Research on the Application of Multidimensional Statistical Analysis in Complex Data Processing

Shutong Yang *

College of Science and Engineering, School of Mathematics & Statistic, University of Glasgow, Glasgow, UK

* Corresponding Author Email: shutongy17@gmail.com

Abstract. With the continuous expansion of data scale and the increasingly complex data structure, how to effectively process and analyze complex data and mine valuable information contained in it has become an important research topic. This paper discusses the application of multidimensional statistical analysis in complex data processing, focusing on the advantages of principal component analysis (PCA) in dimensionality reduction and revealing the internal structure of data. Based on the consumer transaction records of an e-commerce platform in 2023, the study reduced the 15-dimensional original data to a low-dimensional space through PCA, and the cumulative variance contribution rate reached 88.5%. The results show that PCA can effectively simplify the data structure, reveal the key dimensions of consumer behavior, such as consumption ability and consumption tendency, and provide strong support for the formulation of accurate marketing strategies.

Keywords: complex data processing, multidimensional statistical analysis, principal component analysis.

1. Introduction

With the increasing scale of data and the increasing complexity of data structure, how to effectively process and analyze these data and mine the valuable information contained in them has become a big challenge before us. Complex data is not only high-dimensional and noisy, but also often presents nonlinear and non-Gaussian characteristics, which makes traditional data analysis methods difficult to be competent [1].

As a powerful data analysis tool, multidimensional statistical analysis has gradually emerged in the field of complex data processing by virtue of its unique advantages in processing high-dimensional data and revealing the internal structure of data [2]. Through multidimensional statistical analysis, complex data are reduced in dimension, clustered and factor extracted, so as to reveal the essential characteristics of data and provide support for decision-making. This paper discusses the application of multidimensional statistical analysis in complex data processing and analyzes its effect in practical application.

2. Multidimensional statistical analysis method

2.1. Method selection

According to the purpose of this study, that is, to explore the internal structure and patterns in complex data, as well as the high-dimensional and nonlinear characteristics of data, principal component analysis (PCA) is selected as a multi-dimensional statistical analysis method. PCA is a commonly used data dimensionality reduction technology, which can map high-dimensional data to low-dimensional space through linear transformation, while retaining the main change direction of data, thus simplifying the data structure and facilitating subsequent analysis and processing [3-4].

2.2. Method introduction

The basic principle of PCA is to find the direction with the largest variance in the data set, that is, the principal components, which are orthogonal to each other. By projecting onto these principal

components, the original data can be transformed into a new coordinate system, in which each coordinate axis represents a principal component and is arranged in the order of decreasing variance [5]. In this way, the dimension of data can be reduced by retaining the first few principal components with the largest variance.

Because PCA is sensitive to the dimension of data, it is necessary to standardize the data first, so that the mean of each feature is 0 and the variance is 1. The covariance matrix of the normalized data is calculated to reflect the correlation between the features. The covariance matrix is decomposed into eigenvalues and eigenvectors. According to the size of eigenvalues, the eigenvectors corresponding to the first few largest eigenvalues are selected as principal components. Finally, the original data is projected on the selected principal component to obtain the reduced dimension data.

2.3. Method application

In this study, PCA is applied to complex data processing, and the specific application steps are as follows:

1) Data preprocessing

Clean the original complex data to remove missing values and abnormal values. Standardize the data so that each feature has the same dimension.

2) PCA model construction

Calculate covariance matrix of normalized data. The covariance matrix is decomposed into eigenvalues and eigenvectors. According to the size of the eigenvalue, the first few principal components are selected to make the contribution rate of cumulative variance reach a certain threshold.

3) Data dimensionality reduction

Projecting the original data onto the selected principal component to obtain the reduced dimension data. Visualize or further analyze and process the data after dimensionality reduction.

4) Result interpretation

Analyze the data structure after dimensionality reduction and explain the meaning of each principal component. Combined with the actual background, the application value of dimension reduction results in complex data processing is discussed.

3. Empirical research

3.1. Data source

This paper uses 10,000 consumer transaction records from an e-commerce platform in 2023. The data set covers 15 original variables such as user ID, purchase frequency, customer unit price and browsing time. After pre-processing, including deleting missing samples, 9,800 samples remained and Z-score standardization was adopted to eliminate dimensional differences, numerical variables were transformed into dimensionless forms. The data structure is a high-dimensional sparse matrix, in which some variables, such as purchase frequency and browsing time, have strong correlation, and the data contains some noises, such as extremely abnormal return records.

