

Functional Data Clustering Via Functional Mahalanobis Distance

Yangxinzi Zao *

College of Science, Minzu University of China, Beijing, China, 10081

* Corresponding Author Email: zyxz13698730397@163.com

Abstract. As an exploratory data analysis method, functional data clustering aims to identify the underlying features of the observed data. In this context, this paper proposes a functional data clustering method based on functional Mahalanobis distance. As a distance-based non-parametric clustering model, the proposed method can effectively avoid the disadvantages of generative models and has excellent properties of decoupling and dimension standardization. Compared with other functional data clustering models, this method has lower computational complexity. In addition, the proposed method can be applied to any distance-based multivariate clustering method, thus generalizing it to the case of functional data. In practical data analysis, this paper compares the performance of this method with some other functional clustering methods, using k-means clustering as an example, and finds that it has better performance in terms of purity, adjusted Rand index, and computational speed. Finally, the idea of using Mahalanobis distance for functional data distance measurement can also be extended to construct kernel functions for measuring similarity between functional data samples, thus developing non-linear functional data analysis methods based on reproducing kernel theory.

Keywords: Functional Data, Cluster Analysis, Functional Mahalanobis Distance.

1. Introduction

Against the background of the big data era, with the increasingly advanced technologies of data collection and storage, the presentation of data is no longer limited to discrete and sparse point sets under traditional technical conditions, but rather characterized by continuous and dynamic functional curves (or surfaces) supported by a large amount of data. Functional data analysis was first proposed by Ramsay [1-2]. Among several branches, functional data clustering analysis is an exploratory data analysis method that classifies function objects according to their similarities and has a wide range of applications in object classification recognition, medical diagnosis, and numerical control machine monitoring. This paper focuses on the research of functional data clustering analysis, and the following are some application examples of functional data clustering analysis. The first example is to identify the category of residents' household computers based on collected data on residents' electricity usage at home (Figure 1. a). The second example is to identify potential differences and distinguish whether the measured person's heart has myocardial infarction by tracking the electrocardiogram data during one heartbeat [3] (Figure 1. b). The third example is to match the contours of different fish species based on their contour parameters for species identification [4] (Figure 1. c). The fourth example is to identify different meats based on Fourier transform infrared spectroscopy combined with attenuated total reflection sampling [5] (Figure 1. d).

As a statistical analysis method, Cluster analysis has been widely used and has achieved numerous significant research results. Function-based data clustering methods are roughly classified into four categories: direct clustering of raw data, two-step concatenation clustering based on continuous functions, adaptive model clustering, and distance-based clustering [6]. Specifically, Gareth M et al. [7] developed a pioneering EM-type algorithm for parameter estimation. Jank [8] proposed the ascent EM and genetic ascent EM algorithms based on the Monte Carlo EM algorithm. Lim et al. [9] applied wavelet basis functions to cluster multiscale function-based data. García-Escudero and Gordaliza [10] used a variant of the K-means algorithm, namely the d-trimmed K-means algorithm, based on the third-order B-spline basis for data clustering analysis. Ray and Mallick [11] developed a non-

parametric Bayesian analysis framework. Chen [12] used basis function expansion for function-based data clustering and solved the problem of the Pearson similarity coefficient being unable to describe differences in curve shape due to Euclidean distance. Du et al. [13] proposed an innovative functional data clustering method based on directional multiple hypothesis test and information entropy.

