Determination of Aquatic Product Growth Factors Based on PCA with Stepwise Regression Test

Yingkun Zheng¹, Xufeng Hua¹²*, Yu Zhang¹, Yunchen Tian¹²*, Yangyang Xue³ and Manjiang Wang⁴

¹ School of Computer and Information Engineering, Tianjin Agricultural University, Tianjin, 300384, China.
² Tianjin Key Laboratory of Aquatic Ecology and Aquaculture, Tianjin, 300384, China.
³ Tianjin Tongyang Agricultural Technology Co., Ltd, Tianjin, 300270, China.
⁴ Tianjin Yililai breeding Co., Ltd, Tianjin, 300380, China.
Corresponding Author: Xufeng Hua and Yunchen Tian
¹ Email: huaxufeng@163.com

Abstract. For the information contained in the growth data of the pearl gentian grouper, there is overlap and inclusion. When establishing the growth model of the pearl gentian grouper, it is difficult to grasp the inherent laws of the growth factors to affect the establishment of the research target model and the operation efficiency. The data dimensionality reduction method is often used in the data preprocessing stage. The Principal Components Analysis (PCA) method was used to analyze and influence the growth factors of the pearl gentian grouper. The method of regression fitting optimizes the results of selecting principal components under different criteria. Experimental results show that the main growth factors of the pearl gentian grouper selected by this method can meet the needs of the construction of aquaculture growth models, and can provide a reference for reducing the dimensions of aquaculture fish growth data and the complexity of the modeling process.

1. Introduction

With the improvement of detection technology and the popularization of Internet of things, a large number of multi factor data affecting the growth of aquatic products have been collected. The data of each influencing factor can reflect some information of aquatic product growth index, and at the same time, there is a general correlation between the influencing factor data, which leads to the overlap of the information reflected in the data to a certain extent. In view of this kind of phenomenon, many methods of filtering data dimensions have been proposed. From the perspective of spectral analysis, it can be divided into linear and nonlinear dimensionality reduction [1]. In reference [2], it is pointed out that linear dimensionality reduction has many advantages, such as simple calculation, often analytical solution, very effective for data set with linear structure, and easy to explain. Because of the advantages of linear dimensionality reduction method, linear dimensionality reduction gradually becomes the main means of high-dimensional data processing when solving real problems. Dong [3] pointed out in the comparative study of Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) that the dimension of PCA after dimensionality reduction can be specified as needed, and LDA cannot be greater than n-1 dimension. Gao [4] combined PCA with neural network and applied it to the prediction of urban water consumption, which improved the accuracy of the prediction results of the model. In theory, reference [5] pointed out that PCA is a linear dimension reduction method with...
the least loss of original data information. Reference [6] proposed "principal component artificial neural network photometry", which retains the important information needed for prediction in application, reduces the input dimension of neural network, and improves the operation efficiency of the model. It is pointed out in reference [7] that the accuracy of MDS is slightly worse than that of PCA. Through the summary, it is found that the PCA algorithm is a statistical analysis method for screening data dimensions and extracting multi-dimensional vectors of the main features of the data [8]. Such as conceptual simplicity, calculation convenience, and optimal linear reconstruction error, with these excellent characteristics, PCA method became one of the most widely used dimensionality reduction methods in actual data processing.

2. PCA with Stepwise Regression Test

2.1. Calculating of Eigenvalues and Eigenvectors
The eigenvalue $\lambda_i$ of the correlation matrix $R$ and the corresponding orthogonalized unit eigenvector $a_i$ need to be calculated first. The first $m$ larger eigenvalues of $R$, $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_m > 0$ is the variance corresponding to the first $m$ principal components, and the unit eigenvector $a_i$ corresponding to $\lambda_i$ is the coefficient of the principal component $Z_i$ about the original variable. The $i$-th principal component of is $Z_i$, that is, $Z_i = a_i^T X^*$. The information contribution rate $v_i$ reflects the ability of the $i$-th principal component to integrate the original indicator information. The larger the value, the stronger the ability. The calculation of the information contribution rate $v_i$ is shown in formula (1).

$$v_i = \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j}$$

2.2. Calculating of Principal Components Load and Score
The principal component load reflects the degree of correlation between the principal component $Z_i$ and the original variable $X_j$. The original variable $X_j$ is on each principal component $Z_i$. The load $l_{ij}$, is shown in formula (2).

