Analysis and Applications of PCA Information Features

DING Shifei¹,², SHI Zhongzhì²

1. College of Information Science and engineering, Shandong Agricultural University, Taian, 271018, P.R.China
2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, P.R.China

Abstract: This paper makes a brief introduction about the principle of principal component analysis (PCA). Then according to information entropy theory, and making full use of the inherent characteristic of eigenvalues of correlation matrix of data, the 2nd information function, the 2nd information entropy and geometry entropy under PCA are proposed firstly, by which the information features of PCA are metricized. In addition, two new concepts of information rate (IR) and accumulated information rate (AIR) are proposed, which are used to illustrate the degree of information feature extraction of PCA. In the end, through simulated application in practice, the results show that the method proposed in this paper is efficient and satisfactory. It provides a new research approach of information feature extraction for pattern recognition, machine learning, data mining and so on.

Key words: information entropy theory; principal component analysis; information feature extraction; information rate; accumulated information rate

1. Introduction

Pattern recognition is the scientific discipline whose goal is the classification of objects into a number of categories or classes. Depending on the application, these objects can be images or signal waveforms or any type of measurements that need to be classified. We will refer to these objects using the generic term patterns. Pattern recognition has a long history, but before the 1960s it was mostly the output of theoretical research in the area of statistics. As with everything else, the advent of computers increased the demands for practical applications of pattern recognition, which in turn set new demands for further theoretical developments. As our society evolves from the industrial to its post-industrial phase, automation in industrial production and the need for information handling and retrieval are becoming increasingly important. This trend has pushed pattern recognition to the high edge of today’s engineering applications and research. Pattern recognition is an integral part in most machine intelligence systems built for decision-making.

We assumed that a set of training data were available, and the classifier was designed by exploiting this a priori known information. This is known as supervised pattern recognition. However, this is not always the case, and there is another type of pattern recognition tasks for which training data, of known class labels, are not available. In this type of problem, we are given a set of feature vectors X and the goal is to unravel the underlying similarities, and cluster (group) “similar” vectors together. This is
known as unsupervised pattern recognition or clustering. Such tasks arise in many applications in social sciences and engineering, such as remote sensing, image segmentation, and image and speech coding[1-5].

With the development of science and technology, the pattern recognition theory is gotten initial application in surveying and mapping scientific field. Some researchers have proposed methods of surveying and mapping pattern recognition[6,7]. Specially, with the rapid development of geographical information system (GIS), geography data is becoming increasingly richer and richer and contains a great deal of information. In order to exploit and use the information efficiently, it is necessary to establish corresponding theory and method. Consequently, how to use these magnanimity data, analyze, extract useful information and eliminate influence of the related factors is worth studying field. In practice, we often collect a lot of feature indexes, which usually bring us two problems. On one hand, more and more feature indexes give us difficulties to analyze, because too large data can occupy amount of memory space and computerization time. On the other hand, a lot of features include possibly many correlation factors with each other, or redundant which results to information repeat or waste. Therefore, It is necessary for us to take measures to reduce the feature dimensionality under not decreasing recognition effect, this is called the problems of feature optimal selection[8-11]. Through information collecting, information preprocessing, original features are formed, many of which may be huge, or pattern sample is in high dimensional space-measurement space ($X$). By the way of mapping, pattern sample can be represented by low dimensional space-feature space ($Y$). This process is called feature extraction, broadly speaking, is a kind of transform.

About feature extraction, many scholars had been doing a great research for this field at home and abroad. Some methods, such as correlation analysis, principal component analysis (PCA), rough sets and so on, are proposed[12-14]. In this paper, the authors study them continuously. Firstly, we make a brief introduction about principle of PCA. Secondly, applied information entropy theory, the definition of the 2nd information function and the 2nd information entropy and geometry entropy are proposed firstly. Thirdly, the analysis and application of PCA information features are carried on, and therefore the new concepts of information rate (IR) and accumulated information rate (AIR) are set up, by which the degree of information feature extraction is metricized. In the end, the simulated application is carried on in practice.

2. Principle of PCA

Algebraically, principal components (PCs) are particular linear combinations of the $n$ random variables $x_1, x_2, \cdots, x_n$. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with $x_1, x_2, \cdots, x_n$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure[15,16].

