We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

A New Functional Clustering Method with Combined Dissimilarity Sources and Graphical Interpretation

Wenlin Dai, Stavros Athanasiadis and Tomáš Mrkvička

Abstract

Clustering is an essential task in functional data analysis. In this study, we propose a framework for a clustering procedure based on functional rankings or depth. Our methods naturally combine various types of between-cluster variation equally, which caters to various discriminative sources of functional data; for example, they combine raw data with transformed data or various components of multivariate functional data with their covariance. Our methods also enhance the clustering results with a visualization tool that allows intrinsic graphical interpretation. Finally, our methods are model-free and nonparametric and hence are robust to heavy-tailed distribution or potential outliers. The implementation and performance of the proposed methods are illustrated with a simulation study and applied to three real-world applications.

Keywords: depth, insurance, intrinsic graphical interpretation, robustness, statistical rankings

1. Introduction

Cluster analysis is a critical step in exploratory data analysis intended to identify homogeneous subgroups among observations. Cluster analysis is also widely used for functional data in tasks such as the classification of electrocardiogram curves in the diagnosis of cardiovascular ischemic diseases [1] and the extraction of representative wind behavior [2, 3]. The various functional data clustering methods described in the literature can generally be categorized into two subgroups: distance-based methods and filtering-based methods.

The distance-based methods involve the construction of a distance matrix with a specific metric; the clustering results may be derived with hierarchical or centroid-based clustering tools [3, 4]. The filtering-based methods involve the approximation of the curves with linear combinations of finite basis functions, such as splines and functional principal components, and the cluster analysis is conducted based on the coefficients or scores of finite dimensions [5–7]. The focus of this study is on distance-based methods. In this paper, we propose a new family of clustering algorithms based on the chosen functional ordering. The dissimilarity matrix is constructed via the chosen functional ordering, which is applied to the set of

IntechOpen

differences of all pairs of the functional data under investigation. Various functional ordering can be chosen, but we concentrate on orderings with intrinsic graphical interpretation. But any ordering that treats the sources equally can be used, including the modified band depth [8] and the simplicial band depth [9]. The choice of functional ordering with intrinsic graphical interpretation allows us to show the resulting clusters and a central region that attains a natural interpretation. I.e., All functions contained in the central region do not leave the plot of the central region, and all functions not contained in the central region leave the plot of the central region in at least one point. It has to be mentioned that the classical functional orderings mentioned above do not satisfy this natural condition, and therefore we will concentrate on functional orderings defined in [10].

Functional data differ in various ways, such as in magnitude, shape, phase, and dependence structure, and hence they are difficult to analyze when clusters exist from multiple perspectives. The existing methods either focus on a single type of variation or pool the various sources of variation with weightings that rely on a delicate selection procedure. Without balancing, the clustering results could be dominated by the component with the greatest absolute variation. In order to achieve some balancing between the various sources, it is possible to standardize the curves before applying existing methods, such as k-means or model-based methods. By "standardization", we mean that the marginal empirical distributions are standardized so that they have zero mean and unit variance. This approach is used in the simulation study in order to compare the performance of existing methods with the proposed methods.

Since the proposed procedure applies functional ordering, such that every part of the function is treated equally, the different sources of variation are combined in an equal manner. For univariate cases, it may combine the raw curves and the derivatives equally to measure the magnitude and shape variation simultaneously. For multivariate cases, it may combine the marginal curves and the covariance functions equally to account for both marginal and joint variation among curves. Furthermore, the proposed method provides a reasonable graphical interpretation of the clustering result. Finally, it inherits the robustness of functional orderings and can stably recover the clusters when abnormal observations contaminate the data.

The remainder of this paper is organized as follows. In Section 2, we define the new proposed procedure with an arbitrary functional ordering. Further, we review several functional orderings already defined in [10] which satisfy the intrinsic graphical interpretation. Finally, we study the metric properties of derived dissimilarity. In Section 3, we describe the simulation studies we conducted to assess the performance of the proposed methods and compare them with some existing competitors in cases where the combination of the various sources is of interest. In Sections 4–6, we demonstrate the effectiveness of our method with three real-world examples. The proposed methods will be available soon in the R package GET.

2. Description of methods

2.1 Dissimilarity matrix

Assume that the functions $f_i(x), i = 1, \ldots, s$ are observed at a fixed set of points $x_1, \ldots, x_d$, so that the functions can be represented as d-dimensional vectors $\mathbf{T}_i, i = 1, \ldots, s$. If the functions of interest are not observed at the same set of points, a nonparametric smoothing method can be applied to address the situation.

To induce dissimilarity measure from functional ordering, we construct the set of functional differences:

$$D_f = \{df_{ii'} = f_i(x) - f_{i'}'(x), \quad i, i' = 1, \ldots, s\}.$$

We remark here that $df \equiv 0$ is an element of D_f. We then apply a functional ordering to D_f and obtain the induced measure of centrality of $df_{ii'} = f_i(x) - f_{i'}'(x)$ as $M_{ii'}$. Finally, the dissimilarity between $f_i(x)$ and $f_{i'}(x)$ is defined as $d_{ii'} = 1 - M_{ii'}$, and this forms the dissimilarity matrix of $\{ f_i \}_{i=1}^{s}$. Such an ordering can take the form of any functional depth notions or rankings in the literature, such as the band depth and modified band depth [8], the simplicial band depth [9], the spatial functional depth [11], or the curve depth [12]. These notions naturally give equal treatment to the variations at each design point, compared with the norm-based methods such as L_1 or L_2 distances.

After a dissimilarity matrix is established, the partitioning around medoids procedure can be used to determine the given number of clusters. This produces a family of clustering algorithms that depends on the choice of the functional ordering.

In the following, we will discuss the possible choices of functional ordering. First, we assume functional orderings, which take different sources of the data variability equally. We call such ordering combined functional ordering. Such an approach is useful when the investigator wants to join different information about the data and combine them in one universal procedure. Second, we review several functional orderings which satisfy the intrinsic graphical interpretation.

Our proposed procedure then consists of the following steps:

1. Choose the appropriate data sources (e.g., raw data, derivative and second derivative)

2. Choose the functional ordering, which allows for intrinsic graphical interpretation and which gives the same weight to every chosen source (e.g., the studentized maximum ordering, the area rank ordering).

3. Compute the dissimilarity matrix

4. Apply partitioning around medoids

5. Plot the resulted clusters together with their central region with intrinsic graphical interpretation.

2.2 Combined functional ordering

We consider now functions $T_i(x), i = 1, \ldots, s'$ and specify their combined functional ordering. Various perspectives, such as different magnitudes and different shapes of the functions, can be used to order the functions. Here we provide a general method to combine these different perspectives in an equal manner. As suggested by [13], data transformation is an effective method to convert different types of variation into types that are easy to handle by the functional depth. Hence, various transformations could be applied to the raw functions to obtain the transformed data sets of interest, such as $V_1, \ldots, V_k$. These transformations are computed in the same fixed set of points $x_1, \ldots, x_d$; for instance, shifting each curve to zero means eliminates the magnitude variation, normalizing the centered curves by their L_2 norms, respectively, to extract pure shape information. In the case of multivariate functional data, each component of the data and their transformation could be treated similarly. Also, the covariance function

between the components can be added to take into account the dependence structure.

We denote with $V_k(T_{ij})$ the resultant curves of T_{ij} via the transformation V_k, and we can express the long vector as:

$$\mathbf{T}_i = (V_1(T_{i1}), \ldots, V_1(T_{id}), \ldots, V_k(T_{i1}), \ldots, V_k(T_{id})), \quad i = 1, \ldots, s' \tag{1}$$

We can then apply to them the corresponding ordering and hence construct the dissimilarity matrix. Note that each of the orderings to be introduced considers each element equally by ranking or scaling, so the desired perspectives of ordering are considered and treated equally in such a combined ordering. To enhance the interpretability of the clustering results, we focus only on the notions that satisfy the intrinsic graphical interpretation.

2.3 Functional ordering with intrinsic graphical interpretation

The following definition specifies the properties of *the global envelope that has an intrinsic graphical interpretation with respect to an ordering*. This definition was already used in [10] to define global envelope tests and central regions with graphical interpretation.

Definition 1: Assume a general ordering $\prec$ of the vectors $\mathbf{T}_i, i = 1, \ldots, s'$, that is induced by a univariate measure M_i. That is, $M_i \geq M_j$ iff $\mathbf{T}_i \prec \mathbf{T}_j$, which means that $\mathbf{T}_i$ is less extreme or as extreme as $\mathbf{T}_j$. (The smaller the measure M_i, the more extreme $\mathbf{T}_i$.) The $100(1-\alpha)\%$ global envelope $\left[\mathbf{T}^{(\alpha)}_{\text{low}j}, \mathbf{T}^{(\alpha)}_{\text{upp}j}\right]$ has *intrinsic graphical interpretation* (IGI) with respect to the ordering $\prec$ if:

1. $m_{(\alpha)} \in \mathbb{R}$ is the largest of the M_i such that the number of those i for which $M_i < m_{(\alpha)}$ is less than or equal to $\alpha s'$;

2. $T_{ij} < \mathbf{T}^{(\alpha)}_{\text{low}j}$ or $T_{ij} > \mathbf{T}^{(\alpha)}_{\text{upp}j}$ for some $j = 1, \ldots, d$ iff $M_i < m_{(\alpha)}$ for every $i = 1, \ldots, s'$;

3. $\mathbf{T}^{(\alpha)}_{\text{low}j} \leq T_{ij} \leq \mathbf{T}^{(\alpha)}_{\text{upp}j}$ for all $j = 1, \ldots, d$ iff $M_i \geq m_{(\alpha)}$ for every $i = 1, \ldots, s'$.

