A change-point detection method for detecting and locating the abrupt changes in distributions of damage-sensitive features of SHM data, with application to structural condition assessment

Xinyi Lei¹,²,³, Zhicheng Chen¹,²,³, Hui Li¹,²,³ and Shiyin Wei¹,²,³

Abstract
Damage detection or structural condition assessment is an important objective of structural health monitoring (SHM). The damages or adverse changes in structural conditions can usually be manifested as pattern changes in damage-sensitive features (DSFs) extracted from SHM data; this enables us to shift damage detection to DSF change detection. Online monitoring can accumulate huge amounts of data, finding the changes from the massive DSF data through manual inspection is impractical; thus, automatic detection tools are required. If possible, relevant significance test is also desired to make a rational judgment on the existence of a change. In this sense, the change-point detection technique is an attractive choice, which is increasingly proved to be a powerful change detection tool in various SHM applications. However, existing change-point detection methods in SHM are mainly used for scalar or vector data, thus incapable of detecting changes in features represented by complex data, for example, the probability density functions (PDFs). Detecting abrupt changes in the distributions (represented by PDFs) of the DSF data is of crucial concern in structural condition assessment. However, relevant automatic diagnostic tools have not been well developed in the SHM community. To this end, a novel change-point detection method is developed in the functional data analytic framework for this task. The proposed approach has advantages in detecting changes for massive data and directly handling general PDFs. Considering that the major challenge in PDF-valued data analysis comes from the nonlinear constraints of PDFs, the PDFs are embedded into the Bayes space to develop the detection methodology by using the linear structure of the Bayes space. Comprehensive simulation studies are conducted to validate the effectiveness of the proposed method as well as demonstrate its superiority over the competing method. Finally, a case study involving cable condition assessment of a long-span bridge demonstrates its practical utility in SHM.

Keywords
Structural health monitoring, functional data analysis, change-point detection, automatic diagnosis, probability density function, data mining

Introduction
Structural health monitoring (SHM) is an important research field in civil engineering; it involves installing sensors on important structures to obtain a real-time monitoring for structural responses and operational environments and then performing structural health diagnosis through analyzing and mining the monitoring data.¹-⁴ Detecting changes in the features extracted from the measurements of interest is a very important content in SHM data mining,¹-⁷ as it usually contributes to providing useful even critical information for structural condition assessment or damage detection.
especially in the context of statistical pattern recognition (SPR) paradigm.\textsuperscript{3–5} The SPR paradigm contains two main steps: damage-sensitive feature (DSF) extraction and feature change detection.\textsuperscript{3–5} Long-term monitoring can accumulate huge amounts of data, even the DSF data have been extracted using related feature extraction algorithms, it is still impractical to screen out the suspected changes related to damages from the massive feature data via manual inspection. Therefore, developing effective and reliable statistical methods for automatically detecting feature changes is a pressing task for SHM data mining, which can tremendously benefit automatic structural damage detection or condition assessment.

Change-point detection is an attractive unsupervised statistical technique for feature change detection, which concerns detecting and dating abrupt pattern changes of a data sequence over an investigated time span. Change-point detection originates from statistical quality control, now it has become an important research topic in statistics\textsuperscript{3} and is increasingly recognized in the engineering field as a highly effective tool for damage detection, fault detection, abnormal structural behavior detection, and so on.\textsuperscript{9–17} In change-point detection, the decision on a change event is generally made according to the hypothesis testing under a given significance level; thus, the change-point detection technique has more potential in revealing subtler changes that might correspond to minor damage at an early stage. So far, various change-point detection methods have been developed or adopted for different SHM applications. Cathbas et al.\textsuperscript{9} proposed a nonparametric approach for detecting the changes in the correlation information of multi-sensor strain measurements in a damage detection application. Nigro et al.\textsuperscript{10} examined and compared several change-point detection methods used in other industries and then modified them to be applicable in SHM for damage detection. Pozo et al.\textsuperscript{11} combined the principal component analysis with various hypothesis testing tools to detect structural changes using piezoelectric sensor data. Liao et al.\textsuperscript{12} developed a sequential change-point detection-based algorithm for damage detection. Mariani and Cawley\textsuperscript{13} employed the generalized likelihood ratio change-point detection method to develop an automatic damage detection approach using guided wave signals. Most recently, Li et al.\textsuperscript{14} developed a change-point detection-based method for detecting the abnormal changes in crack behavior of concrete dams. Additionally, change-point detection methods have also been successfully applied in tunnel monitoring for abnormal structural behavior identification\textsuperscript{16} as well as in high-speed rail monitoring for large deformation detection.\textsuperscript{17}

To date, the change-point detection methods employed for SHM are generally applicable to scalar or vector data. In other words, the detection methods can only handle the features that can be converted to be represented as scalar- or vector-valued data living in the Euclidean space, thus limiting their applications in detecting the changes in the features represented by complex data objects without Euclidean data structure. The SHM data contain huge amounts of information; rich information is carried by complex data (e.g., high-dimensional data, functional data). By analyzing the complex data extracted from SHM data, one can uncover new information that cannot be clarified by conventional scalar or vector data analysis. Developing effective statistical tools for complex feature data analysis is another urgent task for current SHM data mining.

One representative type of complex and non-Euclidean data that can be widely encountered in SHM applications is the probability distributions. From the emerging object-oriented data analysis perspective, probability distributions (including their representations, such as density functions, histograms) can be regarded as data objects.\textsuperscript{18} Such data objects associated with random probability distributions are called distributional data.\textsuperscript{19} The distributions of the feature variables extracted from structural responses play an important role in structural health diagnosis.\textsuperscript{4,20–22} Generally, the pattern changes in the feature data are essentially caused by the changes of the underlying distributions. An abrupt change occurred in the distributions of DSF data may signify a change in structural conditions or behaviors. Researchers have found that performing structural condition assessment based on distributional information of the extracted DSF data is more effective than conventionally used modal parameters in some situations.\textsuperscript{20–21} So far, distribution-change-based structural condition assessment methods have been successfully applied in SHM for local structural state diagnosis, such as stay cables\textsuperscript{20} and steel box girders.\textsuperscript{21} Moreover, some types of the self-powered sensors (an emerging sensing technique and increasingly being used in SHM)\textsuperscript{23} record the cumulative durations of over-threshold events associated with the monitored structural responses rather than the corresponding time-histories.\textsuperscript{24,25} Such measurement records are actually distribution-like data, namely, the unnormalized histograms; after appropriate processing (i.e., normalization and density smoothing), they can be converted to standard distributional data. One significant advantage of such self-powered sensors is battery-free as they can harvest ambient energy.\textsuperscript{23,24} However, special data analysis technique is required to interpret and harness the recorded distributional information for damage detection. Recently, Azimi and Pekcan\textsuperscript{25} proposed a novel deep learning-based damage detection approach that can effectively utilize the histogram-valued data collected by the self-powered sensors, where the histograms are abstracted by extremely compressed data consisting of three parameters. For all the above SHM applications involving probability distributions, the pattern change of the distributional information is of great concern; however, the SHM community still lacks effective tools for automatically detecting such distributional changes. This practical requirement motivates this research to develop a general distributional change-point detection method for SHM applications.

In the distributional change-point detection method to be developed in this study, the raw DSF data are divided into data
segments and then summarized by their corresponding probability density distributions (PDFs). Generally, the extracted PDFs are continuous functions, which naturally belong to functional data. A directly related branch of modern statistics specifically for analyzing such data object consisting of random functions is functional data analysis (FDA).18,26 This study seeks to develop an FDA-based change-point detection method for the extracted PDF-valued sequence (also called distributional sequence throughout this study). In the field of statistics, related researches on change-point detection for functional data only emerged in the last decade, and they are mainly for ordinary functional data that assumed to lie in linear spaces; representative studies include Berkes et al.27 Aston and Kirch,28 and Aue et al.29 However, PDFs are special functional data, not only subjecting to inherent constrains (i.e., \( f(x) \geq 0 \), \( \forall x \in D \) and \( \int_D f(x) \, dx = 1 \) for arbitrary PDF \( f(x) \) defined on the domain \( D \)) but also not residing in a linear space under ordinary linear operations.19,30,31 Such an inherently nonlinear nature makes statistical analysis for distributional data more challenging. So far, only a few research contributions on change-point detection for distributional data have been reported. Padilla et al.32 developed a nonparametric approach for distributional change detection, the distributions are assumed to be identical within the two data sequences separated by the change-point. Most recently, Horváth et al.33 proposed another novel distributional change-point detection method based on Wasserstein distance, the distributions before (or after) the change-point are also assumed to be equal. Another excellent contribution is the Fréchet change-point detection method proposed by Dubey and Müller,34 which is developed for data objects residing in a general metric space and applicable to random PDFs or networks. However, for engineering applications, existing distributional change-point detection methods still have their imperfections. For instance, the identical distribution assumption made in Padilla et al.32 and Horváth et al.33 is too strong to be realistic for SHM applications. Additionally, existing methods generally lack effective measures to cope with the adverse impacts caused by outliers. The SHM systems operate in complicated environments, data contamination is very common and inevitable1,2; thus, being able to provide reliable results under data contamination is a basic performance requirement for the change-point detectors employed for SHM applications.