$$l(Z_i, X_j) = \sqrt{\lambda_i} a_{ij}$$

The matrix composed of loads $l_{ij}$ is called the principal component load matrix which is shown in formula (3).

$$L = \begin{bmatrix} l_{11} & \cdots & l_{1m} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pm} \end{bmatrix}$$

The component matrix shown in formula (4) is the coefficient matrix when the principal components $Z_1, Z_2, \ldots, Z_m$ represent the normalized original variables $X_1^*, X_2^*, \ldots, X_p^*$.

$$\begin{bmatrix} X_1^* \\ \vdots \\ X_p^* \end{bmatrix} = L \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix}$$

After rotating the component matrix, the scores of the original variables on the $m$ principal components are obtained. The corresponding score matrix is

$$\begin{bmatrix} a_{11} & \cdots & a_{pm} \\ \vdots & \ddots & \vdots \\ a_{1m} & \cdots & a_{pm} \end{bmatrix}$$

Then the principal component $Z_i = a_{i1} X_1 + a_{i2} X_2 + \cdots + a_{ip} X_p, i = 1, 2, \ldots, m$. From the scoring matrix of principal components, we can see the influence of each original variable on the principal
component, and then we can explain the extracted principal component in combination with the actual meaning of the original variable.

2.3. Determination of Principal Components Based on Stepwise Regression

2.3.1. Determining Quantity of Main Components. When the primary variable index is comprehensively transformed to obtain the principal component, the number \( m \) of the principal component needs to be determined at the same time, so that the first \( m \) components \( Z_1, Z_2, ..., Z_m \) with large variance are extracted as the principal component. The criteria and methods commonly used to determine the amount of principal component extraction are cumulative contribution rate, Kaiser criterion, and gravel map. The cumulative contribution rate criterion is an empirical criterion. If the cumulative contribution rate \( G(m) \) of the first \( m \) principal components reaches 80% or more than 85%, and it is considered sufficient to reflect the information of the original variable, the number of principal components is \( m \). The scree plot is a visual representation of the contribution rate of the principal components, and can be used as a reference to determine the number of principal components. Determine the number of principal components by observing the "elbow" of the gravel map. The Kaiser criterion believes that if the eigenvalue of the principal component is less than 1, the principal component is untrustworthy, so the principal component with the eigenvalue greater than or equal to 1 should be retained.

2.3.2. Principal Component Test Based on Stepwise Regression. The methods for determining the quantity of principal components, such as cumulative contribution rate and gravel map, are mostly based on empirical judgment or subjective observation. Although simple and easy to implement, they lack a test of whether the extraction results can reconstruct the original data. In this paper, a stepwise regression analysis method is used to establish regression equations for the principal component factors determined by the three methods of cumulative contribution rate, Kaiser criterion, and gravel map. The optimal number of principal component selection is determined by comparing their respective fitting effects.

The complete model with \( p \) independent variables \( x_1, x_2, ..., x_p \) fitting dependent variable is shown in formula (5).

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon
\]  

(5)

Define \( \Delta R^2_i = R^2 - R_i^2 \), where \( R^2 \) is the remeasurement coefficient of the full model and \( R_i^2 \) is the remeasurement coefficient of this model. When \( \Delta R^2_i \) is almost 0, it means that the explanatory ability of adding independent variable \( x_i \) to \( y \) has not improved significantly; if \( \Delta R^2_i \) is not significantly zero, then \( x_i \) can provide significant explanatory information for the regression model.

3. Determination of Growth Factors for Pearl Gentian Grouper

3.1. Data Preparation

3.1.1. Data Acquisition. The data source of this article is the growth factor data of pearl gentian grouper cultured by a Tianjin marine fish culture enterprise in a factory circulating water environment. There are three main methods to obtain target data: sensor records, manual records and laboratory measurements. They have the characteristics of timing and controllability. The data acquisition process strictly followed the established breeding practices to ensure that the data was sufficient and complete. The data acquisition target selected for this study was a total of 8,500 pearl gentian grouper as a data collection sample for the entire breeding pond. From the beginning of fry to about 600g of adult fish, this type of growth data of 15 types of growth factors throughout the breeding cycle is a variety of data.
3.1.2. Data Standardization. Because the growth data of pearl gentian grouper contains many factors, and the size and unit of each indicator are particularly large, the original data is standardized using the mean and standard deviation to eliminate the dimensional impact, that is, transform the original data as shown in formula (6).

\[ x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, p \]  

(6)

Where \( \bar{x}_j \) is the mean of the j-th factor, and \( S(x_j) \) is the standard deviation of the j-th factor.