Let us call *the ordering with intrinsic graphical interpretation* such ordering, for which exists a global envelope with IGI with respect to this ordering. Remark here that $m_{(\alpha)}$ is not exactly the α quantile of M_i and that points 2 and 3 are equivalent. We kept points 2 and 3 to show the interpretability of the IGI. The simple ordering criterion based on L_∞ distance, $M_i = \max_j |T_{ij} - \overline{T}_j|$, clearly satisfies such a property, but it does not account for the changes in the marginal distribution of $T_{\cdot j}$ for different values of j [14, 15]. To address this problem, Myllymäki et al. [14] proposed studentized and directional quantile scaling of the maximum ordering, which also satisfies IGI. Furthermore, [15, 16] simultaneously defined extreme rank length ordering, which is based on the number of the most extreme pointwise ranks and satisfies IGI. Finally, [10] extended this family with continuous rank ordering, which is based on the continuous extension of pointwise ranking, and area rank ordering, which is based on the area with the most extreme continuous ranks. To the best of our knowledge, no other functional (respective multivariate) orderings satisfy IGI.

The definitions of all previously mentioned orderings are given in [10]. For the sake of completeness, we provide here a short list of these definitions.

2.3.1 Extreme rank length ordering

Let $r_{1j}, r_{2j}, \ldots, r_{s'j}$ be the raw ranks of $T_{1j}, T_{2j}, \ldots, T_{s'j}$, such that the smallest T_{ij} has rank 1. In the case of ties, the raw ranks are averaged. The two-sided pointwise ranks are then calculated as $R_{ij} = \min\left(r_{ij}, s' + 1 - r_{ij}\right)$. Consider now the vectors of pointwise ordered ranks $\mathbf{R}_i = \left(R_{i[1]}, R_{i[2]}, \ldots, R_{i[d]}\right)$, where $\left\{R_{i[1]}, \ldots, R_{i[d]}\right\} = \{R_{i1}, \ldots, R_{id}\}$ and $R_{i[k]} \leq R_{i[k']}$ whenever $k \leq k'$. The extreme rank length measure of the vectors $\mathbf{R}_i$ is equal to:

$$E_i = \frac{1}{s'} \sum_{i'=1}^{s'} (\mathbf{R}_{i'} \prec \mathbf{R}_i) \tag{2}$$

where

$$\mathbf{R}'_i \prec \mathbf{R}_i \Leftrightarrow \exists n \leq d : R_{i'[k]} = R_{i[k]} \forall k < n, \;\; R_{i'[n]} < R_{i[n]}.$$

The division by s' leads to normalized ranks that obtain values between 0 and 1. Consequently, the ERL measure corresponds to the extremal depth as defined in [16].

Let e_α be defined according to point 1 of Definition 2.3, and let $I_\alpha = \left\{i \in 1, \ldots, s' : E_i \geq e_{(\alpha)}\right\}$ be the index set of vectors less extreme than or as extreme as e_α. Then, the $100(1-\alpha)\%$ global extreme rank length envelope (or global extreme rank length central region) induced by E_i is:

$$\mathbf{T}^{(\alpha)}_{\text{low}\,k} = \min_{i \in I_\alpha} \; T_{ik} \quad \text{and} \quad \mathbf{T}^{(\alpha)}_{\text{upp}\,k} = \max_{i \in I_\alpha} \; T_{ik} \quad \text{for} k = 1, \ldots, d. \tag{3}$$

2.3.2 Global continuous rank ordering

The continuous rank measure is:

$$C_i = \min_{j=1, \ldots d} c_{ij} / \lceil s'/2 \rceil,$$

where c_{ij} are the pointwise continuous ranks defined as:

$$c_{ij} = \sum_{i'} \mathbf{1}\left(T_{i'j} > T_{ij}\right) + \frac{T_{[i+1]j} - T_{ij}}{T_{[i+1]j} - T_{[i-1]j}} \quad \text{for } i : T_{ij} \neq \max_{i'} T_{i'j} \text{ and } T_{ij} > \text{median}\left(T_{ij}\right),$$

$$c_{ij} = \exp\left(-\frac{T_{ij} - T_{[i-1]j}}{T_{[i-1]j} - \min_i T_{ij}}\right) \quad \text{for } i : T_{ij} = \max_{i'} T_{i'j},$$

$$c_{ij} = \sum_{i'} \mathbf{1}\left(T_{i'j} < T_{ij}\right) + \frac{T_{ij} - T_{[i-1]j}}{T_{[i+1]j} - T_{[i-1]j}} \quad \text{for } i : T_{ij} \neq \min_{i'} T_{i'j} \text{ and } T_{ij} < \text{median}\left(T_{ij}\right),$$

$$c_{ij} = \exp\left(-\frac{T_{[i+1]j} - T_{ij}}{\max_i T_{ij} - T_{[i+1]j}}\right) \quad \text{for } i : T_{ij} = \min_{i'} T_{i'j}.$$

$$c_{ij} = R_{ij} \quad \text{for } \; T_{ij} = \text{median}\left(T_{ij}\right),$$

Here, $T_{[i-1]j}$ and $T_{[i+1]j}$ denote the values of the functions, which are in a j-th element below and above T_{ij}, respectively (i.e., $T_{[i-1]j} = \max_{i':T_{i'j}<T_{ij}} T_{i'j}$ and $T_{[i+1]j} = \min_{i':T_{i'j}>T_{ij}} T_{i'j}$).

The $100(1-\alpha)\%$ global continuous rank envelope induced by C_i is constructed in the same manner as the global extreme rank length envelope.

2.3.3 Global area rank ordering

The area rank measure:

$$A_i = \frac{1}{\lceil s'/2 \rceil d} \sum_j \min\left(R_i, c_{ij}\right),$$

where.

$R_i = \min_j \{R_{ij}\}$ and R_{ij} are two-sided pointwise ranks defined above. The $100(1-\alpha)\%$ global area rank envelope induced by A_i is constructed in a manner similar to that of the global extreme rank length envelope.

2.3.4 Studentized maximum ordering

Because we construct a symmetric set of functions to compute the dissimilarity matrix, here we use only the symmetric studentized ordering. The above orderings are based on the whole distributions of $T_{\cdot j}, j = 1, \ldots, d$. It is also possible to approximate the distribution from a few sample characteristics. The studentized maximum ordering approximates the distribution of $T_{\cdot j}, j = 1, \ldots, d$ by the sample mean T_{0j} and sample standard deviation $\mathrm{sd}(T_{\cdot j})$. The studentized measure is:

$$S_i = \max_j \left| \frac{T_{ij} - T_{0j}}{\mathrm{sd}(T_{\cdot j})} \right|. \tag{4}$$

The $100(1-\alpha)\%$ global studentized envelope induced by S_i is defined by:

$$\mathbf{T}^{(l)}_{\mathrm{low}j} = T_{0j} - s_\alpha \mathrm{sd}(T_{\cdot j}) \quad \text{and} \quad \mathbf{T}^{(l)}_{\mathrm{upp}j} = T_{0j} + s_\alpha \mathrm{sd}(T_{\cdot j}) \quad \text{for} j = 1, \ldots, d, \tag{5}$$

where s_α is taken according to point 1 of IGI.

2.4 Dissimilarity matrix based on the combined ordering

In this section, we validate the dissimilarity matrix construction defined in Section 2.1 for studentized measure by showing that $d_{ii'} = S_{ii'}$ is a metric and for global area rank measure by showing that $d_{ii'} = 1 - A_{ii'}$ is a semi-metric. The latter means that the $d_{ii'} = 1 - A_{ii'}$ satisfies all properties of metric, except for the triangular inequality, which is violated in specific cases. The metric properties are usually required when choosing the distance measure, but it is not necessary for the partitioning around medoids algorithm, which is used to calculate the clusters afterward. Furthermore, our simulation study demonstrates that these specific cases, where the triangular inequality of global area rank measure is not satisfied, are not realized by functions appearing in real data studies. Furthermore, we provide a thorough check of satisfaction of the triangular inequality for global area rank measure in our implementation of the algorithm. Thus in practice, a user can check

this feature of the metric for particular data of interest. For any dataset considered by us in simulation and data studies, the triangular inequality was satisfied.