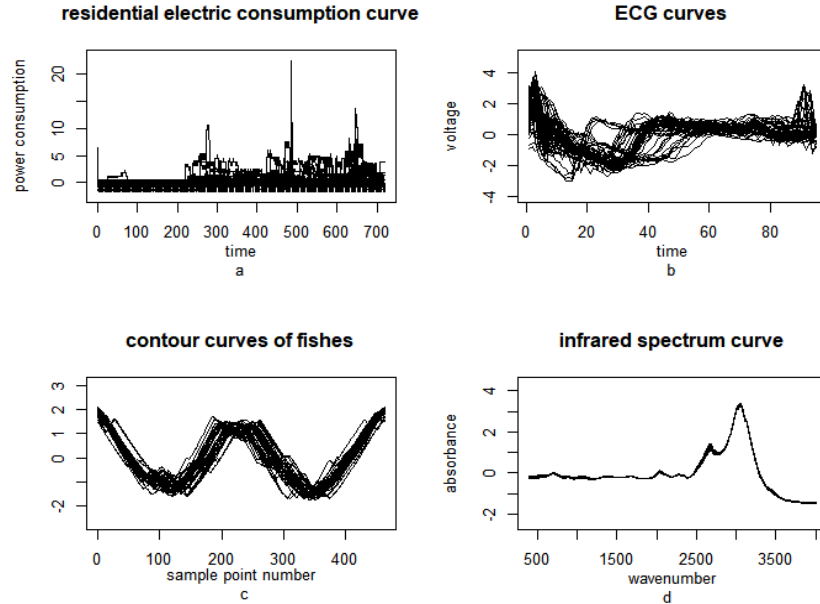


Figure 1. Curves of different application scenarios

Although these methods have good clustering effects, their application scenarios are somewhat limited. For example, function-based clustering methods based on point data do not take into account potential function feature information and can suffer from the problem of dimensionality disaster. Clustering methods based on generative models cannot estimate probability distributions well, and clustering methods based on two-stage models do not have a unified paradigm for selecting the type and number of basic functions. Based on this, this paper will focus on distance-based function-based data clustering methods. Specifically, among several commonly used distances, the Minkowski distance cannot exclude the unit of each variable and the correlation between variables. The Lance distance is only applicable when all variables are greater than 0 and does not consider the correlation between variables. The Mahalanobis distance has received broad attention due to its decoupling and dimensional standardization, for example, Wu et al. [14] improved the covariance matrix in the Mahalanobis distance clustering analysis, resulting in a significant improvement in the clustering accuracy of the new method, Berrendero et al. [15] provided a comprehensive review of various functional Markov distances and proposed a novel Markov distance based on the reproducing kernel Hilbert space (RKHS).

In summary, this paper proposes a function-based data clustering method based on the Mahalanobis distance. The proposed method has several innovative points: firstly, as a non-parametric statistical model based on distance, it can effectively avoid the estimation bias generated by the insufficient sample size in density distribution fitting in generative models. Secondly, compared with other distance-based clustering models, the calculation complexity and cost of the proposed method are relatively low. Finally, this method can be directly applied to distance-based multivariate clustering methods, thereby extending it to function-based data scenarios. In practical data analysis, this paper uses K-means clustering as a simple example and compares it with function-based data clustering methods based on Gaussian mixture models, subspace projection, and other types of distances in the above four practical applications. The results show that the proposed method has certain advantages in purity, adjusted Rand index, and calculation speed.

2. Theory and Methods

2.1. Functional Mahalanobis distance

Galeano et al. [16] extended the Mahalanobis distance [17] to enable its application in functional data analysis. The specific procedure involves transforming discrete data into a functional form through the linear combination of basic functions. In other words, discrete data is smoothed using basis functions such as B-spline bases, wavelet bases, Fourier bases, and so on. The result is an expansion of the basic functions, given by: $X_n(t) \approx \sum_{m=1}^M c_{nm} B_m(t)$ where $1 \leq n \leq N$ and where M is the number of basic functions chosen for the expansion, $B_m(t)$ represents the selected basis functions, and c_{nm} denotes the coefficients corresponding to each basis function. By performing functional principal component analysis, the basic functions are transformed into orthonormal basis functions, and the covariance function $V(s, t) = \sum_{k=1}^{\infty} \lambda_k v_k(s) v_k(t)$ is obtained. The expansion of random functions $X(t)$ using principal component basis functions is expressed as follows:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \theta_k v_k(t) \quad (1)$$

where $\theta_k = \int \{X_i(t) - \mu(t)\} v_k(t) dt$, $v_k(t)$ is a principal component function in L^2 , θ_k is a random function. θ_k is the k th principal component score of $X(t)$ with θ_k having a mean of 0 and a variance of λ_k . The specific derivation process of the above principal component score is based on the method of Ramsay and Silverman [18]. Next, the Mahalanobis distance is extended to the case of functional data. It adds the inverse of the covariance matrix based on the Euclidean distance so that the data can be rotated to the principal component direction to achieve the purpose of eliminating the correlation of each dimension. Specifically, after obtaining the principal component score matrix $\Theta = \int X(t) v(t) dt$ of $N \times K$, the functional Mahalanobis distance is defined as

$$M(X_i(t), X_j(t)) = \sqrt{\frac{\sum_{k=1}^K \left\{ \int [X_i(t) - X_j(t)] \hat{v}_k(t) dt \right\}^2}{\hat{\lambda}_k}} = \sqrt{\frac{\sum_{k=1}^K (\hat{\theta}_{ik} - \hat{\theta}_{jk})^2}{\hat{\lambda}_k}} \quad (2)$$

where $\hat{v}_k(t)$ and $\hat{\lambda}_k$ respectively represent the estimated values of the principal component score function and the variance of the principal component scores.