3.1.3. Correlation Test. According to the definitions of KMO test and Bartlett sphericity test, the correlation of the growth factor data matrix of pearl gentian grouper that has been standardized is tested. After calculation, the KMO value is 0.703, and the statistics of the Bartlett sphericity test is 105. According to the KMO metric, if the KMO value is between 0.7 and 0.8, the factors of the original data are considered to be correlated, and the PCA method can be used. According to the calculated statistics of the Bartlett sphericity test, the chi-square distribution table is queried to obtain an approximate chi-square 3699.777, and the corresponding accompanying probability value is close to 0, that is, there is a significant difference between the growth factor matrix and its unit matrix. According to the relationship between the association probability and the significance level, there is a correlation between the factors of the original data and the PCA method is suitable.

3.2. Determination of Growth Factors for Pearl Gentian Grouper Based on PCA

For the standardized growth factor data of the pearl gentian grouper, the calculated principal component characteristic values, information contribution rate, and cumulative contribution rate are shown in Table 1. The calculated composition matrix coefficients are shown in Table 2.

**Table 1.** Calculation results of components

| Component ID | Feature value | Variance percentage | Accumulation percentage |
|--------------|---------------|---------------------|------------------------|
| 1            | 5.918         | 39.457              | 39.457                 |
| 2            | 2.174         | 14.492              | 53.948                 |
| 3            | 1.653         | 11.019              | 64.967                 |
| 4            | 1.125         | 7.497               | 72.464                 |
| 5            | 1.052         | 7.011               | 79.475                 |
| 6            | 0.932         | 6.216               | 85.691                 |
| 7            | 0.581         | 3.874               | 89.565                 |
| 8            | 0.469         | 3.129               | 92.694                 |
| 9            | 0.325         | 2.169               | 94.863                 |
| 10           | 0.227         | 1.510               | 96.374                 |
| 11           | 0.180         | 1.202               | 97.575                 |
| 12           | 0.171         | 1.140               | 98.716                 |
| 13           | 0.136         | 0.906               | 99.622                 |
| 14           | 0.039         | 0.258               | 99.880                 |
| 15           | 0.018         | 0.120               | 100.000                |
### Table 2. Component matrix (excerpt)

| Growth factors | Component1 | Component2 | Component3 | Component4 | ... | Component15 |
|----------------|------------|------------|------------|------------|-----|-------------|
| Survival days  | 0.957      | -0.196     | 0.071      | 0.005      | ... | 0.044       |
| Dead fish quantity | 0.219 | 0.499     | 0.395      | -0.369     | ... | -0.173      |
| Salinity       | -0.864     | 0.287      | -0.094     | 0.096      | ... | -0.122      |
| Water temperature | -0.611 | 0.033     | 0.129      | 0.166      | ... | -0.079      |
| Feed diameter  | 0.824      | -0.298     | 0.060      | -0.085     | ... | 0.353       |
| Total feed     | -0.321     | -0.771     | 0.013      | -0.227     | ... | 0.010       |
| Chilled feed quantity | 0.630 | 0.566     | -0.364     | 0.218      | ... | 0.023       |
| Individual consumption | 0.726 | 0.334     | -0.323     | 0.332      | ... | -0.001      |
| Dissolved oxygen | -0.050 | 0.099     | -0.075     | 0.556      | ... | 0.164       |
| Nitrate        | 0.253      | 0.598      | 0.408      | -0.402     | ... | 0.066       |
| Ammonia nitrogen | 0.800 | 0.144     | -0.227     | -0.307     | ... | 0.080       |
| Total vibrio   | -0.039     | 0.284      | 0.778      | 0.334      | ... | -0.249      |
| Total bacteria | 0.247      | -0.284     | 0.600      | 0.275      | ... | -0.091      |
| Illumination intensity | -0.888 | 0.261     | -0.106     | -0.054     | ... | 0.023       |
| Culture density | -0.753 | 0.272     | -0.165     | -0.201     | ... | 0.232       |

The number of principal components extracted from all components using the cumulative contribution rate, Kaiser criterion, and scree plot are shown in Table 3.