Theorem 1.1: Define the distance between $\mathbf{T}_i$ and $\mathbf{T}_{i'}$ as:

$$d_{ii'} = 1 - A_{ii'},$$

where $A_{ii'}$ is the global area rank measure of $T_i - T_{i'}$ on D_f. Then $d_{ii'}$ satisfies for any i, i':

1. Non-negativity: $d_{ii'} \geq 0$;
2. Identity of indiscernibles: $d_{ii'} = 0$ iff $\mathbf{T}_i = \mathbf{T}_{i'}$;
3. Symmetry: $d_{ii'} = d_{i'i}$.

Proof:

Non-negativity*: For the set D_f, there are s' curves. The set D_f contains a zero element, which is the deepest point of D_f. I.e. 0 is median in every coordinate. For the area ordering of these curves, we have that two-sided pointwise ranks of curve $\mathbf{T}_i - \mathbf{T}_{i'}$ is $R_{ii'j} \leq \lceil s'/2 \rceil$ and $R_{ii'} = \min_j \{R_{ii'1}, \ldots, R_{ii'd}\} \leq \lceil s'/2 \rceil$. Hence, we have $A_{ii'} \leq 1$, i.e., $d_{ii'} \geq 0$.*

Identity of indiscernibles: *$d_{ii'} = 0 \Leftrightarrow A_{ii'} = 1 \Leftrightarrow R_{ii'j} = \lceil s'/2 \rceil$ for every $j = 1, \ldots, d$ $\Leftrightarrow \mathbf{T}_i - \mathbf{T}_{i'}$ is the deepest curve of $D_f \Leftrightarrow \mathbf{T}_i = \mathbf{T}_{i'}$.*

Symmetry*: This property holds implicitely due to the symmetry of D_f.*

The fourth property of the metric, i.e.

4. Triangle inequality: $d_{ii'} + d_{i'k} \geq d_{ik}$, for any i, i' and k,

is not satisfied when $\mathbf{T_i} \equiv t_i$ for every i. The results of our simulation study suggest that if the system of data provides enough crossings of functions, then the triangle inequality is satisfied.

Theorem 1.2: Define the distance between $\mathbf{T}_i$ and $\mathbf{T}_{i'}$ as:

$$d_{ii'} = S_{ii'},$$

where $S_{ii'}$ is the studentized measure of $T_i - T_{i'}$ on D_f. Then $d_{ii'}$ is a valid metric.

Proof:

The first three properties obviously hold for the studentized difference distance. We prove the triangle inequality for $d_{ii'}$. Note that $df \equiv 0$ is an element of D_f, and hence the sample mean $T_{0j} = 0$ for $j = 1, \ldots, d$. Lets denote the sample standard deviation of the j-th element of D_f by $\mathrm{sd}(D_{\cdot j})$. Then, we have:

$$\begin{aligned} d_{ik} &= \max_j \left| \frac{T_{ij} - T_{kj} - 0}{\mathrm{sd}(D_{\cdot j})} \right| \leq \max_j \left\{ \left| \frac{T_{ij} - T_{i'j}}{\mathrm{sd}(D_{\cdot j})} \right| + \left| \frac{T_{i'j} - T_{kj}}{\mathrm{sd}(D_{\cdot j})} \right| \right\} \\ &\leq \max_j \left| \frac{T_{ij} - T_{i'j}}{\mathrm{sd}(D_{\cdot j})} \right| + \max_j \left| \frac{T_{i'j} - T_{kj}}{\mathrm{sd}(D_{\cdot j})} \right| \\ &= d_{ii'} + d_{i'k}. \end{aligned}$$

This completes the proof.

3. Simulation study

This section describes the intensive simulation studies we conducted to assess the empirical performance of the proposed clustering methods and compares this performance with those of the existing methods when the clusters demonstrate differences from various perspectives. For comparison, we also consider two clustering methods for functional data: the k-means methods available in the R package *fda.usc* [17] and the model-based clustering methods proposed by [18], which are available in the R package *fdapace* [19]. For the fairness of comparison, the standardization procedure is applied to normalize the empirical marginal distributions as described in Section 1 so that they can be combined equally.

Specifically, we consider the following five models on $t \in [0, 1]$:

- Class 1: $X(T) = 2T + e(T)$;
- Class 2: $X(T) = 2 - 2T + e(T)$;
- Class 3: $X(T) = 2\ \mathbf{1}(T > U) + e(T)$;
- Class 4: $X(T) = 1.5 + 2\ \mathbf{1}(T > U) + e(T)$;
- Class 5: $X(T) = 3 - 2.5T + e(T)$.

Here, U follows a uniform distribution on $[0.5, 0.6]$, and $e(T)$ is generated from a Gaussian process with zero mean and covariance function $\gamma(s,t) = \sigma^2 \exp\{-\phi|t-s|^\nu\}$, where $\sigma^2 = 0.2$, $\phi = 2$ and $\nu = 1$.

In addition, to assess the robustness of the proposed methods, we also consider another situation by replacing $e(T)$ with a multivariate-t distribution with two degrees of freedom, $t_2(\mu, \Sigma)$, where $\mu = 0$, and Σ is generated with $\gamma(s,t)$. The heavy

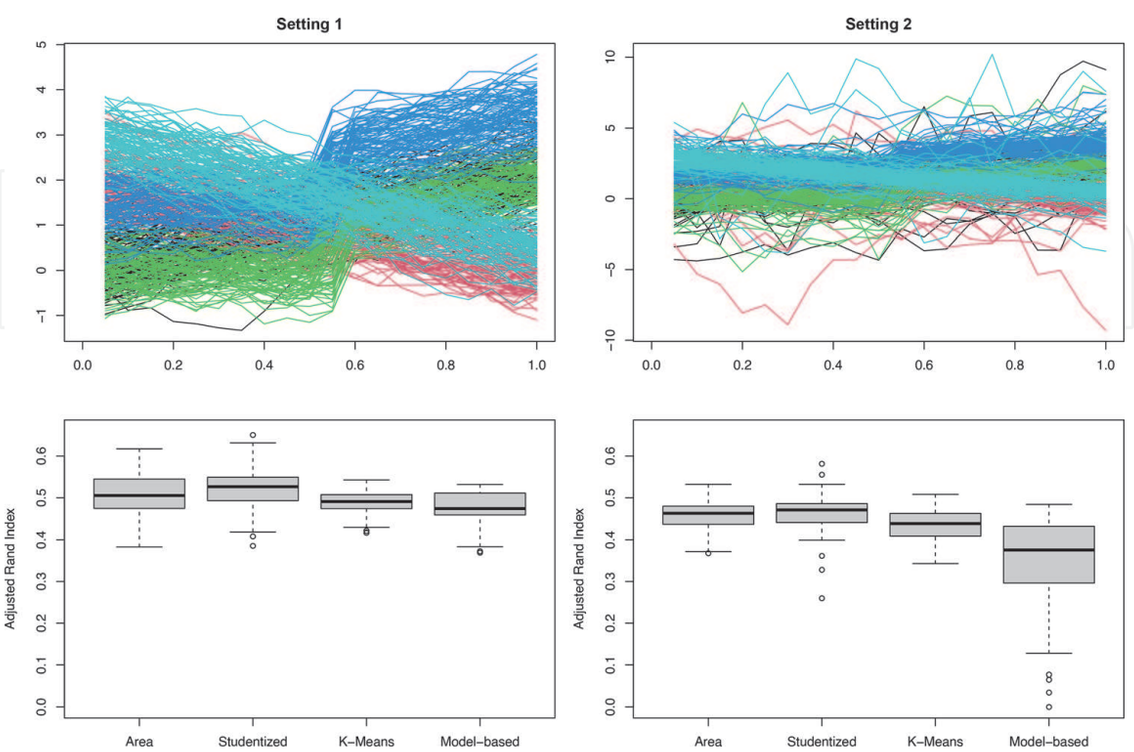


Figure 1.
Top panel: Realizations of two settings. Bottom panel: Adjusted Rand index of four clustering methods with the two settings.

tail property of the marginal distribution allows the data to be viewed as contaminated by some outliers, which are commonly encountered in practice. We generate 100 samples for each of the five classes with 20 equally spaced design points; as a result, 500 curves are clustered into five groups. The top panel of **Figure 1** demonstrates one realization of the simulated samples under two settings. To account for both the magnitude and the shape variation among clusters, we make two