2.2. Functional clustering method based on functional Mahalanobis distance

Distance-based clustering methods can be extended to incorporate the functional Mahalanobis distance, enabling their application in clustering functional data. Here, we provide a detailed explanation of how the Mahalanobis distance can be used to generalize K-means clustering for functional data, as well as briefly discuss extending other distance-based clustering methods for functional data. After obtaining the calculation method for the Mahalanobis distance of functional data, it can be incorporated into the K-means clustering method by replacing the Euclidean distance calculation. This allows for the application of K-means clustering to functional data. The specific calculation process is as follows:

Assuming we have a dataset $D = \{X_1, X_2, \dots, X_N\}$, where $X_i = \{x_{i1}, x_{i2}, \dots, x_{im}\}$. The dataset is randomly divided into K clusters, denoted as $D_j = \{X_{j1}, X_{j2}, \dots, X_{jr}\}$, and we obtain the centroid function $V = \{V_1, V_2, \dots, V_K\}$ for each cluster.

Compute the functional Mahalanobis distance, denoted as

$$M(X_i(t), V_j(t)) = \sqrt{\frac{\sum_{k=1}^K \left\{ \int [X_i(t) - V_j(t)] \hat{v}_k(t) dt \right\}^2}{\hat{\lambda}_k}} \quad i=1, 2, \dots, N, j=1, 2, \dots, K$$

between each cluster of X_i and its corresponding centroid function V_j . This process results in a distance matrix $Mdis$ with N rows and K columns.

Find the cluster corresponding to the minimum distance between X_i and the class center function V_j , and classify X_i to the cluster corresponding to the minimum distance, and then obtain a new cluster data set $D_l = \{X_{l1}, X_{l2}, \dots, X_{lr}\}$ and a new center $V = \{V_1, V_2, \dots, V_K\}$.

Repeat steps 1 to 3 until the functional Mahalanobis distance between the class center function of the previous round and the class center function of the next round approaches 0, and the final clustering result $D_s = \{X_{s1}, X_{s2}, \dots, X_{sr}\}, s=1, 2, \dots, K$ is obtained.

Determining the value of K can be achieved using methods such as the Silhouette Coefficient method, the Elbow method, or criteria like AIC and BIC. Similarly, other distance-based clustering methods can be improved by incorporating the Mahalanobis distance to extend their applicability to functional data clustering. For instance, in the Agglomerative Hierarchical Clustering Algorithm, replacing the Euclidean distance between data clusters with the Mahalanobis distance allows for stepwise clustering to obtain clustering results based on functional data. This improvement method helps mitigate the impact of noise and outliers to a certain extent. Furthermore, in the Mean-Shift Clustering Algorithm, when calculating the distance between data weights and the current centroid function, utilizing the Mahalanobis distance instead of the Euclidean distance can yield more objective weight coefficients. The functional Mahalanobis distance can also be applied to distance-based classical clustering algorithms such as DBSCAN and OPTICS.

3. Data analysis

3.1. Data sources and experimental settings

This article conducts empirical analysis on five real datasets using Mahalanobis distance-based functional data clustering algorithms. The specific datasets are Computers, ECG200, Fish, and Meat, sourced from the timeseriesclassification [19] website.

The computers dataset consists of electricity usage data collected from 250 households as part of the "Powering the Nation" research project. The data includes electricity readings for each household sampled every two minutes over a monthly period, resulting in 720 data points per household. The dataset is divided into two classes: households using desktop computers and households using laptop computers. By exploring the electricity usage data, the potential characteristics of household electricity consumption can be investigated, allowing for clustering analysis of sampled households based on computer usage patterns. The ECG200 dataset is experimental data on time series research conducted by Olszewski [3]. It records the heart activity data of 100 participants, with each sequence representing the electrocardiogram (ECG) activity during a single heartbeat. Each sequence has a length of 96 data points. The participants include individuals with normal heart conditions and patients with myocardial infarction. By analyzing the ECG signals and extracting their latent features, clustering can be performed to explore the ECG characteristics of individuals with different heart conditions. The Fish dataset is from Lee's article on the application of contour matching in fish species recognition [4]. It contains contour data of 175 fish, with each data group having a length of 463. The dataset covers seven different fish species. By clustering the features of different fish contour data, the recognition results of different fish species can be obtained. The Meat dataset is used in Kemsley et al.'s study on meat classification using infrared spectroscopy [5]. The dataset is obtained through

Fourier transform infrared spectroscopy combined with attenuated total reflection sampling. Chicken, pork, and turkey meats exhibit characteristic differences in infrared spectroscopy. The dataset consists of 60 infrared spectroscopy data, with each data point having a length of 448.