As we shall see, PCs depend solely on the covariance matrix $\Sigma$ (or the correlation matrix $R$) of $x_1, x_2, \cdots, x_n$. Their development does not require a multivariate normal assumption. On the other hand, PCs derived for multivariate normal populations have useful interpretations in terms of the constant density ellipsoids. Further, inferences can be made from the sample components when the population is multivariate normal. Let the random vector $X = (x_1, x_2, \cdots, x_n)'$ have the covariance matrix $\Sigma$
with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$.

Consider the linear combinations

$$
\begin{align*}
y_1 &= a'_1 X = a_{11} x_1 + a_{12} x_2 + \cdots + a_{1n} x_n \\
y_2 &= a'_2 X = a_{21} x_1 + a_{22} x_2 + \cdots + a_{2n} x_n \\
&\vdots \\
y_n &= a'_n X = a_{n1} x_1 + a_{n2} x_2 + \cdots + a_{nn} x_n
\end{align*}
$$

(1)

Then, we obtain

$$
Var (y_i) = a'_i \sum_j a_{ij} \quad i = 1, 2, \ldots, n \quad (2)
$$

$$
Cov (y_i, y_k) = a'_i \sum_j a_{jk} \quad i, k = 1, 2, \ldots, n \quad (3)
$$

The PCs are those uncorrelated linear combinations $y_1, y_2, \ldots, y_n$ whose variances in (2) are as large as possible.

The first PC is linear combination $a'_1 X$ that maximizes

$$
Var (a'_1 X) \text{ subject to } a'_1 a_1 = 1
$$

The second PC is linear combination $a'_2 X$ that maximizes

$$
Var (a'_2 X) \text{ subject to } a'_2 a_2 = 1 \quad \text{and}
$$

$$
Cov (a'_i X, a'_2 X) = 0.
$$

At the $i$th step, the $i$th PC is linear combination $a'_i X$ that maximizes

$$
Var (a'_i X) \text{ subject to } a'_i a_i = 1 \quad \text{and}
$$

$$
Cov (a'_i X, a'_k X) = 0 \text{ for } k < i
$$

Lemma 1 Let $\sum$ be the covariance matrix associated with the random vector $X = (x_1, x_2, \cdots, x_n)'$. Let $\sum$ have the eigenvalue-eigenvector pairs $(\lambda_1, \phi_1), (\lambda_2, \phi_2), \cdots, (\lambda_n, \phi_n)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Then the $i$th PC is given by

$$
y_i = \phi'_i X = e_{i1} x_1 + e_{i2} x_2 + \cdots + e_{in} x_n
$$

Where $i = 1, 2, \ldots, n$.

With these choices, we have

$$
Var (y_i) = \phi'_i \sum \phi_i = \lambda_i, \quad i = 1, 2, \ldots, n
$$

$$
Cov (y_i, y_k) = \phi'_i \sum \phi_k = 0, \quad i \neq k
$$

If some $\lambda_i$ are equal, the choices of the corresponding coefficient vectors, $\phi_i$, and hence $y_i$, are not unique.

According to Lemma 1, we have

$$
Y = TX = (\phi_1, \phi_2, \cdots, \phi_n)' X
$$

(4)

Therefore, we can get

$$
Var (Y) = T \sum T' = \Lambda = diag (\lambda_1, \lambda_2, \cdots, \lambda_n)
$$

(5)

That is to say, the PCA makes matrix $\sum$ be diagonal matrix. By PCA, the correlation among components of original vector $X$ is eliminated. So it is possible for us to reach the object that dimensions of characteristic space are decreased while deleting the coordinate axes, which have less information.

Lemma 2 Let $X' = (x_1, x_2, \cdots, x_n)$ have covariance matrix $\sum$, with eigenvalue eigenvector pairs $(\lambda_1, \phi_1), (\lambda_2, \phi_2), \cdots, (\lambda_n, \phi_n)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. Then

$$
\sigma_{11} + \sigma_{22} + \cdots + \sigma_{nn} = \sum_{j=1}^{n} Var (x_j)
$$
Analysis and Applications of PCA Information Features

\[ \lambda_1 + \lambda_2 + \cdots + \lambda_n = \sum_{i=1}^{n} \text{Var}(y_i) \]

According to Lemma 2, we have

Total population variance

\[ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 = \lambda_1 + \lambda_2 + \cdots + \lambda_n \]

and consequently, the proportion of total variance due to (explained by) the kth principal component is

\[ \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}, \quad k = 1, 2, \ldots, n \]  

(6)

If most (for instance, 80 to 90\%) of the total population variance, for large n, can be attributed to the first one, two, or three components, then these components can “replace” the original n variables without much loss of information.