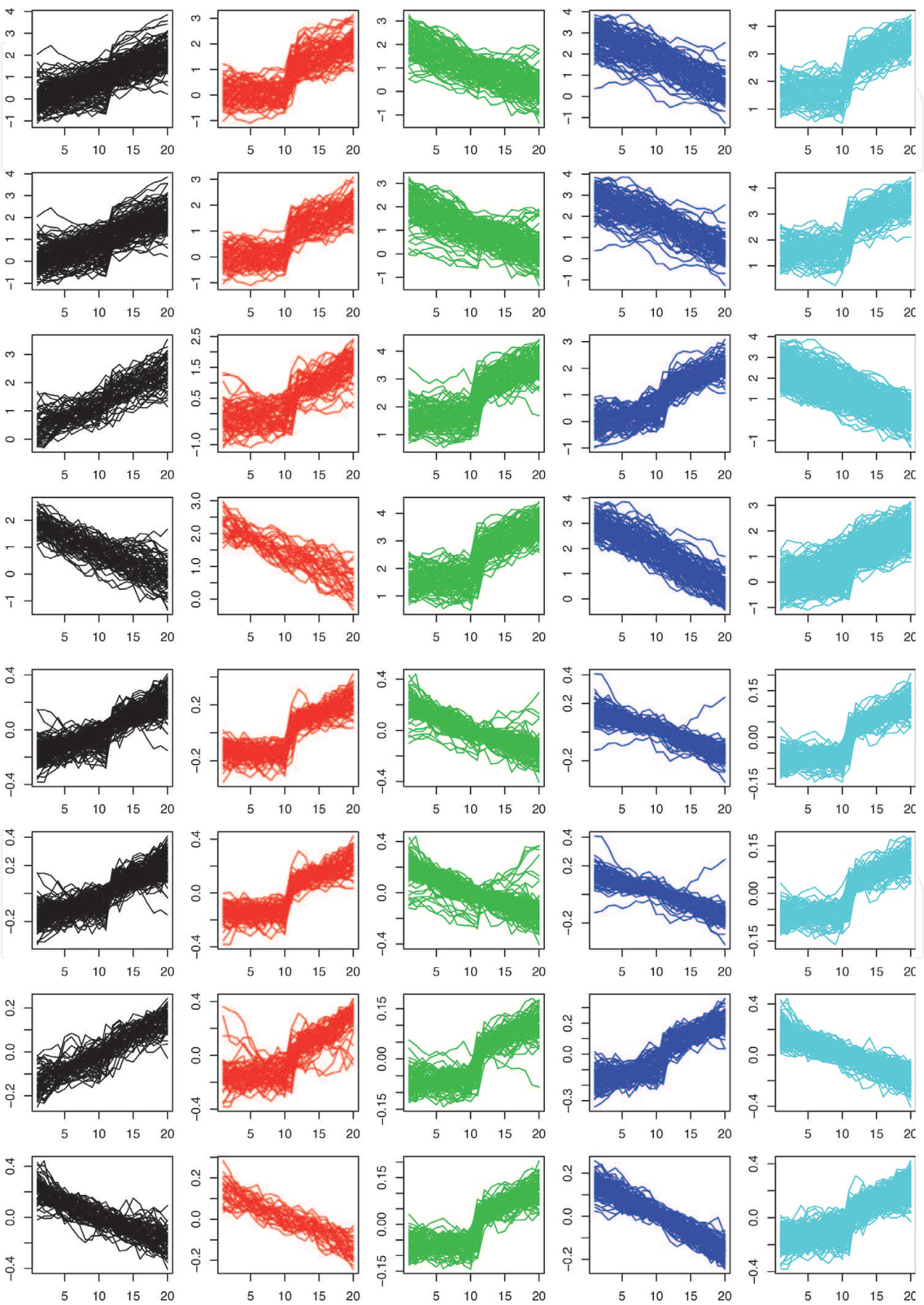


Figure 2.
Clusters for setting 1 visualized on raw curves (top panel) and normalized curves (bottom panel). In each panel, from top to bottom: Area, studentized, k-means, and model based.

transformations suggested by [13] to the raw curves, shifting the curves so that each has a zero mean and then normalizing the centered curves by their L_2 norm. We then bind the three components together as long vectors for clustering. For each run, we use the true number of clusters for all four methods and calculate the adjusted Rand index [20] to compare their clustering results. We repeat the procedure 100 times, and the results are reported in the bottom panel of **Figure 1**.

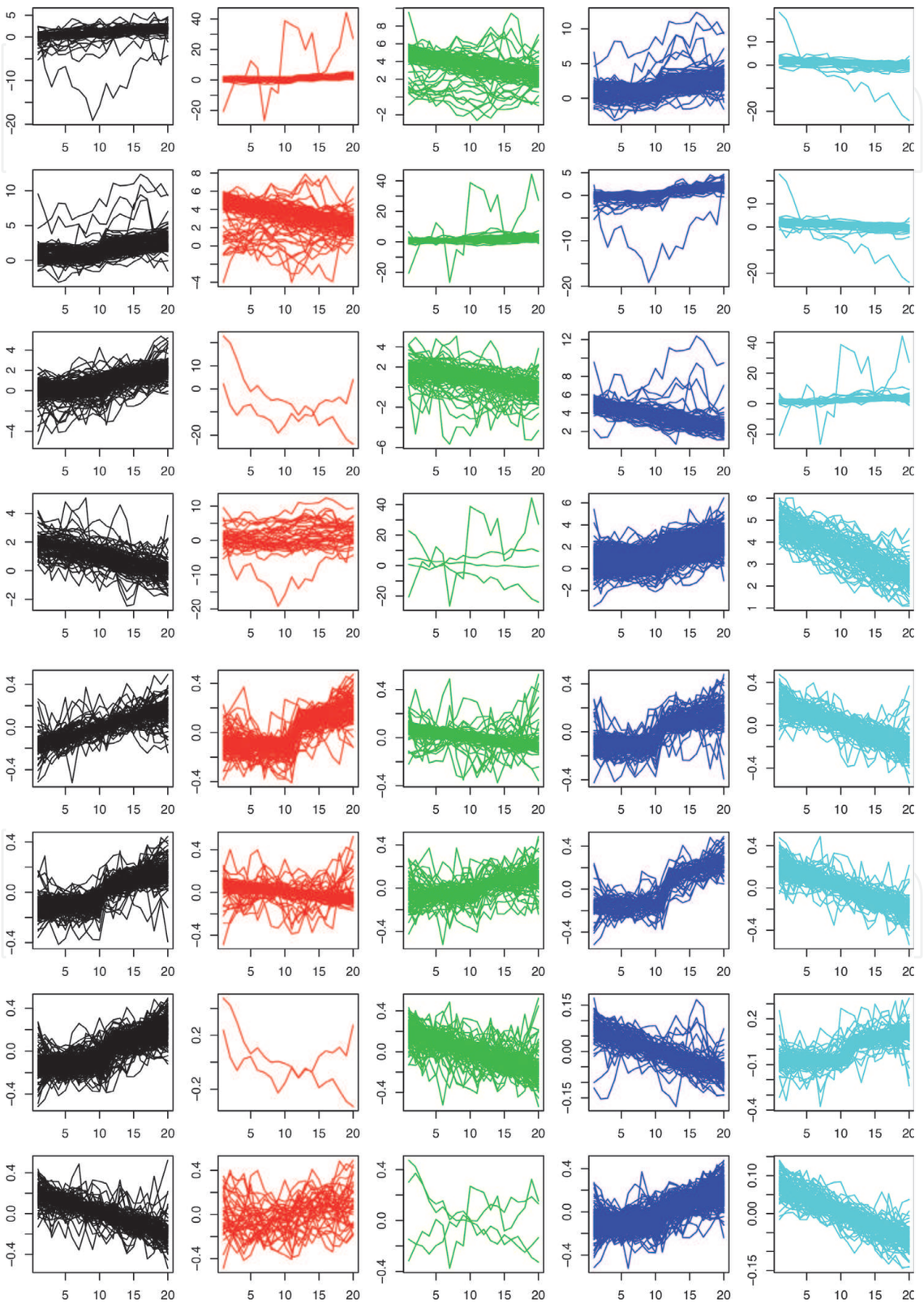


Figure 3.
Clusters setting 2 visualized on raw curves (top panel) and normalized curves (bottom panel). In each panel, from top to bottom: Area, studentized, k-means, and model based.

Note that in all cases of the simulation study, the triangular inequality of the area measure was satisfied for all combinations of curves.

Under the first setting, data are generated from a Gaussian process. With regard to the adjusted Rand index, the four methods are quite comparable but the proposed methods are slightly better than the other two. However, our methods recover much better the characteristics of the true clusters; see **Figure 2**, which illustrates one clustering result for each of the four methods with both raw curves and normalized curves. In contrast, the k-means method merges classes 2 and 5, and the model-based method merges classes 1 and 3.

As for the second setting, in which the marginal distribution becomes heavy-tailed, our methods obtain more robust clustering results than the other two methods and reach higher adjusted Rand indexes (**Figure 3**). The model-based method relies heavily on the Gaussian assumption and thus shows less satisfactory behavior. Again, our methods still accurately recover the patterns of each cluster, whereas the other two methods completely fail to reveal reasonable group structures. Specifically, both k-means and the model-based methods suggest a cluster with only a few curves, which indicates a clear misinterpretation of the situation.

