Journal of Applied Analysis and Computation
Volume 5, Number 2, May 2015, 262–272

Website:http://jaac-online.com/
doi:10.11948/2015024

IMPROVING THE COMPLEXITY OF CHAOTIC SEQUENCE BASED ON THE PCA ALGORITHM

Wei Xu[1], Qun Ding[2,†] and Xiaogang Zhang[3]

Abstract The principal component analysis (PCA) is an effective statistical analysis method in statistical data analysis, feature extraction and data compression. The method simplifies multiple related variables into a linear combination of several irrelevant variables, through the less-comprehensive index as far as possible to replace many of the original data, and can reflect the information provided by the original data. This paper studies the signal feature extraction algorithm based on PCA, and extracts sequences' feature which generated by Logistic mapping. Then we measured the complexity of the reconstructed chaotic sequences by the permutation entropy algorithm. The testing results show that the complexity of the reconstruction sequences is significantly higher than the original sequences.

Keywords Chaos, complexity, permutation entropy, principal component analysis.

MSC(2000) 37M25.

1. Introduction

Chaos, as a classical complex phenomenon of nonlinear dynamic systems, has attracted widespread attention for its broadband, noise-like, and sensitive features. In recent years, with more research on chaos, chaos has replaced the traditional pseudo-random sequence in many commercial and spread-spectrum communication systems [4,10]. The encryption method of a password system based on chaos is simple, fast, easy to be realized. The relationship between cryptograph, plaintext and key is very complicated and any change of plaintext or key will cause great changes in encrypted files, which makes an encryption system has higher security [16].

Chaos of the nonlinear dynamic systems has the geometric and statistical features that deterministic movements usually do not have, such as local instability and overall stability, strange attractor, continuous power spectrum, positive Lyapunov index, fractal dimension, positive measure entropy, and so on. To sum up, chaos has the following three main qualitative characteristics [17]: Inherent randomness, Fractal dimension characteristics, Universality.

The complexity of the data sequences is not only a similarity degree of measurement between chaotic pseudo-random sequence and random sequence, but also a

[†] the corresponding author. Email address:qunding@aliyun.com(Q. Ding)
[1] School of Computer Science and Technology, Heilongjiang University, Harbin 150080, China
[2] School of Electronic Engineering, Heilongjiang University, Harbin 150080, China
[3] Education Office of Heilongjiang Province, Harbin 150001, China

complexity degree of measurement by using part of the sequence to recovery the whole. The bigger complexity of a sequence is, the smaller the possibility of recovery is. Therefore, the complexity of a sequence is an important index to quantify the performance of chaotic sequences. It is important to study the chaotic complexity.

The researches of complexity have received attention worldwide. Kolmogrov [7] defined a measure entropy and used it to measure the disordered degree of system movements. And then Lempel et al. [8] realized the measure entropy method by computer. Pincus [11] proposed the definition of approximate entropy through measuring the complexity of time series, and then Bandt et al. [1] proposed a permutation entropy for measuring time series. Xiao et al. [13] proposed a symbolic dynamics approach for the complexity analysis of chaotic pseudo-random sequences. Next, Larrondo et al. [9] proposed an intensive statistical complexity measure to quantify the performance of chaotic pseudorandom number generators. Chen et al. [2] proposed a new complexity metric to evaluate the unpredictability of the chaotic pseudorandom sequences based on the fuzzy entropy.

Kolmogorov-Sinai entropy proposed can measure the complexity of chaotic systems, but it needs a lot of sample space and heavy computation. The approximate entropy is a method of quantizing the complexity of time series based on edge probability distribution statistics. It can accurately calculate the complexity of the sequences, but the result is influenced by different parameters. The symbolic dynamics approach can reduce the degree of dependence on the parameters, but before we measure the complexity, we must get the size of the symbol space of the initial sequences, which is very difficult to obtain without priori knowledge in practice. Permutation entropy is an appropriate complexity measure for chaotic time series, in particular in the presence of dynamical and observational noise, since the method is extremely fast. It seems preferable when there are huge data sets and there is no time for preprocessing and fine-tuning of parameters.

In order to get higher complexity of chaos random sequences, using principal component analysis algorithm in feature extraction of Logistic chaotic map sequences in this paper, we eliminate the signal correlation between each component to improve the complexity of the sequences. We draw conclusion after calculating the sequence complexities of the original testing sequences and reconstructing sequences by permutation entropy algorithm.

2. Principal component analysis

Principal component analysis (PCA) was first proposed by Pearson in 1901. Then, a large number of papers performed thorough research to it which make its theory gradually perfect [6]. PCA method is a kind of commonly used linear mapping method in pattern recognition. It is based on the data signal analysis of the second-order statistical characteristic. The method simplifies the multiple related variables into a linear combination of several irrelevant variables, under the principle of minimum data information loss, by a linear transformation using abandon part of information, with a few new variables instead of multidimensional variables, so as to realize the high-dimensional variable space mapping to a low-dimensional space [3, 12, 15].

