Function clustering analysis method based on C-Bézier curves

Mao Shi¹,a*, Ruowei Zhang²,b*

¹School of Mathematics and Information Science Shaanxi Normal University Xi'an, China
²Xing Zhi College Xi'an University of Finance and Economics Xi'an, China
a*Corresponding Author e-mail: shimao@snnu.edu.cn
b*Co-Corresponding Author e-mail: ruowei_zhang@163.com

Abstract—With the development of machine learning and big data research, curve clustering analysis has become an important part of multivariate statistics. After analyzing the pros and cons of existing curve clustering methods based on polynomial basis, we present a curve clustering method based on power basis and trigonometric function. Using the $L^2$ norm as a closeness measurement, the C-Bézier curve is introduced into the construction of the curve clustering algorithm to achieve the connection between curve clustering and traditional multivariate statistical clustering methods. As an application, a functional clustering example using the average price of several major foods in 50 cities as the object to verify the results shows that such curves can achieve better clustering results.

1. INTRODUCTION
Cluster analysis is an important multivariate statistical method, which has been applied in many fields. The basic idea is to classify samples or variables according to the closeness between the data. Classical clustering methods include systematic clustering, dynamic clustering and so on. However, with the development of computer software and hardware systems, data collection has become more and more intensive, and the structure of statistical data has become more and more complex. Therefore, a data type with obvious functional properties has appeared, which is called functional data. For this type of data, the traditional method is to use multivariate statistical methods for data analysis. Its disadvantage is that the geometric properties of the data itself are not taken into account, so that the geometric information contained in the function data cannot be studied comprehensively; another disadvantage is that the traditional statistical tools are faced with the instability accompanied by the numerical solution of singular matrix in the process of calculation[1-3].

In order to deal with the above problems, Ramsay and Dalzell formally put forward the concept of function data analysis in 1991, and then many experts and scholars studied it extensively[3-5]. The problem of cluster analysis of function data curves currently includes two methods: One is to construct a reduced-dimensional clustering algorithm based on the spectral decomposition of covariance operator of the fitting curve. Another is to use the finite dimensional function space to approximate the infinite dimensional data function, and then use the method of multivariate statistics to realize clustering analysis. The commonly used basis functions of this kind of methods are power basis function, Bernstein basis function, Fourier basis function and so on. However, the calculation of power basis function is unstable. The Bernstein basis functions are a set of polynomial bases based on the Weierstrass approximation theorem. Although the polynomial curve consisted by Bernstein basis
functions has some properties such as endpoint preserving interpolation and convex hull. It has no local shape adjustment and continuous order differentiable, as well as cannot accurately represent the quadratic curve[6][7]. Fourier basis functions are trigonometric function, which can accurately represent some curves which cannot be expressed analytically by polynomial curves, but it is often used in the clustering study of periodic curves. In this paper, we will discuss the clustering analysis of C-Bézier basis functions consisting of \([l,t,\sin t,\cos t]\). It was first given by Zhang Jiwen of Zhejiang University. This type of curve not only has all the geometric features of Bézier curves, but also can accurately represent quadratic curves[8].

The paper is arranged as follows: For the sake of the integrity of the paper, firstly, the definition of C-Bézier basis functions are given. Secondly, according to the research results of Abraham and other scholars, on the basis of C-Bézier basis, the degree of affinity of the curve is discussed in \(L^2\) space. Finally, it is applied to the field of economic analysis, and the application examples of function clustering of several kinds of main food average prices in 50 cities are given to verify the effectiveness of this method.

2. CONSTRUCTION OF CLUSTERING METHOD
In this section, we first give the definition of C-Bézier curve and its first and second derivative functions, and then we study the corresponding Euclidean distance given in [1], so as to define the closeness measurement of two curves.

2.1. Definition of curve
The C-Bézier curve is a kind of curve similar to Bézier curve, which basis function is composed of \([l,t,\sin t,\cos t]\), and can express some geometry shapes that polynomial curve can't express [7]. The basic definitions are as follows:

\[
x(t) = Z_0(t)q_0 + Z_1(t)q_1 + Z_2(t)q_2 + Z_3(t)q_3,
\]

(1)

where

\[
Z_0(t) = \frac{(\alpha - t) - \sin(\alpha - t)}{\alpha - \sin \alpha}, \quad Z_1(t) = M \left[ \frac{1 - \cos(\alpha - t)}{1 - \cos \alpha} - Z_0(t) \right],
\]

\[
Z_2(t) = M \left[ \frac{1 - \cos(t)}{1 - \cos \alpha} - Z_1(t) \right] \quad \text{and} \quad Z_3(t) = \frac{t - \sin t}{\alpha - \sin \alpha}
\]

are called C-Bézier basis functions, and

\[
M = \begin{cases} 
\frac{1}{S} & \alpha = \pi, \quad S = \sin \alpha, \\
0 \leq \alpha < \pi, \quad S = \sin \alpha, 
\end{cases}
\]

\[
C = \cos \alpha, \quad K = \frac{\alpha - S}{1 - C}. 
\]