### 3. Information Entropy

Shannon, an American electric engineer, proposed firstly the concept of information entropy in the paper of A Mathematical Theory of Communication in 1948, and regarded it as uncertainty of a stochastic event or metric of information content, and so provided scientific theory basis for modern information theory[17].

Information content is a central concept of information theory, basic point to measure information, and regarded information obtained as elimination of uncertainty. Therefore, information content may be denoted by elimination of uncertainty, while the uncertainty of a stochastic event is described by probability distribution as follows[18,19].

Suppose that there is a discrete random variable \( X \), whose \( n \) possible values are \( a_1, a_2, \ldots, a_n \), while the probability of every value is \( p_1, p_2, \ldots, p_n \). The probability space of \( X \) can be expressed as follows

\[ [X \cdot P]: \begin{cases} X: a_1, a_2, \ldots, a_n, \\ P: p_1, p_2, \ldots, p_n, \\ \sum_{i=1}^{n} p_i = 1 \end{cases} \]

(7)

Where \( P(a_i) = p_i \) is probability occurred by event \{ \( X = a_i \) \} and \( 0 \leq p_i \leq 1 \). Because \([X \cdot P]\) has described completely characteristics of discrete information source denoted by \( X \), it is called information source space of \( X \). We call

\[ I(a_i) = -\log p_i, \quad i = 1, 2, \ldots, n \]  

(8)

information function. Before the event \{ \( X = a_i \) \} happens, \( I(a_i) \) denotes uncertainty of event \{ \( X = a_i \) \}; After the event \{ \( X = a_i \) \} happens, \( I(a_i) \) denotes information content included by event \{ \( X = a_i \) \}, and also called self information of \( a_i \).

Shannon regarded statistic average value of information function in information source space \([X \cdot P]\)

\[ H(X) = H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i \]  

(9)

as measurement of uncertainty of information source \( X \), and called information entropy or Shannon entropy. The bigger \( H(X) \) is, the bigger the uncertainty of \( X \) is, and the bigger information content obtained by \( X \) is. We may prove that when all prior probabilities of discrete information source are equal.

### 4. Analysis of PCA Information Features

#### 4.1 Entropy function based on PCA

According to formula (5), we conclude that

1. every components of vector \( Y \) is
orthogonal;

② \( \lambda_i \) is variance of \( y_i \), and that is
\[
\lambda_i = \text{var}(y_i) = E[(y_i - \bar{y}_i)^2] \quad (i = 1, 2, \cdots, n)
\]
where \( \bar{y}_i = E(y_i) \). Transformation of \( \lambda_i \) is carried on, and written as follows.
\[
\rho_i = 1 - \frac{1}{\sum_{i=1}^{n} \lambda_i}
\]
Therefore, \( 0 \leq \rho_i \leq 1 \), \( \rho_i \) has numerical properties of probability. Being similar with the definition of entropy function, we define entropy function called the 2\(^{nd} \) information entropy as follows
\[
I(\lambda_i) = -\log \rho_i, i = 1, 2, \cdots, n \quad (10)
\]
\[
H(T) = H(\rho_1, \rho_2, \cdots, \rho_n) = -\sum_{i=1}^{n} \rho_i \log \rho_i \quad (11)
\]
\( H(T) \) reflects unevenness of \( \rho_i \) or \( \text{var}(y_i) \).