At present, there are two common methods on the choice of the number of principal components [5]: one is the principal component regression method, another is a principal component contribution rate cumulative percentage.

The PCA algorithm is described as follows:

Set $X = [x_1,\ x_2,\ \ldots\ ,\ x_p]^T$ which is a p-dimensional random vector. If the mean of X is 0, and the covariance matrix is C, the PCA method is to put the p random variables integrated into m new variables $y_1,\ y_2, ..., y_m$, i.e.

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p, \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_m = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mp}x_p. \end{cases} \tag{2.1}$$

Denote it by $Y = AX$, where

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mp} \end{bmatrix} = \begin{bmatrix} a'_1 \\ a'_2 \\ \cdots \\ a'_m \end{bmatrix}. \tag{2.2}$$

The coefficient's selection principles of transform matrix A are as follows:

1. The y_1 and y_2 ($i \neq j, i, j = 1, 2, ...$) are unrelated;

2. y_1 is the largest variance in the linear combination of $x_1,\ x_2, ..., x_p$, even the a'_1X has the maximum variance; y_2 has no relation to y_1, and is the second largest variance in the linear combination of $x_1,\ x_2, ..., x_p$; $\ldots$, y_m is the mth largest variance in the linear combination of $x_1,\ x_2, ..., x_p$, and has no relation to the others. The y_m is called the mth principal component of the original random variables $x_1, x_2, ..., x_p$.

According to the above, we get the new variable y_1 by transforming the random variable X,

$$\begin{aligned} Var(y_1) &= Var(a'_1X) \\ &= E(a'_1X - Ea'_1X)(a'_1X - Ea'_1X)' \\ &= a'_1E(X - EX)(X - EX)'a_1 \\ &= a'_1Ca_1, \end{aligned} \tag{2.3}$$

where C is the covariance matrix of X.

The method for solving the first principal component y_1 is seeking the vector a_1 under the condition of $a'_1a_1 = 1$, such that

$$Var(y_1) = Var(a'_1X) = a'_1Ca_1 \tag{2.4}$$

is maximum. For the random vector X, set its covariance matrix is C be

$$\begin{aligned} C_x &= E\{(x_i - \mu_x)(x_i - \mu_x)^T\} \\ &= \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu_x)(x_i - \mu_x)^T]. \end{aligned} \tag{2.5}$$

We use c_{ij} to denote the component of C which represents the covariance of x_i and x_j. The changes of components are the deviation degree to its average. If the two components x_i and x_j are uncorrelated, then their covariance is zero, namely $c_{ij} = c_{ji} = 0$.

We suppose the p characteristic values of C are λ_1,λ_2,$\ldots$,λ_p and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ (because C is nonnegative), and its corresponding orthonormal characteristic vector are u_1,u_2, $\ldots$,u_p. We can see that C is a symmetric matrix according to the nature of the covariance matrix. We can find its orthogonal basis by calculating its eigenvalues and eigenvectors. So, there is an orthogonal array $U = (u_1,\ u_2,\ \ldots\ u_p)$, where $u_i = (u_{1i},\ u_{2i},\ldots,\ u_{pi})^T$,

$$C = U \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix} U^T = \sum_{i=1}^{p} \lambda_i u_i u_i^T. \tag{2.6}$$

If the vector α is the unit vector of p dimension, namely α= $(\alpha_{11},\ \alpha_{12},\ \ldots,\ \alpha_{1p})^T$, then

$$\begin{aligned} \alpha^T C\alpha &= \sum_{i=1}^{p} \lambda_i \alpha^T u_i u_i^T \alpha \\ &= \sum_{i=1}^{p} \lambda_i (\alpha^T u_i)^2. \end{aligned} \tag{2.7}$$

Because λ_1 is the maximum eigenvalue of C , so

$$\begin{aligned} \alpha^T C\alpha &\leq \lambda_1 \sum_{i=1}^{p} (\alpha^T u_i)^2 \\ &= \lambda_1 \alpha^T U U^T \alpha \\ &= \lambda_1 \end{aligned} \tag{2.8}$$

i.e.

$$Var(a_1' X) = a_1' C a_1 \leq \lambda_1. \tag{2.9}$$

When the vector $\alpha = u_i$,

$$\begin{aligned} u_1^T C u_1 &= u_1^T (\sum_{i=1}^{p} \lambda_1 u_i u_i^T) u_1 \\ &= \lambda_1 (u_1^T u_1)^2 \\ &= \lambda_1. \end{aligned} \tag{2.10}$$