The empirical analysis in this article is conducted using R 4.2.1. In each experiment, the distclust, iterSubspace [20], funclust [21], funHDDC [22], fscm [23], and waveclust [24] algorithms from the Fuzzy library (version 1.0.0) [25] are compared. The functional data clustering algorithms based on the Mahalanobis distance are designed by improving the K-means clustering algorithm. For the parameter settings of transforming the data into a functional form, the article uses B-spline bases for smoothing the data, the smoothing degree control parameter is set to 10^{-5} , and the penalty term of the smoothing function is set to the second derivative. When performing principal component analysis on the functional data, the number of principal components is determined based on the cumulative contribution rate, considering a threshold of 80% as reasonable. In the data analysis and comparison, the purity, adjusted Rand index (ARI), and computational speed are evaluated to assess the algorithm's performance compared to the other six algorithms. Higher purity and adjusted Rand index indicate better clustering results.

3.2. Comparative result

The clustering results of the Mahalanobis distance-based functional data clustering method using the computers dataset are depicted in Figure 2, where the red and black curves represent two categories of computers. The figure displays a subset of seventy curves for visualization purposes. From the visual representation, it can be observed that the clustering results based on the Mahalanobis distance capture certain patterns compared to the actual data. To provide a more intuitive evaluation of the method's performance, this study compares it with other clustering methods in terms of purity, adjusted Rand index, and computational speed, as shown in Table 1. In this context, clustering methods with computation times exceeding 900 seconds are considered impractical and are not further discussed.

As evident from Table 1, the Mahalanobis distance-based clustering method exhibits significant advantages in terms of computational speed. Moreover, it demonstrates higher purity and adjusted Rand index values compared to other clustering methods, indicating superior clustering performance. The clustering results of different methods are visualized and compared with the actual data categories in Figure 3. The visual representation in Figure 3 illustrates that the Mahalanobis distance-based clustering method achieves satisfactory results by separating the data into two categories more accurately.

Table.1. Evaluation index of the clustering effect of Computer dataset

Clustering method	Computing time/s	Purity	Adjusted Rand index
ma-distance	0.0775	0.6160	0.0505
distclust	>>900	——	——
iterSubspace	0.3146	0.5240	-0.0015
funclust	23.4189	0.5560	0.0089
funHDDC	0.4294	0.5880	0.0271
fscm	19.0315	0.5520	0.0069
waveclust	>>900	——	——

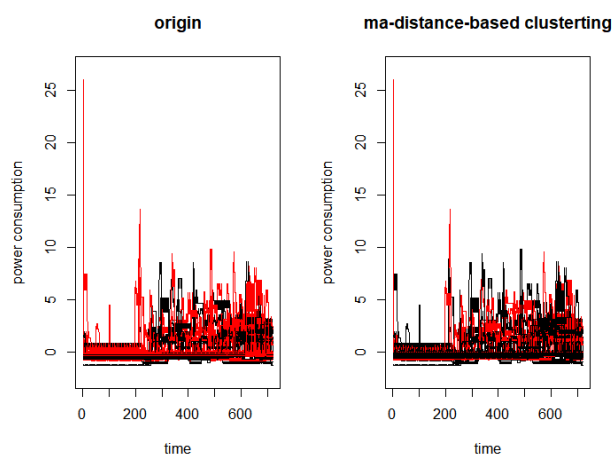


Figure 2. Computer dataset clustering results

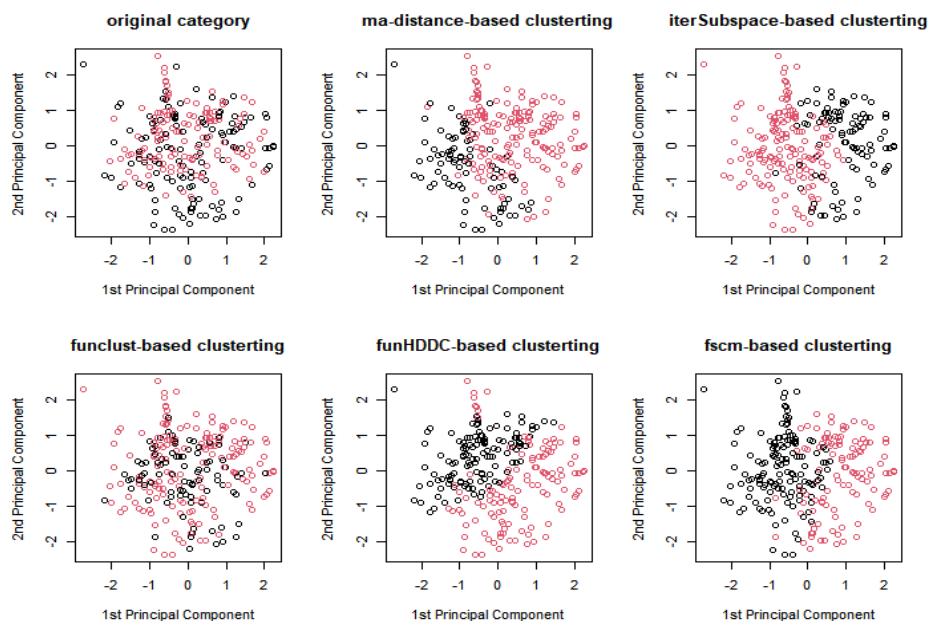


Figure 3. Comparison of the results of different clustering methods

The clustering results of function-based data using Mahalanobis distance on the ECG 200 dataset are shown in Figure 4. The red and black curves represent different electrocardiogram (ECG) features. From the figure, it can be observed that the clustering method based on Mahalanobis distance can effectively identify and cluster problematic curves in the ECG signals, yielding good results. The comparison between the clustering method based on Mahalanobis distance and other clustering methods is presented in Table 2.