This study proposes an alternative distributional change-point detection method for data-driven structural condition assessment applications. The distributions before and after the change-point are assumed to be respectively generated from two different functional stochastic processes. In contrast to the identical distribution assumption made in Padilla et al.32 and Horváth et al.33, the distributions before (or after) the change-point in our method are not necessary to be identical. The distributional change-point model is built on the linear structure of the Bayes space after imbedding the PDFs into the Bayes space. The hypothesis testing procedure for change-point detection is developed by leveraging the isometric isomorphism property of the Bayes space. To obtain more reliable detection results under data contamination, a data cleaning strategy is also presented to handle outlying PDFs. A case study involving cable condition assessment of a long-span bridge reveals its great practical utility in SHM.

### Problem formulation of distributional change-point detection and its merits

This section formulates the problem of distributional change-point detection to be investigated in this study. Suppose the damage-sensitive feature (DSF) data have been extracted from the raw SHM data with the environmental and operational effects eliminated in advance using existing methods. Once the DSF data have been extracted, they are divided into a series of non-overlapping time segments (e.g., a day, a week, or other specific time span) and then the segment-data are abstracted by their corresponding PDFs estimated by kernel density estimator. By this setting, a new functional dataset denoted as \( F = \{ f_i : i = 1, 2, \ldots, n \} \) (formed by the PDFs ordered in time) can be obtained, which is referred to as PDF-valued sequence or distributional sequence. From a statistical perspective, the PDF-valued sequence can be regarded as a realization of an unknown underlying distributional stochastic process. The distributional change-point detection problem investigated in this study concerns on detecting and locating the abrupt changes occurred in the underlying distributional stochastic process based on its observations \( F = \{ f_i : i = 1, 2, \ldots, n \} \). As discussed in the introduction, the sudden pattern change of the distributions associated with the DSF data is an important sign that may signify a change of the structural condition (see, e.g., Li et al.20 and Wei et al.21 for related engineering cases).

Such a distributional change-point detection framework has several attractive aspects for damage-sensitive feature change detection in SHM applications:

1. **Facilitate massive data detection:** In our method, the change detection is performed to the functional dataset rather than the raw DSF data; thus, the sample size of the data processed in the change detection procedure is substantially lower than the number of data objects in the original DSF dataset. For instance, if one-year DSF data (consisting of huge amounts of scalar data points) are divided into daily segments, the resulting PDF-valued dataset only contain 365 data objects. By this setting, one can implement an efficient change detection for the massive SHM data accumulated by a long monitoring period (e.g., years or even decades).

2. **Lower risk of model misspecification:** The distributional change-point detection method going to be developed in the FDA context can directly deal with
the nonparametric estimates of the PDFs under mild assumptions, that is, the PDFs processed in the change detection procedure only have to satisfy the constrains of PDFs (i.e., non-negative and unit integral) and support on a finite interval \([a,b]\). In this framework, the data do not have to be assumed to follow the normal or mixture normal distribution model; thus, the risk of model misspecification is reduced.

(3) The identical distribution assumption is relaxed: In our method, the distributions before and after the change-point are two groups of distributions assumed to be generated by two different distributional stochastic processes. This is more realistic compared to the identical distribution assumption where the data coming before (or after) the change-point are assumed to identically distributed. The SHM data contain various uncertainties, and the identical distribution assumption is too strong to be realistic.

### Distributional change-point detection method for SHM

This section presents the proposed methodology for distributional change-point detection. The “observed” PDFs \(F = \{f_i(x) : i = 1, 2, \cdots, n\}\) are assumed to be finitely supported on the interval \([a,b]\), that is, \(f_i(x) > 0\) if \(x \in [a,b]\) and \(f_i(x) = 0\) if \(x \notin [a,b]\). In SHM applications, such a finite support assumption can be generally satisfied since the measurements of the structural responses cannot tend to infinity. For how to determine the common support of the PDFs using the SHM data, readers are referred to Chen et al.\(^\text{35}\) for a detailed discussion. Denote the underlying distributional stochastic process that governs the “observed” PDF-valued sequence \(F = \{f_i(x) : i = 1, 2, \cdots, n\}\) as \(\sigma_i(x) : i = 1, 2, \cdots, n\), where \(\sigma_i(x)\) is a PDF-valued random variable. Let \(E[\sigma_i(x)]\) denote the mean of \(\sigma_i(x)\), and obviously, \(E[\sigma_i(x)]\) is also a function, which is referred to as the mean function of \(\sigma_i(x)\). The distributional change-point detection concerned in this study is detecting and locating the abrupt changes in the mean function associated with the distributional stochastic process, which falls into the mean break analysis problem in statistics.\(^\text{3}\) For the structural condition assessment problem in SHM applications, an abrupt change in the mean of the distributional stochastic process associated with the DSF data indicate the pattern of the damage-sensitive feature has changed, which is an important symptom that can signify a suspected change of the structural condition.

In the following, the change-point model of the ordinary functional sequence is briefly recalled, along with a discussion on its limitations in distributional data modeling; then, an alternative change-point model constructed in the Bayes space will be proposed specifically for distributional data, based on which the proposed distributional change-point detection method will be developed. Finally, measures for ensuring a reliable detection under data contamination will further be provided.

### Preliminaries

This subsection briefly introduces the mean break (i.e., abrupt change in the mean) model for ordinary functional data described in Aue et al.\(^\text{29}\) Consider a functional sequence consisting of \(n\) random functions taking values in the \(L^2[a,b]\) space (a type of linear space formed by all square integrable functions), namely, \(\{X_i(t) : i = 1, \cdots, n\}\) with \(X_i \in L^2[a,b]\) and \(t \in [a,b]\), the mean break of this functional sequence can be modeled as

\[
X_i(t) = \mu(t) + \chi_{[k^* + 1, n]}(i) \delta_i(t) + \epsilon_i(t), \quad t \in [a,b] \tag{1}
\]

where \(\mu(t)\) is the baseline mean function, \(\delta_i(t)\) is the increment of the mean function, \(k^*\) is the change-point location, \(\epsilon_i(t)\) is the zero-mean error term, and \(\chi_{[k^* + 1, n]}(i)\) is the indicator function defined as

\[
\chi_{[k^* + 1, n]}(i) = \begin{cases} 
1, & \text{if } i \in [k^* + 1, n] \\
0, & \text{if } i \notin [k^* + 1, n]
\end{cases} \tag{2}
\]

Obviously, before the change-point, the mean function of the functional data is \(\mu(t)\); after the change-point, the mean function has changed to \(\mu(t) + \delta(t)\).

However, if this model is employed to characterize the mean break occurred in the PDF-valued sequence \(F = \{f_i(x) : i = 1, 2, \cdots, n\}\), that is

\[
f_i(x) = \mu_i(x) + \chi_{[k^* + 1, n]}(i) \delta_i(x) + \epsilon_i(x), \quad x \in [a,b] \tag{3}
\]

it may fail to ensure that \(\mu_i\), \(\delta_i\), and \(f_i\) can simultaneously satisfy the constrains of PDFs (because the PDF space is not closed under ordinary linear operations, namely, the pointwise addition and scalar multiplication defined for ordinary functional data. For instance, if \(f(x)\) and \(g(x)\) are two PDFs, then \(f(x) + g(x)\) or \(a \cdot f(x)\) would no longer be a PDF unless \(a = 1\). If the mean functions \(\mu_j(x)\) and \(\mu_j(x) + \delta_j(x)\) before and after the change-point are no longer PDFs, the associated distributional change-point model would become uninterpretable.

### A Bayes space-based change-point model for distributional sequence

The main challenge for PDF-valued data modeling is that the PDF space lacks the linear structure in usual sense. However, by introducing the following additive and scalar multiplication operations to PDFs\(^\text{36–38}\)
addition: \((f \otimes g)(x) = \int_{I} f(\tau)g(\tau) d\tau\), \(x \in I\) (4a)

scalar multiplication: \((c \otimes f)(x) = \int_{I} f(\tau)^{c} d\tau\), \(x \in I, c \in \mathbb{R}\) (4b)

where \(f(x)\) and \(g(x)\) are both PDFs defined on the compact interval \(I = [a, b]\), the PDFs can be regarded as the elements of the Bayes space denoted as \(B^{2}(I) = B^{2}([a, b])\) (see the Appendix for the theoretical background and notions). The Bayes space is a linear space.37,38

Next, by using the linear structure (i.e., the linear operations defined in equation (4)) of the Bayes space, the functional mean break model given in equation (1) will be extended to accommodate PDF-valued data. Considering a functional series composed of \(n\) PDFs residing in the Bayes space \(B^{2}(I) = B^{2}([a, b])\), that is

\[
F = \{f_{1}(x), f_{2}(x), \ldots, f_{n}(x)\}, \quad f_{i} \in B^{2}(I), \quad x \in I = [a, b]
\] (5)

Without loss of generality, the common support \(I = [a, b]\) of the PDFs is assumed to be \(I = [0, 1]\). For the PDFs supported on a general domain, namely, \(I = [a, b]\), they can be converted to be supported on \([0, 1]\) through performing a scale transformation (see, e.g., Chen et al.35 for more detailed descriptions). By this setting, the sample mean function of the \(n\) PDFs can be calculated as

\[
\hat{\mu}_{f}(x) = \left(\frac{1}{n} \otimes \left(\sum_{i=1}^{n} f_{i}\right)\right)(x), \quad x \in [0, 1]
\] (6)

It can be verified that \(\hat{\mu}_{f}\) is also an element of the Bayes space \(B^{2}([0, 1])\), because the latter is closed under the linear operations defined in equations (4a) and (4b).