According to formulae (10) and (11), we can come to a conclusion that:

① \( I(\lambda_i) \) denotes information content included by \( \lambda_i \). The bigger \( \lambda_i \), the bigger \( I(\lambda_i) \) also is, and that conforms the properties of information function;

② when all are equal, that is even, the value of the second information entropy reaches maximum value \( \log n \), and at this moment, PCA components of \( X \) reserve maximum information content of \( X \);

③ when all \( \text{var}(y_i) \) are uneven, taken \( m \) eigenvectors in correspondence with \( m \) the biggest eigenvalues as coordinate system, \( H(T) \) is bigger, and PCA components of \( X \) reserve more information content of \( X \) at this moment, while taken the eigenvectors in correspondence with \( m \) the other eigenvalues as coordinate system, \( H(T) \) is smaller, and PCA components of \( X \) reserve smaller information content of \( X \).

4.2 Total information content keeping constant under of any orthogonal transforms

Suppose that another orthogonal matrix \( V = (v_1, v_2, \cdots, v_n) \), so we have orthogonal transform \( z = Vx \). Write
\[
m_i = \text{var}(z_i) = E(z_i - \bar{z}_i)^2 \quad (12)
\]

According to orthogonal basis \( \{u_j\} \), every orthogonal vectors \( v_i \) is expanded as follows
\[
v_i = \sum_{j=1}^{n} a_{ij} u_j, \quad i = 1, 2, \cdots, n \quad (13)
\]

Where coefficient is \( a_{ij} = u_j^t v_i, \quad i, j = 1, 2, \cdots n \).

Let
\[
a_i = (a_{i1}, a_{i2}, \cdots, a_{in})' \quad (14)
\]

From formula (14) and orthogonal \( v_i, v_j (i \neq j) \), we know easily that \( a_1, a_2, \cdots, a_n \) are still orthogonal vectors. So
\[
m_i = \text{var}(z_i) = v_i' \sum_x v_i
\]
\[
= (\sum_{j=1}^{n} a_{ij} \phi_j) \sum_x (\sum_{j=1}^{n} a_{ij} \phi_j)
\]
\[
= (\sum_{j=1}^{n} a_{ij} \phi_j) (\sum_{j=1}^{n} a_{ij} \phi_j)
\]
\[
= (\sum_{j=1}^{n} a_{ij} \phi_j) (\sum_{j=1}^{n} a_{ij} \lambda_j \phi_j)
\]
\[ \sum_{j=1}^{n} a_{ij}^2 \lambda_j = \sum_{j=1}^{n} a_{ij}^2 \text{var}(y_j) \quad (15) \]

Therefore, considering \( \sum_{j=1}^{n} a_{ij}^2 = 1 \), we have

\[ \sum_{i=1}^{n} m_i = \sum_{j=1}^{n} \left[ \sum_{i=1}^{n} a_{ij}^2 \text{var}(y_j) \right] \]

\[ = \sum_{j=1}^{n} \left( \sum_{i=1}^{n} a_{ij}^2 \right) \text{var}(y_j) = \sum_{j=1}^{n} \text{var}(y_j) = \sum_{j=1}^{n} \lambda_j \quad (16) \]

So from the viewpoint of information content, the formula (16) shows that under any orthogonal transform, the total information content included by a pattern keep constant.

4.3 The smallest the second information entropy by PCA

Let formula (12) be transformed as follows. Write

\[ \sigma_i = 1 - m_i \left/ \sum_{j=1}^{n} m_j \right., \quad 0 \leq \sigma_i \leq 1 \quad (17) \]

Consequently, we may also define the 2nd second information entropy as follows

\[ H(V) = H(\sigma_1, \sigma_2, \cdots, \sigma_n) = - \sum_{i=1}^{n} \sigma_i \log \sigma_i \quad (18) \]

From formulae (15) and (16), we can obtain

\[ \sigma_i = 1 - m_i \left/ \sum_{j=1}^{n} m_j \right. = 1 - \sum_{j=1}^{n} a_{ij}^2 \lambda_j \left/ \sum_{j=1}^{n} \lambda_j \right. \]

\[ = \sum_{j=1}^{n} a_{ij}^2 \lambda_j \left/ \sum_{j=1}^{n} \lambda_j \right. \]

\[ = \sum_{j=1}^{n} a_{ij}^2 (1 - \lambda_j) \left/ \sum_{j=1}^{n} \lambda_j \right. = \sum_{j=1}^{n} a_{ij}^2 \rho_j \quad (19) \]

