Validation Methodologies for Structural Health Monitoring: A Comprehensive Review,” *Measurement*, 109, 200–214.