4. Clustering of insurance penetration

Insurance consumption indicates the equilibrium of supply and demand of insurance products. For a given insurance market, the collection of total (Life and non-Life) yearly insurance consumption observations helps to explain the variation of insurance market development over time. A common measure of insurance consumption, and hence of insurance development, is insurance penetration (IP), defined as the ratio of insurance premiums on GDP. The pattern of the development variation is evident when one views the IP as a function of time, known as the IP curve.

In their effort to promote the European single insurance market through the integration process, European policymakers put emphasis on homogeneity and convergence aspects of development patterns of European insurance markets. That is equivalent to saying that they are interested in identifying a single group (cluster) of countries whose IP curves exhibit similarity in magnitude and shape. The clustering of European countries in terms of their IP curves provides a method for testing the magnitude and shape similarity of the insurance industry in Europe. In particular, functional clustering methods are appropriate for our data, given the time dependency in the observations.

IP curves (time-series data on IP) originated from the Swiss Re (2016) Database were analyzed by the proposed functional clustering (FC) method based on Area measure. The exploration concentrated on the IP curves of 34 European countries (EU and non-EU members) observed over 13 years between 2004 and 2016, that is, before, during, and post-financial and sovereign debt crises.

The FC method extracts the partitioning information from both the magnitude and the shape of IP curves. While the magnitude is captured in the IP curves, the shape is not straightforward to be detected. To this end, we performed two types of transformations on the raw IP curves to reveal their shape. First, the raw IP curves were centred relative to each country's average IP rate to mitigate the widely different magnitudes in the IP data. After this, the resulting centred IP curves were then normalized with their L2 norms to a unit norm (to have a length of 1). These transformations are proposed to extract shape information by [13] By normalizing the centred IP curves in this manner, we eliminate their amplitude signal, while we are only left with the shape signal of the raw IP curves.

For the FC method to run properly, the most suitable number of clusters must be determined. We chose 6 clusters even if the median value of all methods presented in the NBclust library of the R software is 5. Our choice is justified as it better serves the analysis and the characterization of the produced clusters.

Given the IP curves of each cluster, the FC method also provides a graphical representation, through the central regions, of the deepest central IP curves within each cluster. We are interested in the so-called marginal plot style approach of the clustering solution. This means that the central regions are computed separately for magnitude and shape to better express each cluster component's shape. Remark here that the proposed method also allows showing the central region with respect to the combined ordering with respect to the magnitude and shape together. The appearance of clusters is demonstrated by the deepest IP curve (solid curve) that corresponds to the medoid IP curve and the envelope of 50% central IP curves (gray area) that reflects the band where 50% of the IP curves surrounding the deepest are varied. See **Figures 4** and **5**. Note that the fraction of combinations of countries satisfying the triangular inequality with Area measure was 1 with respect to all combinations. With this visualization, we can describe the clusters that are produced by the FC method as follows:

Cluster 1: Developed insurance markets with middle-to-high IP levels and decreasing IP patterns in the whole period. This cluster includes Belgium, France, Ireland, Austria*, the UK, Portugal, Switzerland, Malta, Slovakia, and Germany. Cluster 2: Developing insurance markets with low-to-middle IP level and increasing IP pattern until 2010 and varying (decreasing) thereafter. This cluster of countries consists of Cyprus*, Turkey, Greece, and Luxemburg. Cluster 3: Developed insurance markets with middle-to-high IP levels and increasing IP patterns in the whole period. This cluster unites Finland*, Italy, Spain, Denmark, and the Netherlands. Cluster 4: Developing insurance markets with low-to-middle IP levels and increasing IP pattern until 2009 and decreasing thereafter. The within-cluster countries are Croatia*, Slovenia, Iceland, the Czech Republic, Sweden, and Romania. Cluster 5: Developing insurance markets with low-to-middle IP levels and almost quadratic IP

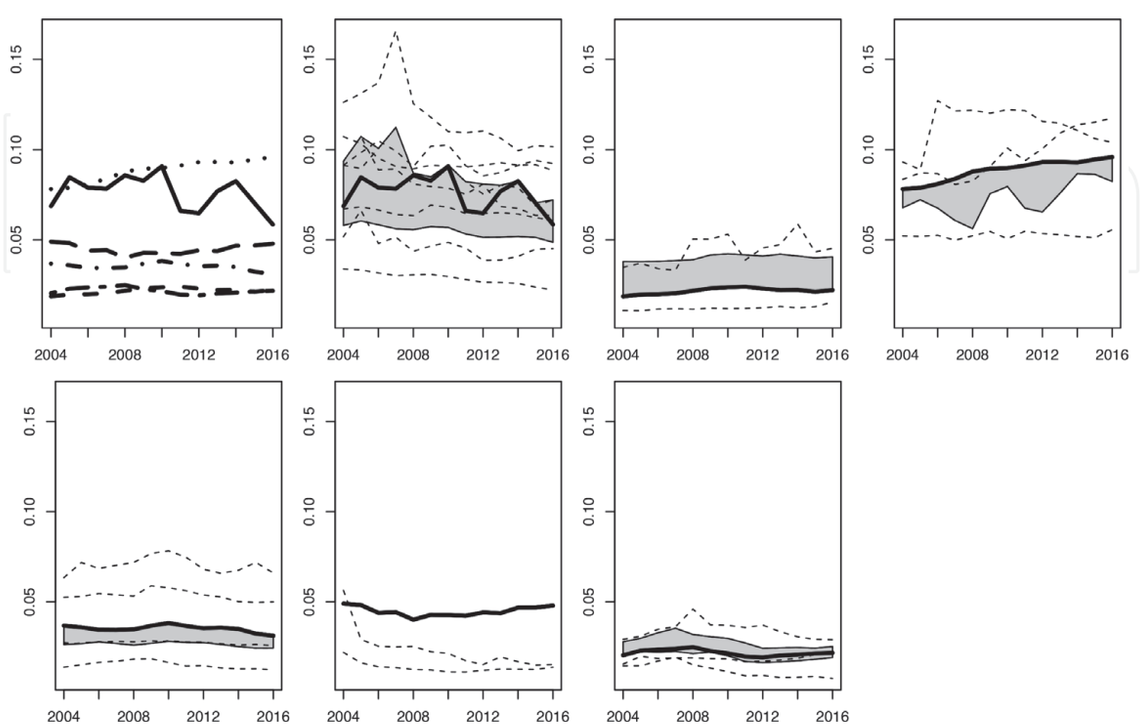


Figure 4.
Clustering results of the IP curves: Magnitude plot.

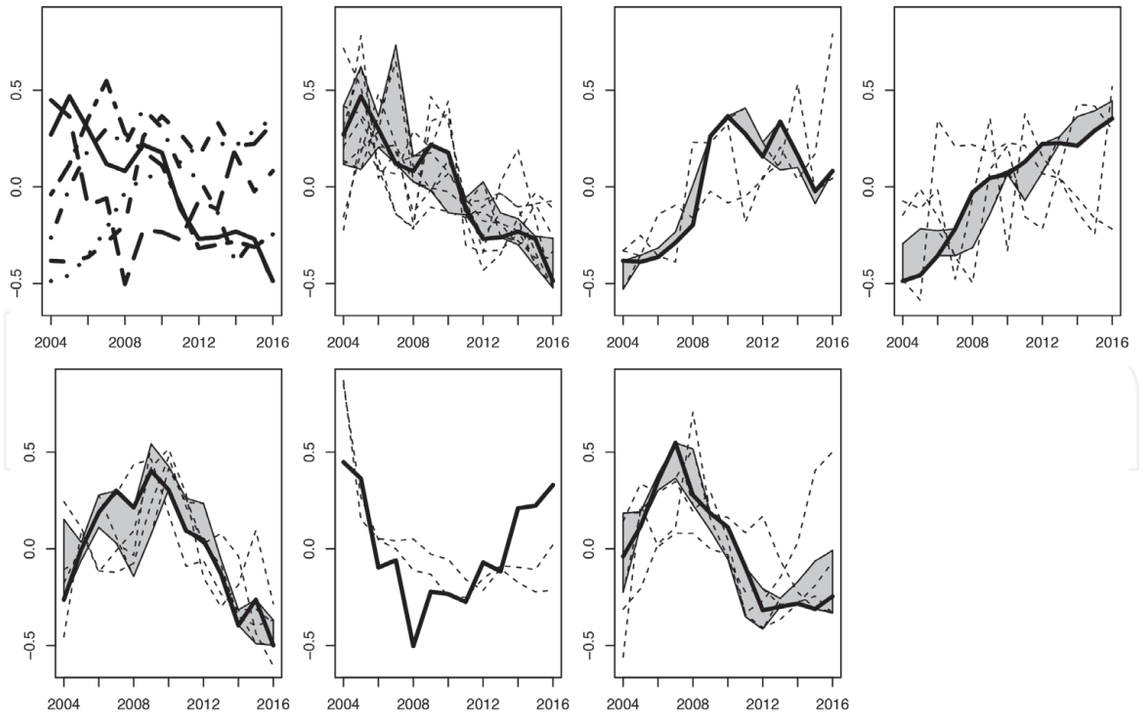


Figure 5.
Clustering results of the IP curves: Shape plot.

pattern and vertex point in 2008. In this cluster, we see countries such as Russia*, Ukraine, and Norway. Cluster 6: Least Developed insurance markets with low IP level and increasing IP pattern followed by a decreasing one initiated in 2007, right on the start of the financial crisis. Members of this cluster are countries such as Lithuania*, Bulgaria, Hungary, Estonia, Serbia, and Poland. The * symbol denotes the medoid IP curve produced by the clustering for each cluster.