After embedding the PDFs into the Bayes space, the change-point model given in equation (1) can be tailored to model the mean break of distributional data as follows

\[
f_{i}(x) = (\mu_{f} \otimes \chi_{k^{*}+1, n}(i) \otimes \delta_{f}) \otimes e_{f}(x), \quad x \in [0, 1]
\] (7)

where \(\mu_{f}\), \(\delta_{f}\), \(e_{f}\), and \(k^{*}\) have the analogous meanings as their counterparts in equation (1). Before the change-point (i.e., \(1 \leq i \leq k^{*}\)), the mean function of the PDF-valued random variables in the distributional stochastic process is \(\mu_{f}\); after the change-point (i.e., \(k^{*} + 1 \leq i \leq n\)), the mean function changes to \(\mu_{f} \otimes \delta_{f}\). Recall that the Bayes space \(B^{2}([0, 1])\) is closed under the operations \(\otimes\) and \(\oplus\), one can verify that \(\mu_{f} \otimes \delta_{f}\) and \(\mu_{f} \oplus \delta_{f}\) can simultaneously satisfy the constrains of PDFs. In this sense, the new model in equation (7) is more interpretable for distributional data.

### Change-point detection method

The distributional change-point detection problem concerns locating the change event occurred in the mean of the distributional sequence as well as testing its significance through hypothesis testing. Only the one that pass the significance test under a pre-specified significance level can be treated as the “true” change-point. The main steps of the change-point detection framework are summarized in the flowchart shown in Figure 1.

Given the change-point model in equation (7), the problem of change-point detection is equivalent to test whether \(\delta_{f} = 0\) under a pre-specified significance level. Thus, the null and alternative hypothesis for change-point detection is

\[
H_{0}: \delta_{f} = 0, \quad H_{1}: \delta_{f} \neq 0
\] (8)

Inspired by the functional cumulative sum statistic (FCSS) constructed in Aue et al.29 for ordinary functional data defined in the \(L^{2}([0, 1])\), the following modified FCSS is constructed in the Bayes space as the basic statistic in subsequent change-point localization and significance test

\[
F_{n,k}^{0} = \frac{1}{\sqrt{n}} \otimes \left(\left(\frac{k}{n} \otimes f_{1}\right) \oplus \left(\frac{-k}{n} \otimes f_{n}\right)\right)
\] (9)

It is worth noting that \(F_{n,k}^{0}\) also resides in the Bayes space \(B^{2}([0, 1])\), and its B-norm \(\|F_{n,k}^{0}\|_{B}\) (induced by the inner product defined in equation (24) of the Appendix, that is, \(\|f\|_{B} = (\langle f, f \rangle_{B})^{1/2} \forall f \in B^{2}([0, 1])\)) is a function of the index \(k\). Similar to the property of the FCSS in the \(L^{2}([0, 1])\) sense,29 \(\|F_{n,k}^{0}\|_{B}\) also tends to reach its maximum value at \(k = k^{*}\); thus, the change-point \(k^{*}\) of the distributional sequence can be estimated by

\[
k^{*} = \min\left\{k : \|F_{n,k}^{0}\|_{B} = \max_{1 \leq k \leq n} \|F_{n,k}^{0}\|_{B}\right\}
\] (10)

Also similar to that in ordinary functional data setting,29 the significance test for the mean break can be realized based on the limiting distribution of the following statistic

\[
T_{f,\alpha} = \max_{1 \leq k \leq n} \|F_{n,k}^{0}\|_{B}^2
\] (11)

Determining the limiting distribution of the test statistic \(T_{f,\alpha}\) directly in the Bayes space is not straightforward. Fortunately, the Bayes space \(B^{2}([0, 1])\) has an attractive property that it is isometrically isomorphic to the \(L^{2}([0, 1])\) space37,38 (see the Appendix for more details). One mapping that can serve as the isometric isomorphism between \(B^{2}([0, 1])\) and \(L^{2}([0, 1])\) is the centered log-ratio (clr) transformation defined as follows37,38

\[
\text{clr}(x) = \log \left(\frac{x}{1-x}\right)
\]
Determine the null and alternative hypothesis for change-point detection

Construct related statistics for change-point location estimation and significance test

Estimate the change-point location $k^*$

Perform a significance test for the pre-found change-point, and make a decision on whether the pre-found change-point is “true” or not based on the significance test result.

Figure 1. Framework of the change-point detection.

$\text{clr}[f](x) = \log f(x) - \int_{0}^{1} \log(t) dt, x \in [0, 1], f \in B^2([0, 1])$

(12)

In the sense of isometric isomorphism, the test statistic given in equation (11) can be equivalently written as

$$T_{f,n} = \max_{1 \leq k \leq n} ||F_{n,k}^0||_B^2 = \max_{1 \leq k \leq n} ||\text{clr}[F_{n,k}^0]||_2^2 = T_{\text{clr}[f],n}$$

(13)

where $|| \cdot ||_2$ stands for the $L^2$-norm defined as $||u||_2 = (\int_{0}^{1} u^2(t) dt)^{1/2}$, $\forall u \in L^2([0, 1])$, $\text{clr}[F_{n,k}^0]$ is the clr-transformation of the FCSS $F_{n,k}^0$ given in equation (9). Using the property given in equation (25) of the Appendix, the expression of $\text{clr}[F_{n,k}^0]$ can be written as

$$\text{clr}[F_{n,k}^0](x) = \frac{1}{\sqrt{n}} \left( \sum_{i=1}^{k} \text{clr}[f_i](x) - \frac{k}{n} \sum_{i=1}^{n} \text{clr}[f_i](x) \right)$$

(14)

Note that the clr-transformed result $\text{clr}[F_{n,k}^0](x)$ also has a constraint of integrating to zero \(38\); thereby, $\text{clr}[F_{n,k}^0]$ actually resides in a subspace of $L^2([0, 1])$. If we neglect this constraint and treat the clr-transformed functions as ordinary functions in $L^2([0, 1])$, then the limiting distribution of the test statistic given in equation (13) under $H_0$ is\(^{29}\)

$$T_{\text{clr}[f],n} = \max_{1 \leq k \leq n} ||F_{n,k}^0||_B^2 \rightarrow \sup_{0 \leq s \leq 1} \sum_{i=1}^{k} B_i^s(x) = T', \quad n \rightarrow \infty, x \in [0, 1]$$

(15)

where $\rightarrow$ denotes convergence in distribution, $B_i(x)$ s are standard Brownian bridges defined on $[0, 1]$ and independent across $l$, and $\lambda_l$ s are the eigenvalues of the covariance operator associated with the clr-transformed error sequence $\{\text{clr}[e_i](x) : 1 \leq i \leq n\}$ (the calculation will be provided later). Since the clr-transformed results belong to a subspace of $L^2([0, 1])$, $T_s = \sup_{0 \leq x \leq 1} \sum_{l=1}^{n} \lambda_l B_l^s(x)$ given in equation (15) is an approximation to the distribution of the test statistic $T_{\text{clr}[f],n}$. Such an approximation may introduce approximation errors; however, related simulation and real data studies conducted later indicate that this approximated limiting distribution is effective in detecting the change-points of the investigated distributional sequences. Moreover, related simulation studies will also validate that such a clr-transforming strategy is superior to an alternative strategy that directly regarding PDFs as ordinary functions.