### Table 3. Number of principal components by different criteria

| Selection criteria       | Number of principal component |
|-------------------------|-------------------------------|
| Cumulative contribution rate | 6                             |
| Kaiser criterion        | 5                             |
| Scree plot              | 4                             |

3.3. Growth Factors Test Based on Stepwise Regression

In this paper, stepwise regression analysis method is used to establish regression equations, and the principal components selected under the three criteria are tested respectively. The regression equation contains all the significant growth factors of the pearl gentian grouper selected by the PCA. The
stepwise regression method is used to fit the actual weight value of the target factor pearl gentian grouper. Using MATLAB 2016a to fit the body weight of the pearl gentian grouper.

By comparing the fitting result with the actual value, it is found that the fitting result of the maximum contribution rate Kaiser criterion is similar, and the fitting effect is better, which can better reflect the relationship between each major growth factor and the body weight of the pearl gentian grouper. The comparison between the fitting results of the selected principal components based on the gravel map and the actual weight shows that the fitting weight fluctuates greatly and the effect is poor.

According to the comparison of regression fitting coefficient, variance R-square and root mean square error RMSE, Table 4 shows that the maximum contribution rate and the weight fitting variance under the Kaiser criterion are close to 1, and the maximum contribution rate criterion weight fitting variance reaches 0.964, the results show that the number of extraction of the main components determined by the maximum contribution rate is 6 optimal, that is, the first 6 main components are selected as the growth factors of the pearl gentian grouper to establish the growth model.

| Selection criteria      | R-square | RMSE |
|------------------------|----------|------|
| Cumulative contribution rate | 0.964    | 32.9 |
| Kaiser criterion       | 0.959    | 35.2 |
| Scree plot             | 0.404    | 135  |

4. Summary
Because there is a certain correlation among the factors that affect the growth and development, the information contained in the growth factors overlaps, which increases the complexity of calculating, building and analyzing the growth prediction model of aquatic products based on such data. The PCA algorithm was used to screen out the main growth factors for the growth data of Pearl gentian grouper. A stepwise regression fitting model is proposed to verify the main growth factors, determine the number of main growth factors, and reduce the dimension of the original data for the purpose of reducing the correlation of growth factors.

5. Acknowledgments
We thank Mr. Xuequan Yu for his help in data sample and analysis from Tianjin Lida sea water resources development Co., Ltd. This research was supported by China Agriculture Research System (CARS-47), Tianjin Technical Expert Project (19JCTPJC58800) and Tianjin Mariculture Research System.

6. References
[1] S Yi, Z Lai, Z He, Y Cheung and Y Liu 2017 Joint sparse principal component analysis Pattern Recognition 61 524
[2] Li Hongqiang, Wei Xiaoming, Wang Youxi, Zhang Zhen, Gong Zheng and Wu Feifan 2019 Application of principal component analysis and linear discriminant analysis to feature extraction and diagnosis algorithm of ECG signal Biomedical Engineering Research 38 145
[3] Dong Husheng 2016 Comparative study on two data dimensionality reduction algorithms of principal component analysis and linear discriminant analysis Modern Computer (Professional Edition) 29 36
[4] Gao Xueping, Chen Lingling and Liu Yinzhu 2017 Urban water consumption prediction based on PCA-RBF neural network model Water Conservancy and Hydropower Technology 48 1
[5] Liu Fen 2019 Efficient face recognition system based on linear and nonlinear algorithm Electronic technology 32 82
[6] Maria Celeste da Silva Sauthier, Erik Galvão Paranhos da Silva, Bruna Rosa da Silva Santos, Emmanuelle Ferreira Requião Silva and Walter Net Lopes dos Santos 2019 Screening of Mangifera indica L. functional content using PCA and neural networks (ANN) Food Chemistry 2731 115
[7] Kensuke Tanioka and Hiroshi Yadohisa 2018 Asymmetric MDS with Categorical External Information Based on Radius Model *Procedia Computer Science* **140** 284

[8] Mohammad Jafarzadegan, Faramarz Safi-Esfahani and Zahra Beheshti 2019 Combining hierarchical clustering approaches using the PCA method *Expert Systems with Applications* **13715** 1