The results bring to surface first the difficulty of the European insurance industry to converge and to exhibit homogeneity among national insurance markets during the whole period. A fact that otherwise could lead to the building of single European insurance industry. Second and final, the differential behavior of European insurance markets under different phases of the macroeconomic environment. For instance, Least Developed non-EU insurance markets faced shrink challenges, especially during and after the financial and sovereign debt crises period. The same challenge with a time lag of approximately two years was obvious for some Developing insurance markets. Russia and Ukraine had their insurance markets running in parallel and separated from the other two Developing insurance markets to follow their own smile-shaped development pattern. A slight improvement in insurance activity was also observed for the remaining Developing insurance markets that lasted almost until the end of the sovereign debt crisis in 2011. However, this improvement was offset by their unstable development pattern thereafter. Over the past years, the overall development of Developed insurance markets has decreased, due to a contraction in life insurance business. However, few of them managed to succeed in an increasing pattern with varying IP rate changes over the years.

5. Clustering of population growth data

Over the last century, the world has seen rapid population growth. Particularly, the global population more than quadrupled. The magnitude of the population rate of change from one year to another is found by the fold change ratio (FCR). Fold

change is calculated simply as the ratio of the year-end over the year-start population of a certain country. We refer to the evolution of FCR over the course of time as the population growth rate (PGF) curve. In this example, our objective is to find clusters of world countries in which their PGF curves share similar magnitude and shape properties. We use the output of the FC method based on Area measure for clustering world countries. This output will also give a hint towards the distribution of the world population and provide the trends or the dynamics that are defining our world, such that policymakers can set sustainable development goals for our societies.

Thus, we consider the world population data (United Nations 2016), which was analyzed by [21]. This dataset includes estimates of the total population (both sexes) in 233 countries, areas, or regions in July 1950–2015. Motivated by these estimates and the arguments needed for the execution of the FC method, we follow three steps. In the first step, we perform the preprocessing of the dataset by selecting those countries with populations of more than one million in July 1950. In total, 134 countries are included in our analysis. For each of these countries, we collect 65 data points that correspond to the FCR of each year interval and propose connecting them to make the PGF curve. In the second step, we derive the shape information from the L2 normalization of the shifted PGF curves towards their center. This particular step is the one that provides the set of PGF pattern (PGFP) curves. In the last step, we specify the input argument for the number of clusters which is required by FC method to start. The optimal number of clusters was arrived at by calculation of the median value of all methods presented in NBclust library of the R software. Based on the result of this calculation, the chosen number of clusters was three.

Figure 6 satisfies the marginal plot style approach followed in our case studies by presenting the output of the FC method in a two-panel display. The first panel is dedicated to magnitude clustering (it helps discern broad trends in PGF curves), and the second to the shape clustering (it helps identify patterns of pace for population rate of change). The first plot of each panel is the plot of the median curves of the clusters. Remark that the fraction of combinations of countries satisfying the triangular inequality with Area measure was 1 with respect to all combinations.

Next, we present both the derived clusters and their characterization, which is based on the United Nations (UN) geographical region and classification of economies. For instance, we see that the population growth rates in Cluster 1 appear to follow an increasing trend or at least maintain a certain degree of stability because of a natural increase and migration. Most countries in this cluster have a developing economy and are mainly located in Sub-Saharan Africa. However, three European countries (Ireland, Norway and Spain) with developed economies are also members of this cluster of countries. In contrast, the other two characteristic population growth trends that are present in both Clusters 2 and 3 paint a picture of a stagnating or shrinking population in the future, the only difference being that the population in Cluster 3 has a faster speed of shrinkage than in Cluster 2. The most populated cluster (that is Cluster 2 with 64 curves) is mostly associated with another set of developing economies (such as those of Brazil, China and Singapore) located, this time, in Latin America and the Caribbean along with East Asia and Pacific. Additionally, the only developed economy that appears to reside in this cluster is that of the United States, while few economies in transition that belong to the Commonwealth of Independent States (such as those of Azerbaijan, Kazakhstan) make their presence visible for a first time.

Finally, Cluster 3 has united mostly the developed economies of Europe and East Asia and Pacific along with the economies in transition of South-Eastern Europe (Albania, Serbia and North Macedonia). Moreover, the population of few

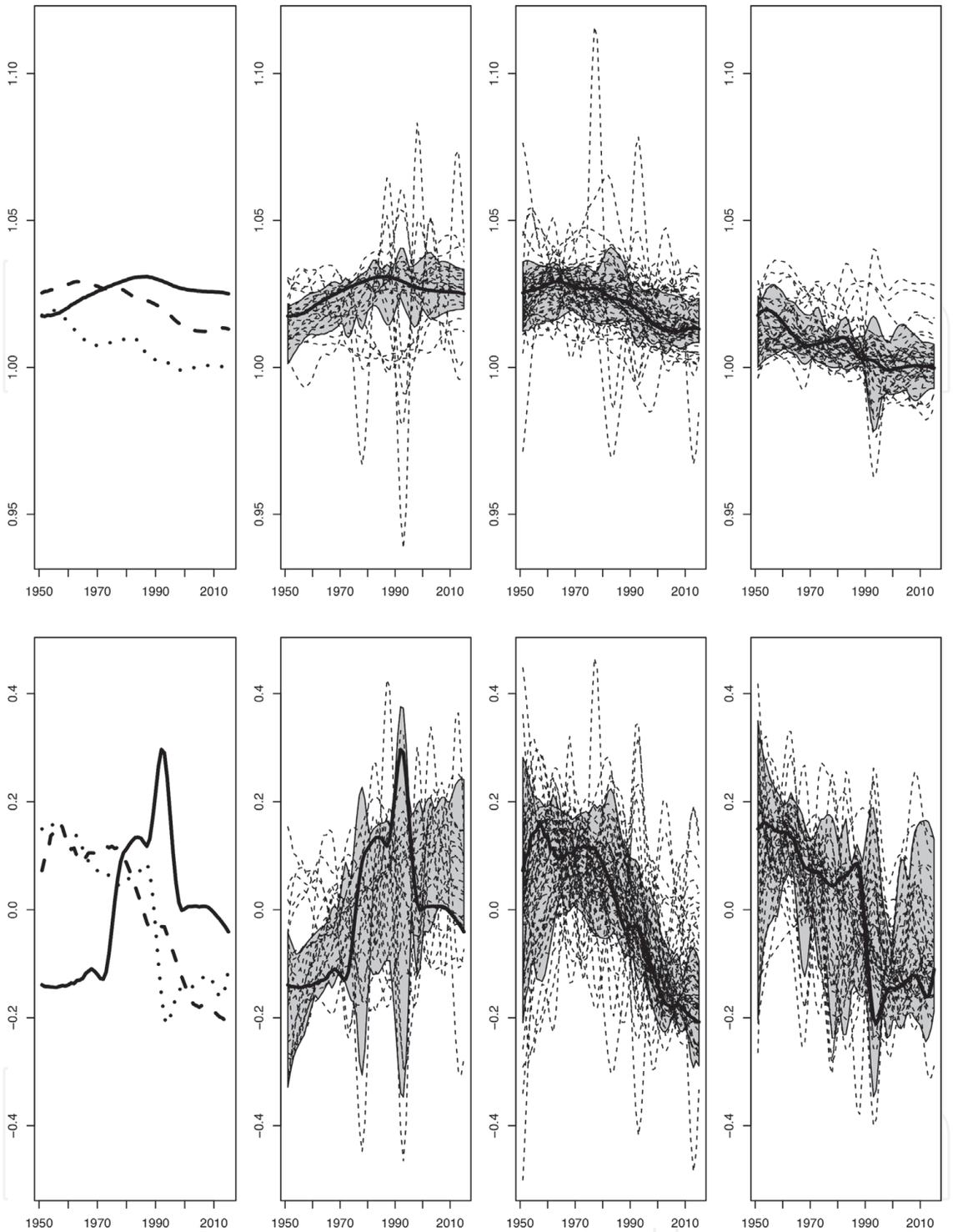


Figure 6.
Clustering results for the population curves. Top panel: Magnitude plot; bottom panel: Shape plot.

developing economies that are located, for example, at Cuba, Jamaica, Puerto Rico, Ghana and Mozambique, have distinguished themselves from the vast majority of developing economies in Cluster 1 or Cluster 2 by following the population behavior of developed economies.