Table.2. Evaluation index of the clustering effect of ECG 200 dataset

Clustering method	Computing time/s	Purity	Adjusted Rand index
ma-distance	0.0480	0.8000	0.3514
distclust	18.8218	0.6400	0.0424
iterSubspace	0.0653	0.6400	0.0424
funclust	1.8715	0.4500	-0.0040
funHDDC	0.3659	0.6400	0.0000
fscm	1.4592	0.6700	0.1031
waveclust	16.0075	0.7400	0.1940

By referring to Table 2 and Figure 5, it is evident that the Mahalanobis distance-based clustering method outperforms other clustering methods in terms of computational speed, purity, and adjusted Rand index. The visual representation of the final clustering results, as shown in Figure 5, demonstrates that the Mahalanobis distance-based method effectively distinguishes the majority of abnormal and normal electrocardiogram (ECG) features. Therefore, in the context of this dataset, the Mahalanobis distance-based clustering method yields the most favorable results.

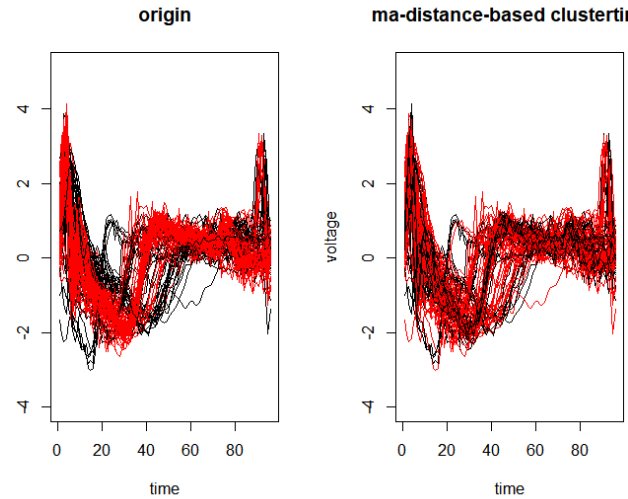


Figure 4. ECG 200 dataset clustering results

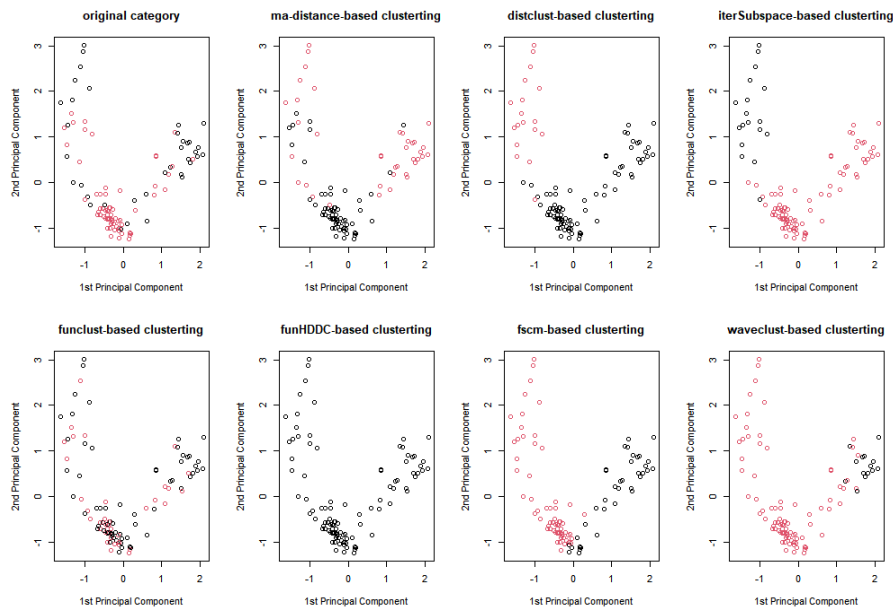


Figure 5. Comparison of the results of different clustering methods

The clustering results of function-based data using Mahalanobis distance on the Fish dataset are illustrated in Figure 6. The curves of different colors represent various species of fish, and the dataset contains recorded characteristic data for seven fish species. From Figure 6, it can be observed that the clustering method based on Mahalanobis distance effectively groups the contour curves of the same fish species into one category. A comparison of the clustering results between this method and six other approaches is presented in Table 3. By examining Table 3 and Figure 7, it is evident that the clustering method proposed in this study outperforms other clustering methods in terms of computational speed and clustering effectiveness.