We now turn to the calculation for the eigenvalue sequence $\{\lambda_l : l = 1, 2, \cdots\}$ involved in equation (15). According to the theory of functional data analysis,\(^{26,29}\) the covariance kernel $C_{\text{clr}[e]}$ and the covariance operator $c_{\text{clr}[e]}$ associated with the clr-transformed error sequence $\{\text{clr}[e_i](x) : 1 \leq i \leq n\}$ are defined as

$$C_{\text{clr}[e]}'(t,s) = \frac{1}{n} \sum_{i=1}^{n} \text{clr}[e_i](t) \text{clr}[e_i](s), \quad (t,s) \in [0, 1] \times [0, 1]$$

(16a)

$$c_{\text{clr}[e]}(\eta) = \int C_{\text{clr}[e]}'(\cdot,\cdot)\eta(\cdot)d\eta \equiv L^2([0, 1]), \quad s \in [0, 1]$$

(16b)

Performing an eigen-decomposition of the covariance operator $c_{\text{clr}[e]}$ can yield the eigenvalue sequence $\{\lambda_l : l = 1, 2, \cdots\}$ in descending order, and the corresponding eigenfunctions satisfy

$$c_{\text{clr}[e]}(\phi_l)(x) = \lambda_l \phi_l(x), \quad l = 1, 2, \cdots$$

(17)
For computational details about such an eigen-decomposition of the covariance operator associated with functional data (see, e.g., the monograph of Ramsay and Silverman). In practical applications, the error sequence is unknown, and it has to be calculated from the clr-transformed functions \{clr[f_i]\} by subtracting the sample mean function. Since the clr-transformed functions are treated as ordinary functional data, the calculation for the empirical covariance kernel \(\hat{C}_{\text{clr}[x]}\) is the same as that in ordinary functional data setting (see, e.g., Aue et al. and Aston and Kirch). In practical applications, the infinite eigenvalue sequence \(\{\lambda_l: l = 1, 2, \cdots\}\) involved in the limiting distribution \(T_s\) is also recommended to be truncated at \(l = L\) with \(L\) determined by

\[L = \min\left\{l \in \mathbb{N} : \sum_{i=1}^{l} \lambda_i \geq \theta\right\},\]

where \(\theta\) is a given threshold and its default value is set to 0.95 throughout this study.

With the approximated limiting distribution of the test statistic at hand, the hypothesis testing for judging whether \(\delta_x = 0\) or not can be realized by checking whether the calculated test statistic \(T_{s,x}\) falls into the rejection region of the limiting distribution \(T_s\) under a pre-specified significance level \(\alpha \in (0, 1)\). However, directly determining the rejection region is not straightforward, a Monte Carlo-based approach is more practical. Specifically, the i.i.d. Brownian bridges \(\{B_t(x): l = 1, 2, \cdots, L\}\) involved in the truncated limiting distribution \(T_{s,L} = \sup_{0 \leq x \leq 1} \sum_{l=1}^{L} \lambda_l B_t^2(x)\) can be first simulated by the Monte Carlo method implemented in the R package sde (https://cran.r-project.org); then, by plugging the simulated \(\{B_t(x): l = 1, 2, \cdots, L\}\) into \(T_{s,L}\), one can obtain a Monte Carlo sample of \(T_{s,L}\). Repeat the above procedure for \(M\) times and denote the resulting Monte Carlo samples as \(\{T_{s,L}^1, T_{s,L}^2, \cdots, T_{s,L}^M\}\); then, the \(p\)-value associated with the hypothesis testing can be calculated by

\[p = \frac{1}{M} \sum_{l=1}^{M} \mathbb{I}[T_{s,L}^l(\tau_{x}) \leq \chi_{l,0}^2] \left(T_{s,L}^l, \tau_{x}\right)\]

where \(T_{s,x}\) is the test statistic calculated by equation (13), and \(\chi_{l,0}^2(\cdot)\) is the indicator function. If the \(p\)-value is less than the pre-specified significance level, then it suggests that the test statistic \(T_{s,x}\) falls into the rejection region of the limiting distribution; thus, \(H_0\) should be rejected. Finally, the implementation details of the significance test procedure for distributional change-point detection are summarized in the flowchart shown in Figure 2.

![Figure 2. Flowchart of the significance test procedure for distributional change-point detection.](image-url)
Measures for robustness improvement under data contamination

The FCSS defined in equation (9) plays a critical role in the change-point detection. Such a statistic is constructed by combining the sample mean function of the PDFs coming before \( k \) with the global sample mean function of the whole sequence. Since the sample mean is sensitive to outliers, the calculated FCSS may also be seriously distorted by the outlying PDFs presented in the distributional sequence. The SHM system usually operates in complicated environment, data contamination is very common and inevitable due to various factors, which may cause some estimated PDFs to exhibit as outliers. Such outliers usually randomly occur in the PDF-valued sequence, or last for a short time then the PDFs return to the normal pattern; in most situations, they are primarily caused by unknown disturbances rather than damages (this phenomenon will be illustrated in the case study). Therefore, effective measures are required to appropriately address the outlying PDFs, so as to reduce their adverse impacts on the change-point detection. For this purpose, we recommend to conduct a data cleaning to the raw distributional data. Specifically, the potential outlying PDFs are first detected by functional outlier detection methods, and then removed from the distributional sequence before implementing the change-point detection.

Algorithm 1. Robust change-point detection

1: Perform the Tree-Distance outlier detection method to \( F = \{ f_1, f_2, \cdots, f_n \} \);
2 (optional): Perform the QF-FDO outlier detection method to \( F = \{ f_1, f_2, \cdots, f_n \} \);
3: Remove the detected outlying PDFs from \( F \), and denote the remaining sub-sequence as \( F^- = \{ f_{k_1}, f_{k_2}, \cdots, f_{k_m} \} \), where \( k_l \) is the original index of the PDF in \( F \);
4: Perform change-point detection to \( F^- \);
5: Convert the change-point location estimated in Step 4 to its corresponding location in the original sequence.

Although functional outlier detection methods for ordinary functional data have been well developed, relative research contributions to distributional outlier detection are still quite rare. Recently, Lei et al.\(^{39}\) conducted a systematic study on distributional outlier detection, and relevant computationally efficient detection tools were developed. Following the recommendation in Lei et al.\(^{39}\), the Tree-Distance distributional outlier detection method\(^{39}\) is selected as the primary tool for outlying PDF detection in this study, and the QF-FDO method\(^{39}\) is utilized as a complementary tool. After fusing the distributional outlier cleaning processing, the new change-point detection procedure is outlined in Algorithm 1.

Simulation studies

Simulation study I

This simulation study aims at validating the effectiveness of the proposed method as well as comparing its performance with an alternative strategy.

Firstly, the following functional change-point model is used to generate a raw functional sequence denoted as \( \{ f_i(x) \}_{i=1}^{n} \)

\[
\tilde{f}_i(x) = f_{\text{Beta}}(x; a_i, b_i) + 0.8 \cdot \chi_{[k^*+1,n]}(i), \quad x \in [0, 1], \quad i = 1, \cdots, n
\]

(19)

where \( f_{\text{Beta}}(x; a_i, b_i) \) is the density of a Beta distribution with parameters \( a_i \) and \( b_i \) generated by

\[
a_i \sim \text{U}(14, 25), \quad b_i = \text{sort}(\{a_i\}_{i=1}^{n}), \quad i = 1, \cdots, n
\]

(20)

where \( \text{U}(u,v) \) denotes the uniform distribution on \([u,v]\), and \( \text{sort}(A) \) represents sorting the elements in set \( A \) in ascending order. Then, \( \{ \tilde{f}_i(x) \}_{i=1}^{n} \) are converted into PDFs to yield the desired distributional sequence \( F = \{ f_1, f_2, \cdots, f_n \} \) as follows

\[
f_i(x) = \frac{\tilde{f}_i(x) - h}{\int_0^1 \tilde{f}_i(\tau) - h \, d\tau}, \quad h = \min_{1 \leq x \leq n} \inf_{x \in [0,1]} \left\{ \tilde{f}_i(x) | i = 1, 2, \cdots, n, x \in [0,1] \right\}
\]

(21)

A representative sample dataset generated by this model \((n = 100 \text{ and } k^* = 50)\) is visualized in Figure 3.

Two detection methods are considered for performance comparison: (1) the proposed method; (2) the competing method, which treats PDFs as ordinary functional data with the change-point detected by using the ordinary functional mean break detection method described in Aue et al.\(^{39}\). In the hypothesis testing, the significance level for each method is set to \( \alpha = 0.05 \), and the parameter \( M \) in equation (18) is set to 2000. The performances of the two methods are evaluated based on their detection accuracies in 500 repeated detection experiments. In each experiment, the distributional sequence is regenerated using the aforementioned data-generating procedure with \( n = 100 \) and \( k^* = 50 \). The accuracy of the change-point estimation is quantified by the absolute error defined as \( e = |\hat{k}^* - k^*| \), where \( \hat{k}^* \) represent the estimate of \( k^* \). To avoid the numerical issue encountered in the logarithmic computation involved in the clr-transformation (see equation (12)) if the PDF takes values near zero, all the PDFs are processed by \( f_i(x) = 0.9f_i(x) + 0.1, \quad x \in [0, 1] \). Unless otherwise stated, this processing will be conducted by default throughout the
rest of this study. In all the 500 repeated experiments, the calculated p-values associated with the two methods are all less than the pre-specified significance level (i.e., \( \alpha = 0.05 \)). The absolute errors of the change-point locations estimated by the two methods in the 500 repeated experiments are displayed in Figure 4(a) and summarized as the boxplots shown in Figure 4(b). Comparing the results, one can see that the proposed method significantly outperforms the competing method, indicating that the change-point model constructed in the Bayes space is more suitable for PDF-valued data.