In conclusion, developing economies and economies in transition are split between two clusters, while the majority of the developed economies belong to one cluster. Based on the characterization of these clusters, it is understood that countries with developing economies experience population growth (or at least population stability). However, the more the economy of a country is developed, the more

its population growth change decreases. This decrease, in certain cases, might have even a severe negative effect on a country's future projected population size. Whereas, in some other cases, the effect of this decrease is smoother without forcing the population size to reach record lows.

6. Multivariate clustering of insurance penetration with ratio of life and total insurance

The insurance industry generates a large volume of multivariate functional data from the simultaneously obtained measurements on variables related to life, non-life, and total insurance activities. In our case of interest, two main country-specific variables that include data on premiums are available. The first is the total IP (TIP) that represents the development of total activities. While the second is the R ratio of life IP to TIP that represents the development of the share of life premiums in total premiums.

Since the insurance industry of a country can be represented by the bivariate variables of TIP and R, it is important to take into account the dependence between them. We compute a variable that describes this dependence through the covariance function:

$$\mathrm{Cov}(t) = sign((IP(t) - m_1(t))(R(t) - m_2(t)))\sqrt{|(IP(t) - m_1(t))(R(t) - m_2(t))|},$$

where $m_1(t)$ is the mean IP over all countries and $m_2(t)$ is the mean R over all countries and represents the development of the link between total and life share dynamics.

There is no doubt that the development of total activities is different from that of life share. Nevertheless, it may be assumed that a common development coordinates these differential developments of different insurance variables. Then, it is of great interest to identify groups of insurance markets with similar joint development patterns. With this consideration in mind, we aim to discover whether the European insurance market is homogeneous when national insurance developments are jointly differential by developing their total activities and their life share.

We obtained again insurance data from Swiss Re (2016) database and for the same European (EU and non-EU) countries as in univariate case. In particular, we employ a dataset of our main variables for 34 European countries sampled at annual frequency between 2004 and 2016. That is to say that the data for each variable can be viewed as curves. Yet, except for the curves related to TIP and R variables, we also included the computed curves for the Cov variable in our dataset and ended up with a set of three-dimensional vectors of curves.

Viewing the curves for each variable as a set of curves, a three-component list of curve sets is constructed to serve as an input for the FC method. This time, the optimal number of clusters is three and consistent with the median value of all methods presented in the NBclust library of the R software. Our proposed method concentrates on visualizing, in the marginal plot style approach, clusters of multivariate insurance functional data with regard to their magnitudes and covariance function (**Figures 7–9**).

The clustering results are summarized as follows:

Cluster 1: Countries of high TIP and high R with no correlation whatsoever between the two variables throughout the study period.

Cluster 2: Countries of high TIP and high R with a positive correlation between the two variables throughout the study period.

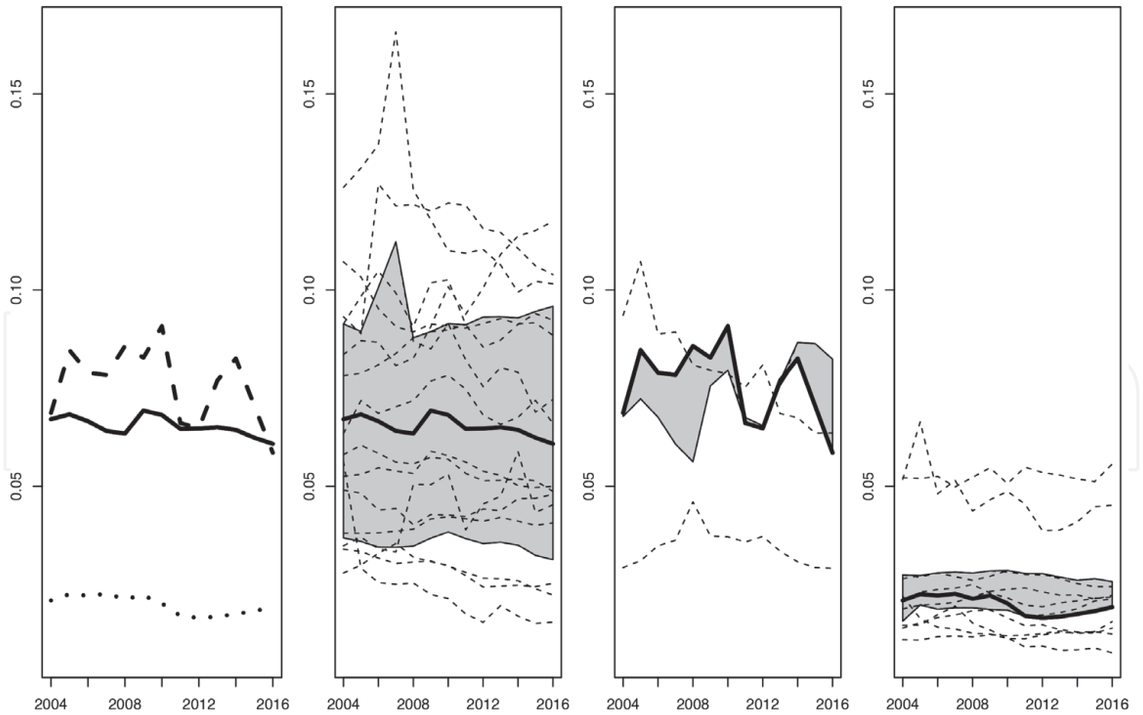


Figure 7.
Clustering results for bivariate curves of TIP and R: TIP plots.

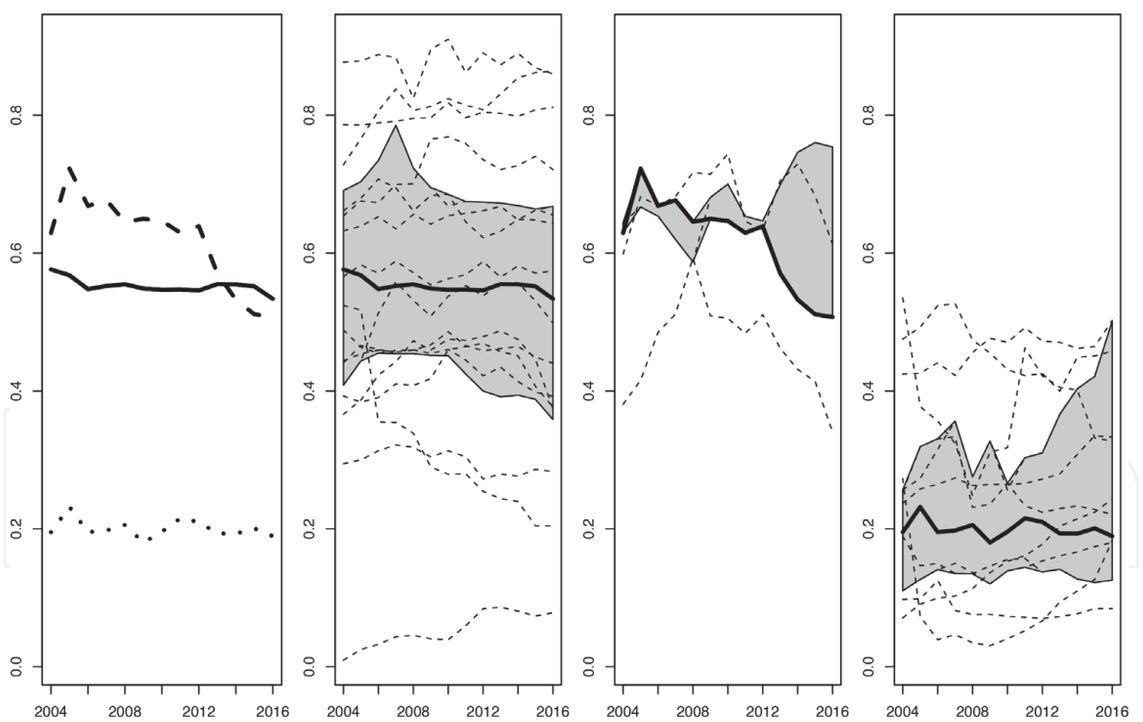


Figure 8.
Clustering results for bivariate curves of TIP and R: R plots.

Cluster 3: Countries of low TIP and low R with no correlation whatsoever between the two variables throughout the study period.