As a result, we select vector α as the corresponding orthonormal characteristic vector u_1 of the largest eigenvalue λ_1 of C, which makes the variance of $\alpha^T X = u_1^T X$ be maximum. The maximum value is λ_1, thus

$$y_1 = u_1 X = (u_{11}, u_{21}, \ldots, u_{p1}) \begin{pmatrix} x_1 \\ \cdots \\ x_p \end{pmatrix}. \tag{2.11}$$

So, y_1 is the first principal component for the random vector X. Similarly, the other principal components can be calculated in turn, and calculation steps are shown as follows: for $i = 2, 3, \ldots, p$,

$$\begin{aligned} Var(u_i^T X) &= u_i^T C u_i \\ &= \sum_{i=1}^{p} u_i^T \lambda_j u_j u_j^T u_i \\ &= \lambda_i (u_i^T u_i)^2 \\ &= \lambda_i. \end{aligned} \tag{2.12}$$

When $i \neq j$,

$$\begin{aligned} Cov(u_i^T X, u_j^T X) &= u_i^T C u_j \\ &= u_i^T (\sum_{k=1}^{p} \lambda_k u_k u_k^T) u_j \\ &= \sum_{k=1}^{p} \lambda_k (u_i^T u_k)(u_k^T u_j) \\ &= 0. \end{aligned} \tag{2.13}$$

We know that $u_2^T X$ is irrelevant to $u_1^T X$ and is biggest variance in all linear combinations of x_1, x_2, $\ldots$, x_p since $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Similarly, $u_m^T X$ is irrelevant to $u_1^T X$, $u_2^T X$,$\ldots$, $u_{m-1}^T X$ and is the biggest variance in all linear combinations of x_1, x_2, $\ldots$, x_p. We set

$$\begin{aligned} y_2 &= u_2 X, \\ y_3 &= u_3 X, \\ &\cdots \\ y_p &= u_p X. \end{aligned} \tag{2.14}$$

We call y_i the ith principal component, and $y_1, y_2, ..., y_p$ are unrelated each other. In practice, we often not take all principal components, but only take the first m ($m < p$) principal components. The selection of m often is according to the contribution rate which we had set. The contribution rate is given by

$$g = \frac{\sum\limits_{i=1}^{m} \lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}. \tag{2.15}$$

Here, g is the contribution rate, generally 0.9 or higher.

As a result, the original p variables are converted to m, which can be accurate enough to describe the original information, where the loss of information is not much, but we can get rid of overlapping information.

The μ is the average of raw data X. If M is a new matrix containing the eigenvalues of the covariance matrix of the raw data, and the mean of data and the covariance matrix had been calculated, then y_i can be obtained by

$$y_i = M(x_i - \mu_x). \tag{2.16}$$

This y_i is a point in the orthogonal system, defined by the feature vector. The component y can be seen as a shift of orthogonal basis. According to the properties of the orthogonal matrix, that is $A^{-1} = A^T$, the source data x can be reconstruct by y:

$$x_i = M^T(y_i + \mu_x). \tag{2.17}$$

Thus, the original vector x in another axis can be solved by the orthogonal basis.

3. The PE algorithm

PE algorithm is based on the complexity of Kolmogorov, by using the concept of information entropy, to calculate the complexity of data sequences. The algorithm uses multidimensional reconstitution space similarity to measure the complexity of the entire sequences, and analyze all the similar characteristics of the embedded dimension [1].

For chaotic sequences, the algorithm is described as follows:

(1) Given a discrete-time sequence of length n, $\{x_i, i = 1, 2, ..., N\}$, generated by the system equations iteratedly, reconstruct a phase space to $\{x_i\}$, and get the sequence of reconstruction:

$$X(i) = [x(i), x(i+\tau), ..., x(i+(p-1)\tau)], \qquad 1 \le i \le N-p+1, \tag{3.1}$$

where p and τ are embedding dimension and delay time, respectively. Using maximum overlapping situation here, set $\tau = 1$, that is, move back a data point to each subsequence to get a child sequence.