Table.3. Evaluation index of the clustering effect of Fish dataset

Clustering method	Computing time/s	Purity	Adjusted Rand index
ma-distance	0.0657	0.5086	0.3126
distclust	>>900	——	——
iterSubspace	0.2836	0.3486	0.0943
funclust	1.0395	0.2000	-0.0023
funHDDC	1.755	0.2514	0.0284
fscm	25.4156	0.3371	0.0815
waveclust	220.4327	0.4343	0.1838

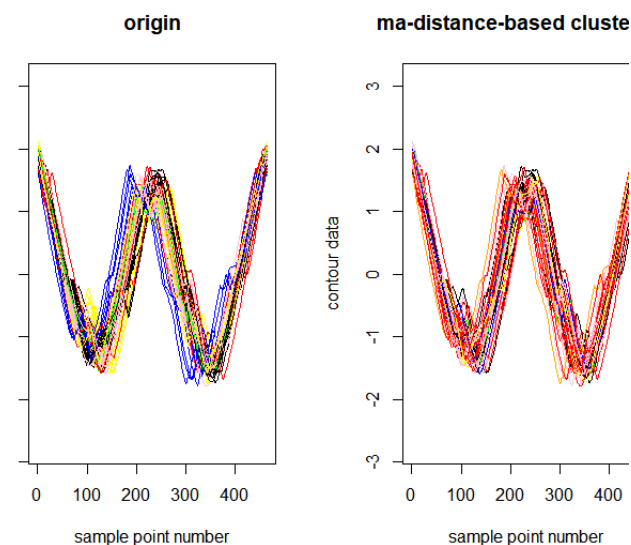


Figure 6. Fish dataset clustering results

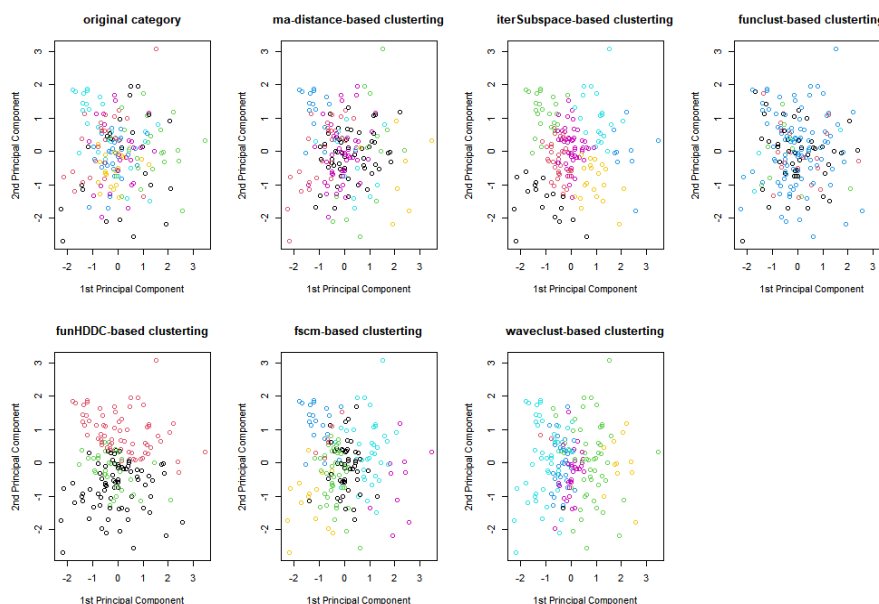


Figure 7. Comparison of the results of different clustering methods

The clustering results of function-based data using the Meat dataset based on the Mahalanobis distance are shown in Figure 8. The dataset consists of infrared spectroscopy data from three types of meat, with different colors representing different meat types. A comparison was made between the clustering results obtained using the Mahalanobis distance and six other clustering methods, as presented in Table 4. From the above results, it can be observed that the clustering method based on the Mahalanobis distance outperforms the five alternative methods in terms of overall effectiveness. It is capable of effectively distinguishing the infrared spectroscopy features of different meat types.