**Simulation study II**

This simulation study aims at investigating the impacts of outlying PDFs on the distributional change-point detection, as well as validating the effectiveness of the distributional outlier cleaning strategy in ensuring the detection quality under data contamination.

The following three synthetic data-generating models are considered:

- **Model I**: Before the change-point (i.e., \( 1 \leq i \leq k^* \)), the distributional data are generated by a Beta distribution model \( \text{Beta}(a_i, b_i) \) with parameters \( a_i \) and \( b_i \) independently generated by \( a_i \sim U(10, 15) \), \( b_i \sim U(10, 15) \); after the change-point (i.e., \( k^* + 1 \leq i \leq n \)), the distributional data are generated by a mixture Beta distribution model \( 0.5\text{Beta}(\tilde{a}_i, \tilde{b}_i) + 0.5\text{Beta}(a_i^b, b_i^b) \) with the parameters \( \tilde{a}_i, \tilde{b}_i \), \( a_i^b \), and \( b_i^b \) independently generated by \( \tilde{a}_i \sim U(25, 40) \), \( \tilde{b}_i \sim U(15, 20) \), \( a_i^b \sim U(2, 4) \), and \( b_i^b \sim U(4, 6) \). The resulting distributional sequence is denoted as \( F_I = \{f_i^I(x) : i = 1, \cdots, n\} \). Figure 5 visualizes the
representative sample data generated by this model with $n = 100$ and $k^* = 50$.

Model II: The distributional sequence is generated by a Beta distribution model $\text{Beta}(a_i,b_i)$ with the parameters following different distributions before and after the change-point. Before the change-point, the parameters $a_i$ and $b_i$ are generated by $a_i \sim U(15,25)$ and $b_i = \beta a_i$ with $\beta = \frac{1}{c} - 1$; after the change-point, the parameters $a_i$ and $b_i$ are generated by $\tilde{a}_i \sim U(5,10)$ and $\tilde{b}_i = \beta \tilde{a}_i$ with $\beta = \frac{1}{c} - 1$. The resulting distributional sequence can be written in a unified form as follows

$$f_i^{II}(x) = \chi_{[1,k^*]}(i)f_{\text{Beta}}(x;a_i,b_i) + \chi_{[k^*+1,n]}(i)f_{\tilde{\text{Beta}}}(x,\tilde{a}_i,\tilde{b}_i)$$  \hspace{1cm} (22)

According to the theory of Beta distribution, the parameter $c$ involved in the above data-generating procedure is actually the expectation (i.e., the first-order moment) of the random variable $X \sim \text{Beta}(a,b)$ with $b = \beta a = \left(\frac{1}{c} - 1\right)a$, that is, $c = \int xf_{\text{Beta}}(x;a,b)dx$. Therefore, the underlying random variables associated with the distributions generated by Model II have the same expectations. Throughout this simulation study, the value of $c$ is fixed at 0.45, the resulting distributional sequence is denoted as $F_{II} = \{f_i^{II}(x) : i = 1, \cdots, n\}$ and the corresponding representative sample data ($n = 100$ and $k^* = 50$) are visualized in Figure 6.

Model III: The distributional sequence is also generated by the Beta distribution model, but the parameters are simulated using different principles. Before the change-point, the parameters $a_i$ and $b_i$ of the Beta distribution are generated by $a_i \sim U(15,25)$ and $b_i = \beta a_i$ with $\beta_i$ independently drawn from $U(0.85, 1.0)$; after the change-point, the parameters $\tilde{a}_i$ and $\tilde{b}_i$ of the Beta distribution are generated by $\tilde{a}_i \sim U(15,25)$ and $\tilde{b}_i = \beta \tilde{a}_i$ with $\beta_i$ independently drawn from $U(1.0 + q, 1.15 + q)$, where $q \sim U(0.005, 0.015)$. Then, the desired distributional sequence can be obtained by substituting the parameters of the Beta distributions generated above into equation (22), and the result is denoted as $F_{III} = \{f_i^{III}(x) : i = 1, \cdots, n\}$ and the corresponding representative sample data ($n = 100$ and $k^* = 50$) are visualized in Figure 7.

Comparing the simulated distributional sequences generated by the three models shown in Figures 5–7, the PDF-valued data associated with the first two models can be viewed as the strong-change case since the
corresponding mean functions (represented by the dashed bold lines) have changed dramatically after the change-point. In contrast, the mean function of the PDFs generated by Model III only exhibits a slight change after the change-point; thus, it can be viewed as the mild-change case. Moreover, one can see from Figure 6 that the PDFs generated by Model II are aligned functional data, because the underlying Beta distributed random variables share the same expectation. It is worth noting that the expectation, namely, the mean of the underlying random variable, is different from the mean function of the PDFs calculated by equation (6). Such a model is to simulate the fact that the monitoring data may share the same mean but follow different distributions. One can learn from this model that the changes in the distributions of the monitoring data cannot be detected by merely using the scalar sample mean of the raw measurements (belong to scalar data) in some situations, because the scalar sample mean only reflects the information of the first-order moment of the distribution. In contrast, in our proposed functional data analytic-based distributional change-point detection method, the PDFs themselves are directly analyzed as the random objects; thus, more complete information of the probability distributions can be preserved for investigation. In practical applications, the mild-change case represented by Model III is analogous to the distributional information break of DSF data caused by a minor condition change of the structure. In this case, the abrupt change in the distributions is not easy to be discovered by manual inspection, whereas the proposed method can provide a more effective tool for detecting such a minor change hidden in the distributional sequence because the decision is made according to the result of hypothesis testing at a given significance level.

Algorithm 2. Generate $N_o$ outlying PDFs

1: Set $\text{Out}_{PDF} = \emptyset$
2: for $i = 1$ to $N_o$ do
   Generate $z \sim U(0,1)$
   if $z > 0.7$ then
   \begin{align*}
   \text{Generate} \quad \mu_1 & \sim U(0.3,0.4), \quad \mu_2 \sim U(0.6,0.7), \\
   a_1 & \sim U(8,14), \quad a_2 \sim U(15,20) \\
   \text{Compute} \quad f_i^{\text{out}}(x) &= 0.5 \cdot f_{\text{Beta}} \left( x; a_1 \frac{a_1}{\mu_1} - a_1 \right) + 0.5 \cdot f_{\text{Beta}} \left( x; a_2 \frac{a_2}{\mu_2} - a_2 \right) \\
   \text{else} \quad \text{Generate} \quad y & \sim U(0,1), \quad a \sim U(2,5), \quad b \sim U(13,16), \\
   c & \sim U(17,22) \text{ and } d \sim U(2,5) \\
   \text{Compute} \quad f_i^{\text{out}}(x) &= f_{\text{Beta}}(x; a,b) \cdot \chi_{(0.5,1)}(y) + f_{\text{Beta}}(x; c,d) \cdot \chi_{[0,0.5]}(y) \\
   \text{end if}
   \end{align*}
   Set $\text{Out}_{PDF} \leftarrow f_i^{\text{out}}(x)$
3: Output $\text{Out}_{PDF} = \{f_i^{\text{out}}(x)\}_{i=1}^{N_o}$

To simulate data contamination, the procedure outlined in Algorithm 2 is firstly used to generate $N_o$ outlying PDFs; then, $N_o$ elements in the raw distributional sequence are randomly selected to be replaced by the generated outlying PDFs. For notation simplification, the three contaminated distributional sequences are still denoted as $F_i, F_{II}$, and $F_{III}$.

Next, the proposed distributional change-point detection method is applied to the contaminated distributional sequences to investigate its performance in resisting the effects of outlying PDFs. For each of the three models, the number of PDFs in the simulated distributional sequence is set to 100, the change-point location $k^*$ is fixed at 50, and the outlier contamination rate is set to $a = 20\%$ (i.e., $N_o = 20$). In the change-point detection, the same argument setting as that in Simulation study I is adopted. The above detection experiments are repeated for 500 times with the distributional data regenerated in each repetition. The absolute errors of the estimated change-point locations (i.e., $e = |\hat{k} - k^*|$) in the 500 repeated experiments are summarized as the boxplots displayed in Figure 8(a) for the three data-generating models. For comparison, the absolute errors of the estimated change-point locations obtained by using the uncontaminated data (i.e., the raw simulated distributional data before inserting the synthetic outlying
PDFs) are also presented as the boxplots shown in Figure 8(b). Comparing the results, one can see that the detection results associated with the three models have all been affected by the outlying PDFs, because a portion of the estimated change-points using the contaminated data have deviated from the true change-points especially for the mild-change case (i.e., Model III). In the significance tests, the calculated p-values associated with Models I and II are all less than the pre-specified significance level (i.e., $\alpha = 0.05$) in the 500 repeated experiments, suggesting that all the detected change-points in the two strong-change cases have passed the significance tests (meaning that the distributional sequences have change-points at the given significance level). However, for Model III, a certain number of the detected change-points using the contaminated data fail to pass the corresponding significance tests. Obviously, this result is not consistent with the facts that all the distributional sequences generated by Model III did exist mean breaks, and the invalid detections are mainly attributed to the interference of data contamination.