Using Jensen inequality for lower convex function \( f(x) = x \log x \), we can get

\[ \sigma_i \log \sigma_i = (\sum_{j=1}^{n} a_{ij}^2 \rho_j) \log(\sum_{j=1}^{n} a_{ij}^2 \rho_j) \]

\[ \leq \sum_{j=1}^{n} a_{ij}^2 \rho_j \log \rho_j \quad (20) \]

According to formulae (18) and (20), we can get

\[ H(V) = - \sum_{i=1}^{n} \sigma_i \log \sigma_i \geq - \sum_{j=1}^{n} (\sum_{i=1}^{n} a_{ij}^2 \rho_j \log \rho_j) \]

\[ = - \sum_{j=1}^{n} (\sum_{i=1}^{n} a_{ij}^2) (\rho_j \log \rho_j) \]

\[ = - \sum_{j=1}^{n} \rho_j \log \rho_j = H(T) \]

Because of arbitrariness of \( V \), we have

\[ H(T) = \min_{V} \{ H(V) \} \quad (21) \]

From formula (21), we know that the value of the 2nd information entropy by PCA is the smallest, and so the nonzero variances of the components generated by PCA of \( X \) tend to be more uneven than which generated by any other orthogonal transform. That is to say, PCA makes information content concentrate on the components transformed relatively.

4.4 Tending to be definitude for vectors transformed by PCA in total

From probability and statistic theory, we know that the variance of a random variable describes its stochastic strength. The smaller the variance of random variable is, the smaller its stochastic strength is, and the bigger the definitude is. The function being equal value to the 2nd information entropy \( H(\cdot) \) is called geometry entropy and defined as follows

\[ H_T(T) = \prod_{i=1}^{n} \text{var}(y_i) \quad \text{and} \]
\[
H_R(V) = \prod_{i=1}^{n} \text{var}(z_i) \tag{22}
\]

The formula (22) denotes a volume including statistical meaning defined by variance of the components. We can also know from this formula that PCA makes \( H_R(T) \) be minimum, and the probability at unit volume is maximum. Therefore, as far as statistical meaning, PCA makes the vector transformed in total tend to be definitude.

4.5 Information feature extraction under PCA

By PCA, we can calculate the second information entropy values and order as follows

\[
I(\lambda_1) \geq I(\lambda_2) \geq \cdots \geq I(\lambda_m) \geq \cdots \geq I(\lambda_n)
\]

In order to measure the degree of information extraction, the new concepts of information rate (IR) and accumulated information rate (AIR) are proposed firstly. We define them respectively as follows

Information rate (IR):

\[
IR(\lambda_i) = I(\lambda_i) / \sum_{i=1}^{n} I(\lambda_i), \quad (i = 1, 2, \cdots, n)
\]

Accumulated information rate (AIR):

\[
AIR(\lambda_1, \lambda_2, \cdots, \lambda_n) = \sum_{i=1}^{m} I(\lambda_i) / \sum_{i=1}^{n} I(\lambda_i)
\]

For given \( m \), we select \( m \) eigenvectors in correspondence with the front \( m \) eigenvalues as K-L coordinate system. That is to say, the new features after \( x \) being compressed are

\[
y = (\phi_1, \phi_2, \cdots, \phi_m)'x = T'mx
\]

Where \( T_m \) is called \( n \times m \) extraction matrix. Therefore

\[
y_i = \phi'_ix, \quad i = 1, 2, \cdots, m; m << n
\]

Therefore, we reach the aim of information feature extraction.

5. Numerical Example

Suppose that there are data vector \( \mathbf{X} = (x_1, x_2, \cdots, x_7)' \) composed of 7 features, and the original measured data. Generally speaking, covariance \( \Sigma \) of data matrix \( \mathbf{X} \) is unknown. In surveying and mapping application, \( \Sigma \) is estimated by covariance matrix (\( S \)) of sample. In order to eliminate the influence of different unit of every feature, data matrix is standardized. In this moment, covariance matrix (\( S \)) of sample is correlation coefficient matrix (\( R \)). That is \( S = R \). By computer computation, we can get \( R \) as follows.