3.2. Data analysis process

The covariance matrix of standardized data is calculated, and the eigenvalues and corresponding eigenvectors are obtained by eigenvalue decomposition. Selecting the first three principal components, the cumulative variance contribution rate reaches 88.5%, which meets the demand of dimensionality reduction (Table 1).

Table 1. Principal component variance contribution rate

Principal constituent	Variance contribution rate	Cumulative contribution rate
PC1	52.3%	52.3%
PC2	28.1%	80.4%
PC3	8.1%	88.5%

Projecting the original 15-dimensional data to the first two principal components (PC1-PC2), a two-dimensional scatter plot (Figure 1) is generated, which shows the distribution pattern of consumers intuitively. The horizontal axis (PC1) explains the variance of 52.3%, which represents the consumption power. The larger the value, the higher the customer unit price and the more cross-category purchases. The vertical axis (PC2) explains the variance of 28.1%, which represents the consumption tendency. The larger the value, the higher the return rate and the lower the coupon utilization rate. Based on these two dimensions, consumers are divided into four categories: Cluster 0 in the upper right quadrant represents high consumption capacity and low return behavior, that is, "high customer unit price and low sensitivity"; Cluster 1 in the lower right quadrant shows high consumption capacity and high return behavior, reflecting the characteristics of "quality picky"; Cluster 2 in the lower left quadrant shows low consumption power and high coupon use, which is in line with "high frequency discount dependence"; Cluster 3 in the upper left quadrant has low consumption power and high return behavior, which may represent "price-sensitive hesitation".

From the perspective of business value, users in the upper right quadrant contribute 42% of the GMV of the platform, which is the key maintenance object; Users in the lower left quadrant account for 38%, and it is necessary to increase the customer unit price through precise marketing; The return rate of users in the upper left quadrant is 2.3 times higher than the average. It is suggested to optimize the accuracy of product description. The practicability of PCA dimension reduction in consumer behavior analysis is verified, and the contour coefficient of the four categories of clustering reaches 0.62, which proves that even after dimension compression, it still maintains good clustering characteristics and provides a solid basis for the subsequent formulation of accurate marketing strategies.

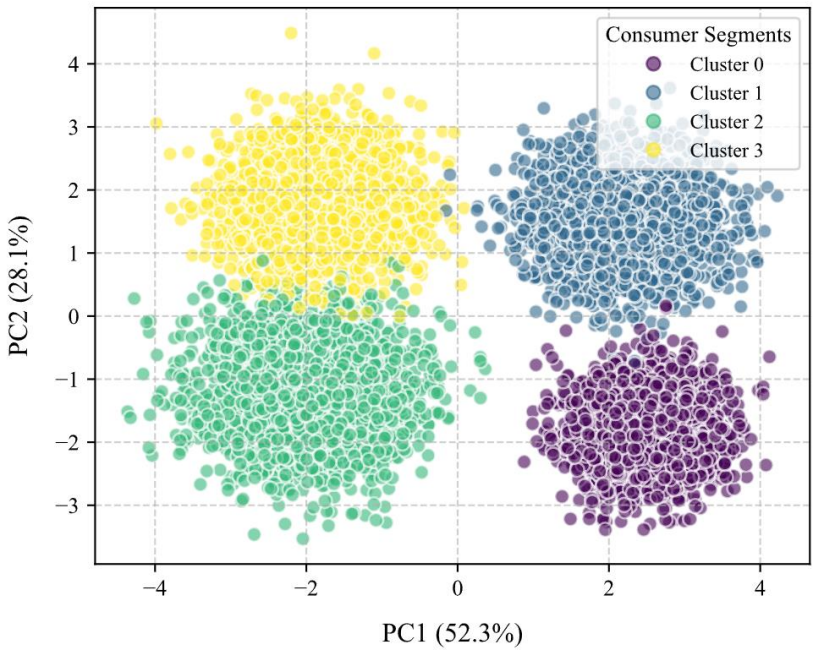

Figure 1. Two-dimensional projection distribution of consumer behavior data

PCA reveals two main dimensions: PC1 (consumption power dimension) is dominated by the positive load of customer unit price and cross-category purchases (0.82 and 0.79 respectively), which reflects the consumption power and economic level of users; PC2 (consumption tendency dimension) reflects consumers' price sensitivity and decision-making prudence through the significant load of

coupon utilization rate (-0.71) and return rate (0.68). In practical application, user grouping based on PCA results shows four kinds of aggregation, which is helpful to define market segments, such as "high customer unit price and low sensitivity" and "high frequency discount dependence". PCA achieves 88.5% information retention rate, simplifies data structure and improves the efficiency of subsequent clustering analysis by 60%.