The above simulation results indicate that effective measures are generally required to deal with the outlying PDFs, otherwise the detection results might be deceptive, especially for the case of mild-change. To address this issue, we have recommended the distributional outlier cleaning strategy outlined in Algorithm 1. Next, the contaminated distributional data associated with Models I-III simulated above will further be utilized to evaluate the performance of the proposed change-point detection method after fusing the distributional outlier cleaning strategy. The Tree-Distance distributional outlier detection method proposed by Lei et al. is employed to detect and remove the outlying PDFs using the default parameter settings (listed in Table A-2 of the online supplement in Lei et al.). After implementing the distributional outlier cleaning, the accuracies of the change-point detection in the 500 repeated experiments are displayed in Figure 9. Compared to their counterparts shown in Figure 8(a), one can see that the absolute errors of the change-point location estimates have been concentrated around zero, similar to the case without contamination shown in Figure 8(b). Moreover, after implementing the distributional outlier cleaning, the calculated p-values in the 500 repeated experiments associated with the three Models are all less than the pre-specified significance level of $\alpha = 0.05$. Such results suggest that the alternative hypothesis (i.e., there exist change-point in the distributional sequence) in each experiment should be accepted at the given significance level. In other words, the outlier-induced invalid detections encountered earlier have disappeared after distributional outlier cleaning.

**Case study using the cable-tension monitoring data of a long-span bridge**

In this section, the proposed method is applied to detect and locate the abrupt changes occurred in the distributions of the DSF data (extracted from the cable-tension monitoring data) involved in cable condition assessment for a long-span cable-stayed bridge. Researchers have found that the ratio of the cable-tensions associated with a cable pair located at the same cross-section of the bridge deck is a feature that is sensitive to the change of cable condition but less sensitive to the change of external loads (e.g., traffic loads). The abrupt changes in the distributions of the cable-tension ratios can provide critical information for diagnosing the changes in the structural conditions of stay cables (see Li et al. for a detailed discussion). In bridge engineering, except for stay cables, similar distribution-change-based
structural condition assessment method has also been successfully applied in diagnosing the local health conditions of the bridge’s steel deck (see Wei et al. for a detailed discussion). To date, effective methods for automatically detecting such distributional changes have not been well developed in SHM. It is worth noting that automatic diagnosis is of crucial interest for practical engineering applications, especially for large scale structural systems. Take the cable-stayed bridge investigated in Li et al. for an example: the cable system is composed of 84 pairs of cables; it is impractical to examine the changes in the distributions of the DSF data by manual inspection for all cable pairs during the whole monitoring period, whereas the distributional change-point detection method developed in this study has good potential for such applications.

In this case study, the investigated data are also the cable-tension monitoring data of the long-span cable-stayed bridge studied in Li et al. The main axis of the bridge is north-south directed, and the 84 pairs of cables are symmetrically located at the edges of the upstream and downstream sides of the bridge. For facilitating cable pair localization, the 84 pairs of cables are labeled as CP-1 to CP-84 from the south to north (see Figure 10). The cable-force monitoring data associated with three representative cable pairs indexed by CP-67, CP-26, and CP-19 are selected for investigation. For each cable pair, 300 days of monitoring data are extracted from the time periods listed in Table 1 to calculate the cable-tension ratio defined in Li et al. and the latter can be regarded as the damage-sensitive feature. Obviously, the 300 selected days are not consecutive during the investigated time span, because some days’ data with poor quality are abandoned. It is worth noting that the calculation of the cable-tension ratio involves the measurements from the two cables located at the same cross-section; thus, only the data with high quality from the both cables are selected for study. Using the same data processing procedures (including the data pre-processing for removing the ambient factor) described in Subsection 3.1 of Li et al., the feature data can be extracted from the raw data. The only difference is that the logarithm computation given in equation (13) in Li et al. is abandoned, the main reason is that the distributions in our methods are assumed to be finitely supported on a compact interval $[a,b]$ and the logarithm map will take the data to be valued in the interval of $(-\infty, +\infty)$. Therefore, the extracted feature data in this study is the exponential of the cable-tension ratio defined in equation (13) in Li et al. However, for convenience, the extracted feature data is also called the cable-tension ratio (CTR) data throughout the rest of this study.

Before estimating the distributions, a further data processing consisting of scalar outlier filtration and scale transformation is conducted to the CTR data. Take the CTR data of CP-67 as an example to illustrate the procedure, the CTR data of the other two cable pairs can be processed in a similar way. Firstly, the classic boxplot is employed to detect and filter out the outliers contained in the raw CTR data. It is worth noting that the outlier handled here is scalar outlier, which is different from the functional outlier addressed in Algorithm 1. Figure 11 displays the comparison
Table 1. The basic information of the extracted distributional data associated with cable pairs CP-67, CP-26, and CP-19.

| Cable pair | CP-67 | CP-26 | CP-19 |
|------------|-------|-------|-------|
| Time period | March, 2008–April, 2012 | April, 2008–March, 2012 | Nov, 2007–Jun, 2010 |
| Number of extracted PDFs (n) | 300 | 300 | 300 |
| True change-point (k∗) | 162 | 145 | 213 |
| Occurrence time of the true change-point | 26 Sep, 2011 | 26 Sep, 2011 | 02 Mar, 2010 |

![Figure 11](image-url) Visualizations for a piece of the raw CTR data associated with cable pair CP-67: (a) before scalar outlier filtration and (b) after scalar outlier filtration.

for a piece of CTR data before and after the scalar outlier filtration. Such outliers randomly appear in the CTR data sequence, which are most likely to be caused by unknown interferences rather than damages. If the structure experiences a damage, the pattern of the CTR data would be expected to change. Then, the common support of the distribution (i.e., the domain where the corresponding PDF defined on) followed by the CTR data (To make the estimated common support more general, the common support is estimated using the CTR data from a wider time span (about six years) than that listed in Table 1) is estimated using the method given in equation (15) of Chen et al., and then, the CTR data are scaled to \([0, 1]\) through the support transformation described in Appendix 4 of Chen et al. Finally, the extracted 300-day CTR data are split into daily segments, and the PDF associated with each day is estimated by using the kernel density estimation technique.

By the above processing, a total of 300 PDFs can be extracted from the CTR data for each cable pair, and the PDFs are ordered in time to form the distributional sequence for further investigation. The resulting distributional sequences associated with CP-67, CP-26, and CP-19 are denoted as

\[ F_{67} = \{ f_{i}^{67}(x) : i = 1, \ldots, 300 \}, \]

\[ F_{26} = \{ f_{i}^{26}(x) : i = 1, \ldots, 300 \}, \]

\[ F_{19} = \{ f_{i}^{19}(x) : i = 1, \ldots, 300 \}, \]

respectively. Figure 12 visualizes the functional curves (first column) and the heatmaps (second column) of the distributional sequences. In the heatmap, the horizontal axis corresponds to the indices of the PDFs in the distributional sequence, and the vertical axis corresponds to the independent variable \(x\) of the PDF.

One can see from the right panel of Figure 12(a) that the heatmap of the distributional sequence associated with CP-67 exhibits a significant dislocation near the middle position, indicating that a sharp change has occurred to distributions of the CTR data, which is a symptom that the cable probably experienced a significant condition change at that moment. For CP-26, the colors of the heatmap shown in the right panel of Figure 12(b) are obviously different between data coming before and after the middle position, which is mainly attributed to the heights of the PDF-curves have reduced significantly after the middle time of the investigated period. Such a phenomenon is similar to that exhibited in the data generated by Model II in Simulation study II. The strong abruptly change in the pattern of the distributions of the DFS data of CP-26 is a symptom that this cable also probably experienced a significant condition change at that moment. For CP-19, a slight dislocation, located near \(i = 200\), can be observed in the corresponding heatmap shown in the right panel of Figure 12(c), which may correspond to a mild condition change event experienced by the cable.

The extracted distributional sequences associated with CP-67 and CP-26 can be viewed as the strong-change case, whereas the one associated to CP-19 can be viewed as the mild-change case. Further manual inspection finds that the true change-points in the three distributional sequences are located at \(i = 162, i = 145, \) and \(i = 213\) for CP-67, CP-26, and CP-19, respectively, and the results along with the occurrence time are reported in the fourth and fifth rows of Table 1, respectively. The change-points in the two strong-change cases occurred in the same day (i.e., 26 Sep, 2011), and the occurrence time is consistent with one representative feature change event caused by suspected cable condition change shown in Fig 8(c) of Li et al., (i.e., Cable SJ11, which corresponds to CP-32 in this study). The change-point in the mild-change case (i.e., CP-19) occurred in 02 Mar, 2010, and this occurrence time is consistent with a representative feature change event shown in Fig 8(e) of Li et al., (i.e., Cable SA03. Actually, it is the same cable with CP-19 in this study). By comparing to the feature change events reported in Li et al., the three change-points listed in Table 1 are highly suspected to be attributed to cable condition changes. Therefore, these change-points are of
critical concern in structural condition assessment, automatically detecting and locating them would definitely benefit the self-diagnostics.