For the simplicity, equation (1) can be rewritten in matrix form as

\[
X(t) = \frac{1}{\alpha - S} \begin{pmatrix} \sin t & \cos t & 1 \end{pmatrix} \begin{pmatrix} C & 1 - C - M & M - 1 \\ -S & -(\alpha - K)M & KM - 0 \\ -1 & M - M & 1 \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \\ q_2 \end{pmatrix}
\]

\[
= \frac{1}{\alpha - S} ABQ
\]

The corresponding first and second derivative functions are

\[
X'(t) = \frac{1}{\alpha - S} \begin{pmatrix} \cos t & \sin t & 0 \end{pmatrix} \begin{pmatrix} C & 1 - C - M & M - 1 \\ -S & -(\alpha - K)M & KM - 0 \\ -1 & M - M & 1 \end{pmatrix} \begin{pmatrix} q_0 \\ q_1 \\ q_2 \end{pmatrix}
\]

\[
= \frac{1}{\alpha - S} A'BQ
\]

and
\[
X'(t) = \frac{1}{a - S} \begin{pmatrix} -\sin t & \cos t & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} C & 1 - C - M & M - 1 \\ 1 - S & (\alpha - K)M & -K M \\ -1 & M & 0 \end{pmatrix} \begin{pmatrix} q_a \\ q_b \\ q_c \end{pmatrix}
\]

\[
= \frac{1}{a - S} A'BQ.
\]

Especially, when \(\alpha = 0\), the C-Bézier curve becomes the Bézier curve.

2.2. Function data generation

Given \(n\) observed data

\[
Y = \begin{bmatrix} y_{i1} & y_{i2} & \cdots & y_{in} \end{bmatrix}^T, \quad (i = 1, 2, \ldots, n)
\]

which is determined by the following formula:

\[
y_i(t) = X_i(t_k) + \varepsilon_i, \quad j = 1, 2, \ldots, n
\]

where \(t_k\) are discrete variables, \(\varepsilon_i\) is an independent random variable with the same variance and zero mean and \(X_i(t_k)\) are C-Bézier curve. In order to solve the control points \(q\), and to construct the clustering of curves, the least square method is used.

2.3. Distance of functions

Let two curves \(X_i(t)\) and \(X_j(t)\) being C-Bézier curve. The control points are \(q_i = [q_a^i, q_b^i, q_c^i, q_d^i]\) and \(q_j = [q_a^j, q_b^j, q_c^j, q_d^j]\), respectively. Its Euclidean square distance is as follows:

\[
d^2(X_i, X_j) = \int_0^T (X_i(t) - X_j(t))^2 dt
\]

\[
= [q_a^i - q_a^j]^T C [q_a^i - q_a^j]
\]

where

\[
C = \begin{bmatrix}
(Z_1, Z_1) & (Z_2, Z_1) & (Z_3, Z_1) \\
(Z_2, Z_2) & (Z_2, Z_1) & (Z_2, Z_2) \\
(Z_3, Z_3) & (Z_3, Z_2) & (Z_3, Z_2) \\
(Z_4, Z_4) & (Z_4, Z_3) & (Z_4, Z_2) \\
\end{bmatrix}
\]

is a \(3 \times 3\) real symmetric matrix, and \(\int_0^T Z_i \cdot Z_j dt\).

In order to realize the application of curve clustering for multivariate statistical clustering, Cholesky triangular matrix decomposition is generally done on the real symmetric matrix, that is,

\[
L = U^T U
\]

Letting

\[
b = U(q^a - q^j)
\]

equation (4) can be written as

\[
d^2(X_i, X_j) = b^T b.
\]

2.4. Distance of derivatives

Because the analysis of function clustering cannot get the geometric information of the curve, sometimes it is necessary to analyze the curve derivatives. Then according to equations (2) and (3), we can get the corresponding derivative distance formulas

\[
d^2(X'_i, X'_j) = \int_0^T (X'_i(t) - X'_j(t))^2 dt
\]

\[
= \frac{1}{a - S} [q_a^i - q_a^j]^T B^T C' B (q_a^i - q_a^j)
\]

and

\[
d^2(X''_i, X''_j) = \int_0^T (X''_i(t) - X''_j(t))^2 dt
\]

\[
= \frac{1}{a - S} [q_a^i - q_a^j]^T B'' C'' B (q_a^i - q_a^j)
\]
where both \( C' = \int_0^1 (A' A') dt \) and \( C'' = \int_0^1 (A'' A') dt \) are real symmetric matrices. Similar to the simplified expression of the distance formula, equations (5) and (6) can also be simplified as
\[
d^2(A', A'') = b^T b'^T
\]
and
\[
d^2(A'', A''') = b'^T b''
\]
where
\[
b'_q = \frac{1}{a-S} U(q'-q'), \quad b''_q = \frac{1}{a-S} U(q''-q')
\]
and
\[
L' = (U')^T U', \quad L'' = (U'')^T U''.
\]