Additionally, the FC method suggests that the total and life share developments in Cluster 1 and Cluster 3 have independent paths since curves for Cov variable almost coincide with the x-axis of **Figure 9**. Simultaneously, it succeeded not to clustered them together due to different magnitude levels. On the contrary, in

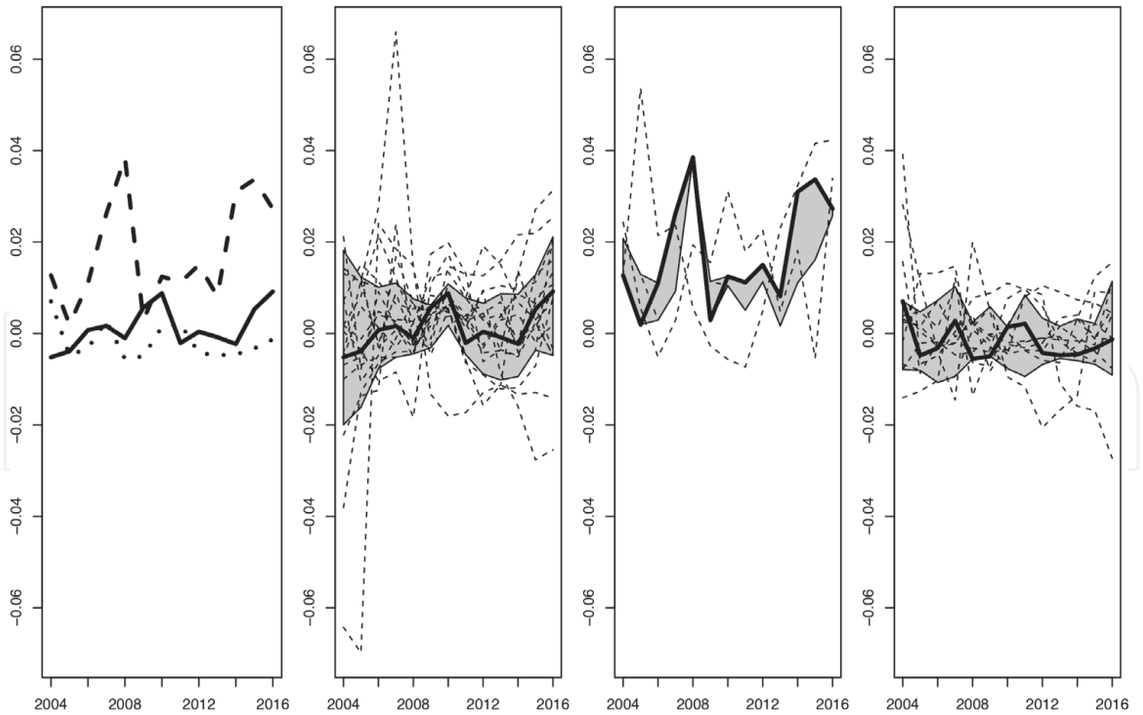


Figure 9.
Clustering results for bivariate curves of TIP and R: Covariance plots.

Cluster 2, the curves for Cov variable are positioned above x-axis, which means that total and life share are dependent functions (positively correlated) over the years.

The functional cluster analysis revealed some differences in the dynamics of insurance markets in Europe. The clustering results clearly reject the hypothesis on the homogeneity of the European insurance market. Europe continues in a two-speed insurance market, with countries with high development and independent paths of total and life insurance business, and others with low. For both speed markets, detecting an increasing pattern in total insurance business does not guarantee that the life premiums will also follow at the same time the same pattern. Any similarity in their patterns could be explained by socio, economic, demographic, or other factors and not by the total business pattern itself. However, there is another high-speed market where the increase of total insurance business in the economy is an additional factor that accelerates the development of life business share.

7. Conclusions

In this study, we introduce a new class of functional cluster analysis methods based on functional orderings. We intended to work with methods that allow intrinsic graphical interpretation to obtain a natural interpretation of clusters via their central regions. Therefore, we propose the use of a studentized measure that forms a metric on the set of functional differences. Also, We suggest the use of the area measure, which orders the functions according to the area of the most extreme continuous rank and considers the entire distribution of the functions. This measure does not form a metric on any set of functions, but the simulation study results and the real data study suggest that it is a metric on any real data set of functions. The check for the satisfaction of the triangular inequality is provided for the given set of functions.

This study's primary aim is to introduce methods that combine the various functional information sources equally. It is possible to study clustering while showing equal concern for both magnitude and shape, as shown in the first and

second data examples. In other words, it is possible to study the clustering of multivariate functions when the marginal functions are taken equally. It is also possible to add to the study term, which summarizes the covariance between the marginals of the multivariate function, as shown in the third data example.

The simulation study suggests that the proposed method is robust and more powerful than studied alternatives that give equal treatment to various sources. The studied alternatives are the K-means method, with pre-standardization of every coordinate by its mean and variance, and the model-based method, which assumes a normal distribution of data and considers marginals means, variances, and the covariance function.

Our proposed methods consider the covariance structure of the functional data via the ordering of the entire functional differences. Our proposed methods are also nonparametric and, as such, have no model requirement. Our simulation study also showed that our proposed methods are quite robust to heavy-tailed functions, which can be considered as a type of functional cluster outlier. The data studies show that our methods can cluster the functions with respect to magnitude and shape and that it provides a sensible graphical interpretation of the resulting clusters. The third example shows that the clusters can be also constructed with respect to the covariance of the marginals in the multivariate function. This study does not examine methods to choose the number of clusters in an optimal manner, and this problem is left to the user's choice or further development.

Acknowledgements

Wenlin Dai has been financially supported by National Natural Science Foundation of China (Project No. 11901573). T. Mrkvička has been financially supported by the Grant Agency of Czech Republic (Project No. 19-04412S).

Author details

Wenlin Dai[1], Stavros Athanasiadis[2] and Tomáš Mrkvička[2*]

1 Center for Applied Statistics, Institute of Statistics and Big Data, Renmin University of China, Beijing, China

2 University of South Bohemia, České Budějovice, Czech Republic

*Address all correspondence to: mrkvicka.toma@gmail.com

IntechOpen

© 2021 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

References

[1] Ieva, F., A. M. Paganoni, D. Pigoli, and V. Vitelli (2013). Multivariate functional clustering for the morphological analysis of electrocardiograph curves. Journal of the Royal Statistical Society: Series C 62 (3), 401–418.

[2] Kazor, K. and A. S. Hering (2015). Assessing the performance of model-based clustering methods in multivariate time series with application to identifying regional wind regimes. Journal of Agricultural, Biological, and Environmental Statistics 20(2), 192–217.

[3] Tupper, L. L., D. S. Matteson, C. L. Anderson, and L. Zephyr (2018). Band depth clustering for nonstationary time series and wind speed behavior. Technometrics 60(2), 245–254.

[4] Ferraty, F. and P. Vieu (2006). Nonparametric Functional Data Analysis: Theory and Practice (1 ed.). Springer Series in Statistics. Springer.

[5] Chiou, J.-M. and P.-L. Li (2007). Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society: Series B 69(4), 679–699.

[6] Jacques, J. and C. Preda (2014). Model-based clustering for multivariate functional data. Computational Statistics and Data Analysis 71, 92–106.

[7] Zeng, P., J. Q. Shi, and W.-S. Kim (2019). Simultaneous registration and clustering for multidimensional functional data. Journal of Computational and Graphical Statistics 28(4), 943–953.

[8] López-Pintado, S. and J. Romo (2009). On the concept of depth for functional data. Journal of the American Statistical Association 104(486), 718–734.

[9] López-Pintado, S., Y. Sun, J. K. Lin, and M. G. Genton (2014). Simplicial band depth for multivariate functional data. Advances in Data Analysis and Classification 8(3), 321–338.

[10] Myllymäki, M. and T. Mrkvička (2020). Get: Global envelopes in r. arXiv preprint arXiv:1911.06583.

[11] Sguera, C., P. Galeano, and R. Lillo (2014). Spatial depth-based classification for functional data. TEST 23(4), 725–750.

[12] de Micheaux, P. L., P. Mozharovskyi, and M. Vimond (2020). Depth for curve data and applications. Journal of the American Statistical Association 0(0), 1–17.

[13] Dai, W., T. Mrkvička, Y. Sun, and M. G. Genton (2020). Functional outlier detection and taxonomy by sequential transformations. Computational Statistics and Data Analysis 149, 106960.

[14] Myllymäki, M., P. Grabarnik, H. Seijo, and D. Stoyan (2015). Deviation test construction and power comparison for marked spatial point patterns. Spatial Statistics 11, 19–34.

[15] Myllymäki, M., T. Mrkvička, P. Grabarnik, H. Seijo, and U. Hahn (2017). Global envelope tests for spatial processes. J. R. Statist. Soc. B 79(2), 381–404.

[16] Narisetty, N. N. and V. N. Nair (2016). Extremal depth for functional data and applications. Journal of American Statistical Association 111 (516), 1705–1714.

[17] Febrero-Bande, M. and M. Oviedo de la Fuente (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software 51(4), 1–28.

[18] Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant

analysis, and density estimation. Journal of the American Statistical Association 97(458), 611–631.

[19] Carroll, C., A. Gajardo, Y. Chen, X. Dai, J. Fan, P. Z. Hadjipantelis, K. Han, H. Ji, H.-G. Mueller, and J.-L. Wang (2021). fdapace: Functional Data Analysis and Empirical Dynamics. R package version 0.5.6.

[20] Hubert, L. and P. Arabie (1985). Comparing partitions. Journal of classification 2(1), 193–218.

[21] Nagy, S., I. Gijbels, and D. Hlubinka (2017). Depth-based recognition of shape outlying functions. Journal of Computational and Graphical Statistics 26(4), 883–893.

IntechOpen