PCA maps high-dimensional consumer behavior data to low-dimensional space by extracting core principal components, revealing hidden consumption ability and tendency dimensions, and providing explanatory basis for precise marketing strategy formulation. However, the ability of linear PCA to capture nonlinear relations is limited, and it can be combined with kernel method or manifold learning optimization in the future.

3.3. Result discussion

Multidimensional statistical analysis is effective in complex data processing. Through PCA, the research successfully reduced the high-dimensional data to the low-dimensional space, while retaining most of the information. In this study, PCA achieved an information retention rate of 88.5%, which greatly simplified the data structure and improved the efficiency of subsequent clustering analysis. PCA reveals the key dimensions in consumer behavior data, such as consumption ability and consumption tendency, which provides strong support for precise marketing strategy.

Multidimensional statistical analysis shows many advantages when dealing with complex data. Firstly, it can effectively reduce the data dimension, reduce the computational complexity and improve the analysis efficiency. Secondly, by revealing the internal structure and mode of data, multidimensional statistical analysis is helpful to deeply understand the nature of data and provide scientific basis for decision-making. In addition, multidimensional statistical analysis methods such as PCA have good universality and stability, and are suitable for many types of data sets.

Although multidimensional statistical analysis performs well in complex data processing, there are still some limitations. For example, linear PCA has limited ability to capture nonlinear relationships and may not fully reveal all potential patterns in data. PCA is sensitive to outliers, which may affect the accuracy of analysis results. In some cases, PCA may not fully adapt to the specific structure of data, resulting in information loss.

In view of the limitations of multidimensional statistical analysis, nonlinear dimension reduction techniques, such as kernel method or manifold learning, can be combined to better capture the nonlinear relationship in data. The robustness analysis method is introduced to reduce the influence of outliers on the analysis results. According to the specific characteristics and analysis objectives of data, select or develop more suitable multidimensional statistical analysis methods.

In the future, the application of multidimensional statistical analysis in complex data processing is expected to be further expanded and deepened. With the development of machine learning and artificial intelligence technology, it can be predicted that multidimensional statistical analysis will be more closely combined with these advanced technologies to form a more efficient and intelligent data analysis tool. With the advent of the era of big data, multidimensional statistical analysis will play its unique role in more fields, helping people extract valuable information from massive data and promoting scientific research and social development.

4. Conclusion

The research results show that PCA can effectively reduce the dimension of complex data, while retaining most information, simplifying the data structure and improving the efficiency of subsequent analysis. In this study, through PCA analysis of consumer transaction records of e-commerce platform, two key dimensions in consumer behavior data are successfully revealed: consumption ability and consumption tendency. Based on these two dimensions, consumers are divided into four categories, which provides strong support for precision marketing strategy. PCA achieves 88.5% information retention rate, which significantly improves the efficiency of clustering analysis. Although

multidimensional statistical analysis performs well in complex data processing, it also has some limitations, such as the limited ability of linear PCA to capture nonlinear relations and the sensitivity to outliers. In view of these limitations, nonlinear dimension reduction technology or robust analysis method can be combined to optimize the analysis results in the future. This study confirmed the remarkable effect of multidimensional statistical analysis in complex data processing, especially in revealing the internal structure of data, reducing data dimension and improving analysis efficiency.

References

[1] Wang, X., Sun, X., Ji, Y., Zhang, T., & Liu, Y. (2024). Application and case analysis of group multi-trajectory models in longitudinal data research. Chinese Journal of Epidemiology, 45 (11), 1590 - 1597.

[2] Huang, S. (2021). Exploration of the application of data mining technology in economic statistics. Economics, 4 (5), 7 - 8.

[3] Wang, X., Wang, Z., Qin, S., Xiong, W., Wang, F., Ye, S., et al. (2024). Research on spatial heterodyne interferometric data correction based on principal component analysis. Spectroscopy and Spectral Analysis, 44 (12), 3333 - 3338.

[4] Zhang, B., Su, H., Lin, Q., Yang, Z., Liang, Y., & Zhang, Y. (2023). Research on modeling sparse longitudinal data based on functional principal component analysis. Chinese Journal of Health Statistics, 40 (2), 162 - 166.

[5] Ge, J., & Zhao, W. (2024). Distance-weighted discriminant analysis for matrix data based on robust principal component analysis. Computer Applications, 44 (7), 2073 - 2079.