Table.4. Evaluation index of the clustering effect of Meat dataset

Clustering method	Computing time/s	Purity	Adjusted Rand index
ma-distance	0.0445	0.7667	0.5325
distclust	>>900	——	——
iterSubspace	0.1516	0.5167	0.4145
funclust	0.0595	0.3667	0.0002
funHDDC	0.5027	0.6667	0.2673
fscm	3.8727	0.5500	0.3600
waveclust	4.5766	0.9667	0.9018

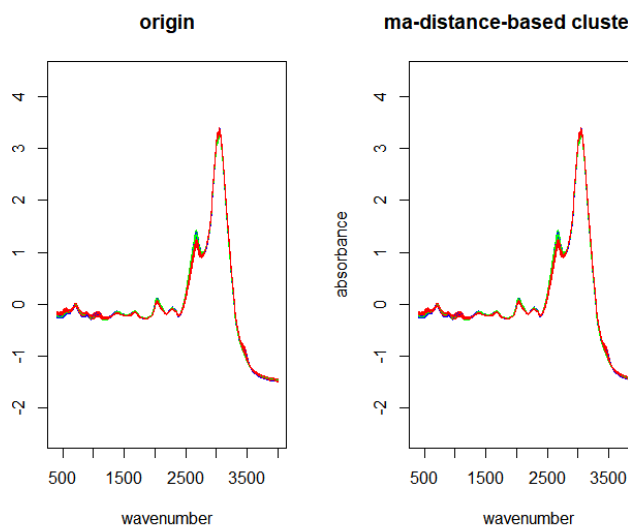


Figure 8. Meat dataset clustering results

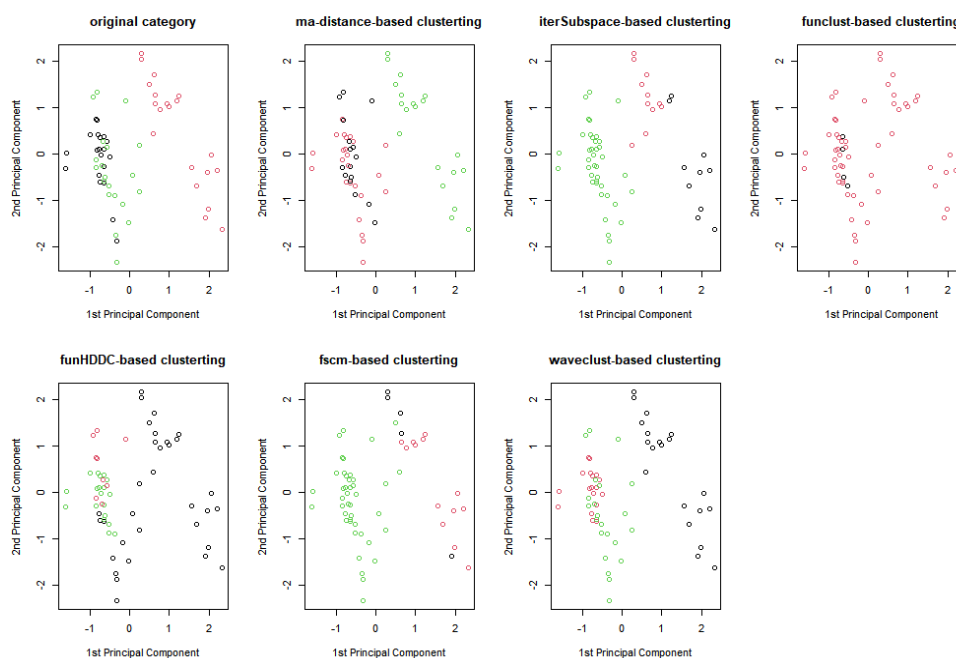


Figure 9. Comparison of the results of different clustering methods

In conclusion, in the practical data analysis phase, the clustering method based on the Mahalanobis distance proposed in this paper has proven to be the optimal approach in terms of computational speed, purity, and adjusted Rand index. Therefore, the algorithm that improves the function-based clustering method using the Mahalanobis distance can be considered successful, as it demonstrates excellent clustering performance and holds significant practical relevance.

4. Conclusion and future directions

This paper proposes a clustering method for function-based data based on the Mahalanobis distance. The main idea is to improve distance-based clustering methods by using the function-based Mahalanobis distance, enabling its application to function-based data clustering. For example, in the K-means clustering algorithm, the Euclidean distance that measures the distance from each point to the cluster centroid is replaced with the function-based Mahalanobis distance, allowing it to be applied to the K-means clustering of function-based data. The Mahalanobis distance has superior decoupling and dimensionality normalization properties compared to other types of distances. Therefore, while improving distance-based methods for function-based data, the excellent properties of the Mahalanobis distance also contribute to its good clustering performance. The advantage of the proposed method lies in its lower computational complexity compared to other non-parametric clustering methods, effectively avoiding the shortcomings of generative models. Additionally, the idea of using the Mahalanobis distance for clustering function-based data can be generalized to various clustering models, making it widely applicable.