\[
R = \begin{pmatrix}
1.0000 & -0.2912 & -0.5705 & -0.5676 & 0.8513 & -0.3886 & 0.0000 \\
1.0000 & 0.2725 & 0.2593 & -0.3444 & 0.1386 & 0.0000 \\
1.0000 & 0.9869 & -0.7485 & 0.6695 & 0.0000 \\
1.0000 & 0.7356 & 0.7779 & -0.5855 & 0.0000 \\
1.0000 & -0.4729 & 0.2793 & 0.8875 & 0.0000 \\
1.0000 & -0.8875 & 0.2793 & 0.8875 & 0.0000 \\
1.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000
\end{pmatrix}
\]

According to \( R \), applied Jacobin method, 9 eigenvalues of \( R \) can be gotten. IR, AIR and total contribution rate (TCR) are listed in table 1.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>IR</th>
<th>AIR</th>
<th>TCR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0000</td>
<td>0.2912</td>
<td>-0.5705</td>
<td>-0.5676</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.2725</td>
<td>0.2593</td>
<td>-0.3444</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.9869</td>
<td>-0.7485</td>
<td>0.6695</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.7356</td>
<td>0.7779</td>
<td>-0.5855</td>
</tr>
<tr>
<td>1.0000</td>
<td>-0.4729</td>
<td>0.2793</td>
<td>0.8875</td>
</tr>
<tr>
<td>1.0000</td>
<td>-0.8875</td>
<td>0.2793</td>
<td>0.8875</td>
</tr>
<tr>
<td>1.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Table 1: Computation of IR, AIR and TCR
From table 1, if we take $m = 3$, then AIR=93.66%, while TCR=91.65%. That is to say, if the eigenvectors in correspondence with the front 3 eigenvalues are regarded as coordinate system, the information content contained by $Y$ reach 93.66% of the total information content of $X$.

The values of the second information entropy and geometry entropy are of difference for selecting different eigenvalues, that is to say, selecting different coordinate system. The calculated results are listed in table 2 and table 3.

**Table 2** The values of entropy function by Selecting three maximum eigenvalues

<table>
<thead>
<tr>
<th>No</th>
<th>$\lambda$</th>
<th>$\rho$</th>
<th>$H(T)$</th>
<th>$H_R(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.2444</td>
<td>0.3385</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.2513</td>
<td>0.8050</td>
<td>0.9722</td>
<td>4.8872</td>
</tr>
<tr>
<td>3</td>
<td>0.9202</td>
<td>0.8566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>6.4159</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3** The values of entropy function by Selecting the other eigenvalues

<table>
<thead>
<tr>
<th>No</th>
<th>$\lambda$</th>
<th>$\rho$</th>
<th>$H(T)$</th>
<th>$H_R(T)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.2444</td>
<td>0.1959</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.9202</td>
<td>0.8257</td>
<td>0.7196</td>
<td>0.4437</td>
</tr>
<tr>
<td>5</td>
<td>0.1136</td>
<td>0.9785</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5.2782</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From table 2 and table 3, we can conclude that the values of the second information entropy and geometry entropy by Selecting the maximum eigenvalues are bigger than the other values, so the most information content of $X$ is reserved and can be reached AIR=93.66%.

**6. Conclusions**

By discussion above, we can conclude as follows:

(1) The 2nd information entropy and geometry entropy based on PCA are minimum. This shows that PCA makes information content concentrate on the components transformed by $X$ relatively.

(2) By PCA, selecting different eigenvalues, the 2nd information entropy and geometry entropy are different. If taken $m$ eigenvectors in correspondence with the biggest eigenvalues as coordinate system, $H(T)$ and $H_R(T)$ are bigger, and the PCs of $X$ reserve more information content of $X$ at this moment, while taken the eigenvectors in correspondence with the other eigenvalues as coordinate system, they are smaller, and the PCs of $X$ reserve smaller information content of $X$.

(3) On the basis of analysis of information features of PCA, The new concepts of IR and AIR are first proposed, by which the degree of information extraction is measured. A new method is provided in practice for information feature extraction, and a new attempt is performed in this research field.

(4) Comparing with PCA, we can see that the principal components are selected according to TCR, and information content is not measured from the viewpoint of information. But information feature extraction in this paper is selected according to AIR, which is proposed firstly by authors, and can be metricized the degree of information extraction to a great extent. So we can say that the new method proposed in this paper is a further extension and perfection for PCA.

**References:**


Analysis and Applications of PCA Information Features