Moreover, all the three distributional datasets contain a certain number of outlying PDFs (see, e.g., the data presented in the dashed rectangles of Figure 12 that possess a feature distinctly incongruous with the majority of the data). In contrast to the global pattern change phenomenon observed in the distributional sequences demarcated by the change-points listed in Table 1, the feature change represented by the outlying PDFs shown in the dashed rectangles of Figure 12 only last for a short time, and then, the data quickly get back to obey the common pattern followed by the majority of the data. Obviously, such distributional outliers are most probably attributed to some unknown interferences rather than the changes of cable conditions, thus they should be classified as data contamination. Meanwhile, such a phenomenon also demonstrates that the scalar outlier filtration performed previously in the data pre-processing procedure cannot effectively remove all
The change-point detection results of the extracted distributional sequences associated with cable pairs CP-67, CP-26, and CP-19.

| Cable pair | CP-67 | CP-26 | CP-19 |
|------------|-------|-------|-------|
| True change-point ($k^*$) | 162   | 145   | 213   |
| Estimated change-point ($\hat{k}$) | Scheme I: 162 | Scheme I: 145 | Scheme I: 185 |
| Scheme II: 162 | Scheme II: 145 | Scheme II: 211 |
| Hypothesis testing | Scheme I: Accept $H_A$ | Scheme I: Accept $H_A$ | Scheme I: Accept $H_A$ |
| Scheme II: Accept $H_A$ | Scheme II: Accept $H_A$ | Scheme II: Accept $H_A$ |

kinds of data corruptions. This is the reason why we also recommend an additional distributional outlier cleaning processing as described in Algorithm 1.

Next, the proposed method is employed to detect and locate the change-points contained in the three distributional sequences of the CTR data. For comparison, the following two detection schemes are considered:

- Scheme I: Do not perform distributional outlier cleaning before change-point detection.
- Scheme II: Perform distributional outlier cleaning before change-point detection.

Scheme II corresponds to the detection strategy in Algorithm 1. Here, the distributional outlier detection methods in Step 1 and Step 2 of Algorithm 1 are both considered. For the Tree-Distance detection method in Step 1, the argument setting is the same as that in Simulation study II; for the QF-FDO outlier detection method (see Lei et al.39 for the implementation details) in Step 2, the detection region is selected as [0.2, 0.8], and the whisker parameters of the boxplot-based detectors associated with the MO- and VO-directions are set to 1.5 and 2.5, respectively.

In the change-point detection, the argument setting is the same as that in Simulation study I. The detection results are reported in Table 2. In the significance tests, the calculated $p$-values associated with the three distributional sequences are all less than the pre-specified significance level (i.e., $\alpha = 0.05$) for both the detection schemes, suggesting that the distributional sequences contain change-points (i.e., the alternative hypothesis $H_A$ should be accepted). Comparing the estimated locations of the change-points, one can see that the results associated with CP-67 and CP-26 are both consistent with the true locations; however, for CP-19, the change-point identified by the first detection scheme significantly deviates from the true location, whereas the one identified by the second detection scheme is very close to the true location.

This case study further validates the effectiveness of the proposed distributional change-point detection method as well as demonstrates its practical utility in SHM. In addition to providing an automatic detection tool for testing the existence of abrupt changes in the distributions of the DSF data, it can also make a more objective judgment on the existence of a change through the hypothesis testing. It is worth noting that an abrupt distributional change detected in the DSF data can only signify that the cable may have experienced a damage event, but it is meaningful to structural condition diagnosis in the sense of automatically uncovering the suspected symptoms of crucial interest. In practical applications, further inspection is usually required to investigate and verify whether the detected change is attributed to a damage event or not as well as determine which type of the damage is. From this point of view, the change-point detection technique is an attractive data mining tool for providing useful information for aiding the condition-based inspection or maintenance.

Additional discussions on change-point technique in SHM applications

As defined by Farrar et al.,3,4 the damage (in a broad sense) of a structure system is a change occurred to the system that will cause adverse effects to the system’s performance. The structural changes of the system can usually be reflected as the changes in the damage-sensitive feature data extracted from the SHM data. This enables us to shift the problem of damage detection to the problem of pattern change detection in the DSF data, and the latter is a core content in the data-driven framework for damage detection. Such data-driven approaches do not require to establish sophisticated physical models (e.g., the finite element model) but highly rely on advanced data analysis techniques for damage-sensitive feature extraction and feature-pattern change detection. The change-point detection technique is of particular interest in such applications, in addition to provide an automatic tool for DSF-pattern change detection, the associated significance testing also makes the change-point technique an attractive choice in discovering the subtle change in the DSF data that may signify the occurrence of an early damage. The case study in this article regarding structural condition assessment for stay cables of a long-span bridge provides concrete evidence to demonstrate the practical utility of the change-point technique in engineering.
The change-point detection method can not only make a judgment about the existence of a sudden change in the DSF data under a given significance level, it can also locate the occurrence time of the change through the change location estimation. Change-time locating is another attractive feature of the change-point detection technique for SHM applications. It can provide important information for the inspector to quickly screen out the abnormal or extreme events (e.g., unexpected truck collision, ship collision, passing over of overloading trucks, and typhoon) that may cause the abrupt change in the DSF data, so as to investigate and verify whether the detected change is attributed to a damage event of the structure.

In addition to structural condition assessment, the distributional change detection method can also provide important guidance for model switching for some statistical methods dealing with distributional data employed for SHM applications. So far, several different distributional data analytic methods have been successfully applied in SHM for different tasks [35,40]; however, if the investigated distributional data changed, the corresponding statistical models should might be adjusted or re-trained accordingly; otherwise, it would lead to unreliable analysis or predictions.

Conclusion

This paper proposed a functional data analytic method for detecting abrupt changes occurred in the distributions of extracted damage-sensitive feature (DSF) data, and it is applied to distributional pattern change detection involved in data-driven condition assessment of structures. The effectiveness of the proposed method has been validated by both synthetic and real monitoring data. The following conclusions can be drawn:

1. The proposed change-point model is constructed using the linear structure of the Bayes space after embedding the PDFs into the Bayes space. Such a model not only has an advantage in the sense of interpretability but also shows better detection accuracy. In the simulation study, the proposed method significantly outperforms a competing method that directly treats PDFs as ordinary functional data.

2. The proposed change-point detection method can provide an effective tool for automatically diagnosing the abrupt change occurred in the distributions of the DSF data extracted from the measurements of structural responses. The decision about the existence of the abrupt change is based on the results of hypothesis testing at a given significance level; thus, it can provide a more rational judgment especially when the change is small and indistinguishable by manual inspection.

3. The change-point detection for the strong-change case is less sensitive to the outlying PDFs presented in the raw distributional data, whereas the detection results of the mild-change case can be seriously affected. The distributional outlier cleaning strategy is a simple and effective measure to cope with the outlying PDFs, and relative simulation and real data studies have demonstrated that it can significantly improve the detection performance under data contamination.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (Grant Nos. 51908166), the National Key R&D Program of China (Grant No. 2018YFB1600202), China Postdoctoral Science Foundation (Grant No. 2019M661287), and Postdoctoral Science Foundation of Heilong Jiang province.

ORCID iDs

Zhicheng Chen https://orcid.org/0000-0002-1587-5892
Hui Li https://orcid.org/0000-0001-9198-3951
Shiyin Wei https://orcid.org/0000-0001-6204-7128

References

1. Sun L, Shang Z, Xia Y, et al. Review of bridge structural health monitoring aided by big data and artificial intelligence: from condition assessment to damage detection. J Struct Eng 2020; 146(5): 4020073.

2. Yi TH, Huang HB and Li HN. Development of sensor validation methodologies for structural health monitoring: a comprehensive review. Measurement 2017; 109: 200–214.

3. Farrar CR, Duffey TA, Doebling SW, et al. A statistical pattern recognition paradigm for vibration-based structural health monitoring. In: Proceedings of the 2nd international workshop on structural health monitoring, Stanford, CA, 8–10 September 1999.

4. Farrar CR and Worden K. Structural health monitoring: a machine learning perspective. Chichester, UK: John Wiley & Sons, 2012.

5. Sohn H, Farrar CR, Hunter NF, et al. Structural health monitoring using statistical pattern recognition techniques. J Dyn Sys Meas Control 2001; 123(4): 706–711.

6. Worden K and Cross EJ. On switching response surface models, with applications to the structural health monitoring of bridges. Mech Syst Signal Process 2018; 98: 139–156.
7. Jung HK, Jo H, Park G, et al. Relative baseline features for impedance-based structural health monitoring. J Intel Mat Syst Str 2014; 25(18): 2294–2304.