For future research, the proposed idea can be extended to construct kernel functions specifically designed for function-based data using the function-based Mahalanobis distance. This would allow for further application of this idea in machine learning methods based on the reproducing kernel theory, such as kernel logistic regression, kernel support vector machines, kernel linear discriminant analysis, as well as kernel-based principal component analysis and other nonlinear dimensionality reduction methods. Therefore, the idea of optimizing and improving algorithms based on the function based Mahalanobis distance proposed in this paper still holds research and generalization value.

Acknowledgments

The authors gratefully acknowledge the data support from the timeseriesclassificationwebsite. Also thank Mr. Li for his guidance.

References

- [1] Ramsay J. When the data are functions [J]. *Psychometrika*, 1982, 47(4): 379-396.
- [2] Ramsay J, Dalzell C. Some tools for functional data analysis [J]. *Journal of the Royal Statistical Society: Series B*, 1991, 53(3):539-561.
- [3] Olszewski R T. Generalized feature extraction for structural pattern recognition in time-series data [M]. Carnegie Mellon University, 2001.
- [4] Lee D J, Archibald J K, Schoenberger R B, et al. Contour matching for fish species recognition and migration monitoring [J]. *Applications of Computational Intelligence in Biology: Current Trends and Open Problems*, 2008: 183-207.
- [5] Al-Jowder O, Kemsley E K, Wilson R H. Mid-infrared spectroscopy and authenticity problems in selected meats: a feasibility study [J]. *Food Chemistry*, 1997, 59(2): 195-201.
- [6] Sun Lirong, Zhuo Weijie, Wang Kaili, and so on. Research on functional clustering analysis method [J]. *Journal of Applied Mathematics of Universities A*, 2020,35 (02): 127-140.
- [7] James G M, Sugar C A. Clustering for sparsely sampled functional data [J]. *Journal of the American Statistical Association*, 2003, 98(462): 397-408.
- [8] Jank W. Ascent EM for a fast and global solution to finite mixtures: An application to curve-clustering of online auctions [J]. *Computational Statistics & Data Analysis*, 2005, 51: 747-761.
- [9] Lim Y, Oh H, Cheung Y K. Multiscale clustering for functional data [J]. *Journal of Classification*, 2019, 36: 368-391.
- [10] Garcia-Escudero L A, Gordaliza A. A proposal for robust curve clustering [J]. *Journal of classification*, 2005, 22(2): 185-201.
- [11] Ray S, Mallick B. Functional clustering by Bayesian wavelet methods [J]. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2006, 68(2): 305-332.

- [12] Chen Xiaofeng, Yin Ruifei. Functional data clustering method based on basis function expansion [J]. Statistics and Decision Making, 2009 (19): 10-12.
- [13] Du Xiuli, Jiang Xiaohu, Sun Chentong, et al. A New Functional Clustering Method Based on Directional Multiple Hypothesis Test and Information Entropy [J]. Journal of Nanjing Normal University (Natural Science Edition). 2022; 45(4): 1-9.
- [14] Wu Xianghua, Niu Shengjie, Wu Chengou, etc. Improvement of covariance matrix estimation in Mahalanobis distance clustering analysis [J]. Mathematical statistics and management, 2011,30 (02): 240-245.
- [15] Berrendero J, Bueno-Larraz B, Cuevas A. On mahalanobis distance in functional settings [J]. Journal of Machine Learning Research, 2020, 21:1-33.
- [16] Galeano P, Joseph E, Lillo R E. The Mahalanobis distance for functional data with applications to classification[J]. Technometrics, 2015, 57(2):281-291.
- [17] Mahalanobis P C. On the generalized distance in statistics [J]. Proceedings of the National Institute of Sciences, 1936, 2:49-55.
- [18] Ramsay J, Silverman W B. Functional data analysis [M]. Springer Science and Business Media, 2005.
- [19] <http://www.timeseriesclassification.com>
- [20] Chiou Jeng-Min and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data [J]. Journal of the Royal Statistical Society: Series B. 69 (4). 679-699. 2007
- [21] Jacques J, Preda C. Funclust: A curves clustering method using functional random variables density approximation [J]. Neurocomputing, 2013, 112: 164-171.
- [22] Bouveyron C, Jacques J. Model-based clustering of time series in group-specific functional subspaces [J]. Advances in Data Analysis and Classification, 2011, 5: 281-300.
- [23] Jiang H, Serban N. Clustering random curves under spatial interdependence with application to service accessibility [J]. Technometrics, 2012, 54(2): 108-119.
- [24] Giacomini M, Lambert-Lacroix S, Marot G, et al. Wavelet-based clustering for mixed-effects functional models in high dimension [J]. Biometrics, 2013, 69(1): 31-40.
- [25] Yassouridis C, Ernst D, Leisch F. Generalization, combination and extension of functional clustering algorithms: the R package fnc [J]. Journal of Statistical Software, 2018, 85: 1-25.