2.5. Algorithms
With the above preparations, we can perform cluster analysis on the curves. If we only focus on the curves itself, we can use the curve distance equation (4). If we want to study the speed of the curve, we can use the derivative distance equations (5) or (6) construct the corresponding cluster analysis. In order to achieve the acquisition of information such as the curve itself or the characteristics of the speed of change. Specific steps are as follows:

- Step 1: Select the fitted interval;
- Step 2: Give the value of the parameter \( \alpha \);
- Step 3: Construct the corresponding C-Bézier curve;
- Step 4: Calculate the vector, or according to equations (4), (5) or (6), and then perform clustering;
- Step 5: Restore the clustering result to the clustering result of the curve.

3. Example Analysis
In order to explain the idea of the above method more clearly, we have selected the price data of several kinds of major food in 50 cities of China for analysis, and the data is shown in Table 1.

| Time  | Eggs | Beef | Pork | Hairtail | Cabbage |
|-------|------|------|------|----------|---------|
| 09.01-10 | 11.02 | 67.15 | 30.49 | 32.35 | 3.44 |
| 09.11-20 | 10.96 | 67.05 | 30.49 | 32.32 | 3.39 |
| 09.21-30 | 10.69 | 67.04 | 29.88 | 32.11 | 3.22 |
| 10.01-10 | 10.22 | 67.24 | 29.67 | 32.51 | 3.00 |
| 10.11-20 | 9.82 | 67.32 | 29.31 | 32.34 | 2.72 |
| 10.21-30 | 9.67 | 67.32 | 29.57 | 32.32 | 2.52 |
| 11.01-10 | 9.62 | 67.21 | 28.58 | 32.21 | 2.41 |
| 11.11-20 | 9.68 | 67.24 | 29.00 | 32.25 | 2.38 |
| 11.21-30 | 9.80 | 67.23 | 28.97 | 32.50 | 2.53 |
| 12.01-10 | 9.86 | 67.21 | 28.29 | 32.58 | 2.47 |
| 12.11-20 | 9.86 | 67.16 | 28.29 | 32.70 | 2.40 |
| 12.21-30 | 9.89 | 67.38 | 29.06 | 32.82 | 2.36 |
| 1.01-10 | 9.88 | 67.03 | 29.09 | 32.56 | 2.31 |
Because the data in Table 1 is discrete, first fit it with C-Bézier, then analyze the function characteristics of the fitting curve, so as to determine the division interval of the function data, and finally get the sample clustering results on different intervals as shown in Figures 1-5.

Figure 1. The price of eggs

Figure 2. The price of beef

Figure 3. The price of pork
4. CONCLUSION

Function data analysis greatly enriches the field of data analysis. From the perspective of clustering analysis, this paper discusses the application of a kind of mixed basis function in the field of clustering analysis. At the same time, it gives the clustering model of the main food prices in 50 major cities in China, and gets better results. However, there are still some problems that need to be improved, such as keeping the price in balance, using polynomial or C-Bézier curve fitting, there is a phenomenon of convex up or convex down. In this paper, we have made artificial correction. The in-depth study of clustering algorithm will be one of our main work in the future.

ACKNOWLEDGMENT

The work is supported by the Fundamental Research Funds for the Central Universities (GK201703007) and Sample data processing and analysis system based on Artificial Intelligence (H190074)

REFERENCES

[1] H. J. HUANG, “Curves Clustering Using B-splines Expansion”. Statistics & Information Forum. 28(9) 3-8. 2013.
[2] J.O. Ramsay, “When the Data are Function”. Psychometrika, 47:379-396. 1982.
[3] J.O. Ramsay, B.W. "Silverman Function Data Analysis”, 2ed. New York: Springer. 2005.
[4] Y. Y. Zeng, J. Z. Weng, “Initiative Research of the Clustering Approach About Function Data”. Statistics & Information Forum. 22:10-14,2007.
[5] J. Wang, K.F. Huang, H. W.Wang “A Cluster Method of Functional Data Analysis”. Application of Statistics and Management, 28:839-844. 2009.

[6] L.Piegl, W.Tiller, The NURBS Book, 2ed, New York: Springer 1997.

[7] G.Farin, Curves and Surfaces for Computer Aided Geometric Design, Academic Press, San Diego. 1988.

[8] J.W. Zhang, C-Curves: An extension of Cubic Curves. Computer Aided Geometric Design, 13:199-217. 1996.