8. Aue A and Horváth L. Structural breaks in time series. J Time Ser Anal 2013; 34(1): 1–16.

9. Catbas FN, Gokce HB and Gul M. Nonparametric analysis of structural health monitoring data for identification and localization of changes: Concept, lab, and real-life studies. Struct Health Monit 2012; 11(5): 613–626.

10. Nigro MB, Pakzad SN and Dorvash S. Localized structural damage detection: a change point analysis. Comput Aided Civ Inf 2014; 29(6): 416–432.

11. Pozo F, Arruga I, Mujica LE, et al. Detection of structural changes through principal component analysis and multivariate statistical inference. Struct Health Monit 2016; 15(2): 127–142.

12. Liao Y, Kiremidjian AS, Rajagopal R, et al. Structural damage detection and localization with unknown postdamage feature distribution using sequential change-point detection method. J Aerospace Eng 2019; 32(2): 4018149.

13. Mariani S and Cawley P. Change detection using the generalized likelihood ratio method to improve the sensitivity of guided wave structural health monitoring systems. Struct Health Monit 2021; 20(6): 3201–3226.

14. Parker BE Jr, Ware HA, Wipf DP, et al. Fault diagnostics using statistical change detection in the bispectral domain. Mech Syst Signal Process 2000; 14(4): 561–570.

15. Li Y, Bao T, Shu X, et al. Data-driven crack behavior anomaly identification method for concrete dams in long-term service using offline and online change point detection. J Civ Struct Health 2021; 11: 1449–1460.

16. Vagnoli M and Remenyte-Prescott R. An ensemble-based change-point detection method for identifying unexpected behaviour of railway tunnel infrastructures. Tunn Undergr Sp Tech 2018; 81: 68–82.

17. Wang QA, Zhang C, Ma ZG, et al. SHM deformation monitoring for high-speed rail track slabs and Bayesian change point detection for the measurements. Constr Build Mater 2021; 300: 124337.

18. Müller H-G. Special issue on “functional and object data analysis”: guest editor’s introduction. Can J Stat 2022; 50(1): 8–19.

19. Petersen A, Zhang C and Kokoszka P. Modeling probability density functions as data objects. Economet Statist 2021; 21: 159–178, DOI: 10.1016/j.eosta.2021.04.004

20. Li S, Wei S, Bao Y, et al. Condition assessment of cables by pattern recognition of vehicle-induced cable tension ratio. Eng Struct 2018; 155: 1–15.

21. Wei S, Zhang Z, Li S, et al. Strain features and condition assessment of orthotropic steel deck cable-supported bridges subjected to vehicle loads by using dense FBG strain sensors. Smart Mater Struct 2017; 26(10): 104007.

22. Worden K and Manson G. The application of machine learning to structural health monitoring. Phil Trans R Soc A 2007; 365: 515–537.

23. Salehi H, Burgueño R, Chakraborty S, et al. A comprehensive review of self-powered sensors in civil infrastructure: State-of-the-art and future research trends. Eng Struct 2021; 234: 111963.

24. Alavi AH, Hasni H, Jiao P, et al. Fatigue cracking detection in steel bridge girders through a self-powered sensing concept. J Constr Steel Res 2017; 128: 19–38.

25. Azimi M and Pekcan G. Structural health monitoring using extremely compressed data through deep learning. Comput Aided Civ Inf 2020; 35(6): 597–614.

26. Ramsay JO and Silverman BW. Functional Data Analysis. 2nd ed. New York, NY: Springer, 2005.

27. Berkes I, Gabrys R, Horváth L, et al. Detecting changes in the mean of functional observations. J Roy Stat Soc Ser B 2020; 71(5): 927–946.

28. Aue A, Rice G and Sonmez O. Detecting and estimating changes in dependent functional data. J Multivariate Anal 2012; 109: 204–220.

29. Aue A, Rice G and Sonmez O. Detecting and dating structural breaks in functional data without dimension reduction. J Roy Stat Soc Ser B 2018; 80(3): 509–529.

30. Kokoszka P, Miao H, Petersen A, et al. Forecasting of density functions with an application to cross-sectional and intraday returns. Int J Forecast 2019; 35(4): 1304–1317.

31. Chen Y, Lin Z and Müller H-G. Wasserstein regression. J Amer Statist Assoc 2021; DOI: 10.1080/01621459.2021.1956937

32. Padilla OHM, Athey A, Reinhart A, et al. Sequential nonparametric tests for a change in distribution: an application to detecting radiological anomalies. J Amer Statist Assoc 2019; 114(526): 514–528.

33. Horváth L, Kokoszka P and Wang S. Monitoring for a change point in a sequence of distributions. Ann Statist 2021; 49(4): 2271–2291.

34. Dubey P and Müller H-G. Fréchet change-point detection. Ann Statist 2020; 48(6): 3312–3335.

35. Chen Z, Bao Y, Li H, et al. A novel distribution regression approach for data loss compensation in structural health monitoring. Struct Health Monit 2018; 17(6): 1473–1490.

36. Egozcue J, Diaz-Barrero J and Pawlowsky-Glahn V. Hilbert space of probability density functions based on Aitchison geometry. Acta Math Sin (Engl Ser) 2006; 22: 1175–1182.

37. Van den Boogaart K, Egozcue J and Pawlowsky-Glahn V. Bayes Hilbert spaces. Aust New Zealand J Stat 2014; 54: 171–194.

38. Hron K, Menafoglio A, Tempf M, et al. Simplicial principal component analysis for density functions in Bayes spaces. Comput Statist Data Anal 2016; 94: 330–350.

39. Lei X, Chen Z and Li H. Functional outlier detection for density-valued data with application to robustify distribution to distribution regression. Arxiv, 2021, https://arxiv.org/abs/2110.00707

40. Lin Q and Li C. Kriging based sequence interpolation and probability distribution correction for gaussian wind field data reconstruction. J Wind Eng Ind Aerod 2020; 205: 104340.
Appendix

Introduction to the Bayes space

A detailed introduction to the Bayes space is beyond the scope of this article, this appendix only provides the basic notions and properties related to this study. Readers can refer to Egozcue et al.\textsuperscript{36} Van den Boogaart et al.\textsuperscript{37} and Hron et al.\textsuperscript{38} for more detailed descriptions of the Bayes space.

The Bayes space, denoted as $\mathbb{B}^2([a,b])$, is a functional space consisting of functions given by the following set

$$B^2([a,b]) = \left\{ f: [a,b] \rightarrow \mathbb{R}, \forall x \in [a,b] \text{ and } \int_a^b |\log f(x)|^2 ds < +\infty \right\} \quad (23)$$

where $\mathbb{R}$ denotes the real numbers. $B^2([a,b])$ is a linear space with its linear operations taking the same form as in equation (4). It can be easily verified that a PDF defined in the compact interval $[a,b]$ belongs to $B^2([a,b])$. $B^2([a,b])$ can further become a separable Hilbert space if it is equipped with the following inner product\textsuperscript{37,38}

$$f \cdot g = \frac{1}{2(b-a)} \int_a^b \log f(t) - \log g(t) dt ds, f, g \in \mathbb{B}^2([a,b]) \quad (24)$$

The inner product can naturally induce a norm denoted as $\|f\|_B = \sqrt{\cdot \cdot}$. Using the linear operations defined in equation (4), a distance can also be induced for the Bayes space, namely, $d_B(f, g) = \|f \oplus (-1) \otimes g\|_B$.

$\mathbb{B}^2([a,b])$ is isometrically isomorphic to $L^2([a,b])$.\textsuperscript{37,38} In other words, there exists a mapping from $B^2([a,b])$ to $L^2([a,b])$, denoted as $T : \mathbb{B}^2([a,b]) \rightarrow L^2([a,b])$, that satisfies the following properties

$$T(f \oplus g) = T(f) + T(g), \quad \forall f, g \in \mathbb{B}^2([a,b]) \quad (25a)$$

$$T(c \odot f) = c \cdot T(f), \quad \forall c \in \mathbb{R} \text{ and } \forall f \in \mathbb{B}^2([a,b]) \quad (25b)$$

$$d_B(f, g) = d_{L^2}(T(f), T(g)), \quad \forall f, g \in \mathbb{B}^2([a,b]) \quad (25c)$$

where $d_{L^2}$ is the $L^2$ distance associated with the $L^2([a,b])$ defined as

$$d_{L^2}(\xi, \eta) = \left( \int_a^b (\xi(t) - \eta(t))^2 dt \right)^{1/2}, \quad \xi, \eta \in L^2([a,b]) \quad (26)$$

One mapping that satisfies the properties given in equation (25) is the centered log-ratio (clr) transformation\textsuperscript{37,38} given by

$$\text{clr}[f](x) = \log f(x) - \frac{1}{b-a} \int_a^b \log f(t) dt, \quad x \in [a,b] \text{ and } f \in B^2([a,b]) \quad (27)$$