(2) The pth reconstructed components $[x(i), x(i+\tau), ..., x(i+(p-1)\tau)]$ of $X(i)$ are arranged in ascending order, giving the relation:

$$[x(i+j_1-1)\tau) \le x(i+j_2-1)\tau) \le ... \le x(i+j_p-1)\tau)], \quad 1 \le j \le N-p+1. \tag{3.2}$$

If any two values $x(i)$ of the sequence are equal, according to the size of the j values to sort, we can get a set of symbol sequences by an arbitrary vector X_i:

$$A(g) = [j_1, j_2, ...j_p], \qquad 1 \le g \le N-p+1. \tag{3.3}$$

(3) There are $p!$ arrangements for p different symbols, that is, a total of $p!$ different symbol sequences, in which symbol sequence $A(g)$ is one of them. To arrange all of the same symbol sequence $A(g)$ as a group, there are a total of k group different symbol sequences in the $N-p+1$ group sequences. The number of each group of sequences ar $Num_1, Num_2, ..., Num_k$ respectively. Calculate the probability P_1, P_2,...,P_k of each symbol sequence:

$$P_k = \frac{Num_k}{N-p+1}. \tag{3.4}$$

(4) According to the form of Shannon entropy, the PE of k different symbol sequences for the time sequence $\{x_i\}, i = 1, 2, ..., n$, can be defined as:

$$H(p) = -\sum_{i=1}^{k} P_k \ln P_k. \tag{3.5}$$

(5) Theoretically, when $P_k = 1/p!$, $H(p)$ reaches a maximum value of $\ln(p!)$. Actually $H(p) \le \ln(N-p+1)$ according to the literature [8]. For convenience, standardizing $H(p)$ by using $\ln(N-p+1)$:

$$0 \le h(p) = \frac{H(p)}{\ln(N-p+1)} \le 1. \tag{3.6}$$

The general steps for calculating the PE of a chaotic pseudo-random sequence are similar to the steps mentioned above. The difference is that the chaotic sequence should be quantized into a chaotic pseudo-random sequence in the first step (1). Then, reconstruct the chaotic pseudo-random sequence. In step (2), we don't need to sort to the reconstructed sequence, since the size of the chaotic pseudo-random sequence itself has a certain relationship, so just need to calculate the number of the same sequence numbers Num_k, and then go directly to step (3).

Obviously, the changes of $H(p)$ reflects the randomicity of the original sequence. The value of $H(p)$ is smaller means the complexity of the sequences is smaller, and it works the other way as well. Ref. [13,14] discussed the validity of the calculation for length N of a sequence and the value of p. If the value of N is too small, it would lose its statistical significance. Generally, $1000 \leq N \leq 10000$; $3 \leq p \leq 15$. Compared with other algorithms, the PE algorithm has many good characteristics such as clear conception, quick calculation, etc.

4. Generated chaotic sequences

4.1. The Logistic map

The Logistic map is given by:

$$x_{n+1} = \mu x_n(1 - x_n), 0 \leq x \leq 1, 0 < \mu \leq 4. \tag{4.1}$$

Figure 1 displays the sequence diagram of the Logistic map, where iteration and initial values respectively are 1024 and 0.3. Figure 2 displays the Lyapunov exponent' trend of the Logistic map with the change of parameter μ.

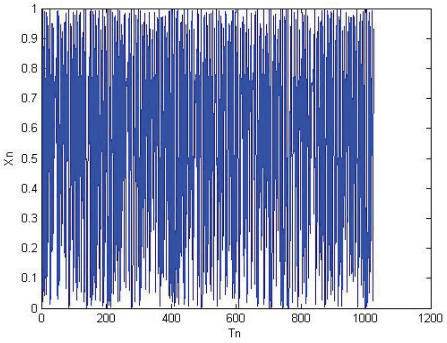

Figure 1. Sequence diagram of the Logistic map

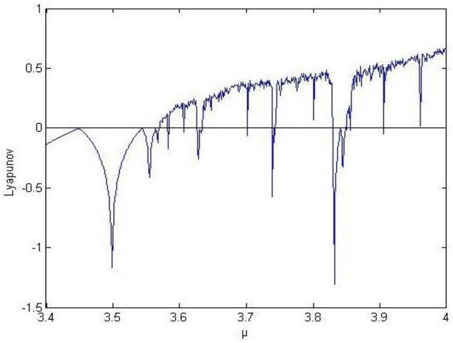

Figure 2. Lyapunov exponent of the Logistic map as a function of parameter μ

We got a sequence with length of 3072 generated by the Logistic map. The sequence is transformed into a 64 × 48 matrix through phase space reconstruction. Then, we extracted features to the matrix, where the selection of contribution rate is more than 90% for the first 35 principal components, according to each component contribution rate. After that, we reconstructed the sequences. The diagram of the first 35 principal components contribution rates and the sequence chart of reconstruct sequence are shown in Figure 3, Figure 4, respectively. We can see from Figure 4, that the refactoring sequence extracted from the original chaotic sequence retains the sensitivity and uncertainty of the original sequence.

Next, we test the complexity of the two sequences by using the PE algorithm. The complexities of the two kinds of sequence test results are shown in Figure 5.

We can see from Figure 5, that the complexity of the reconstructed chaotic sequence is significantly higher than the original. In the same way, we take the sequences with lengths of 5000 and 10000 points respectively to test the complexity

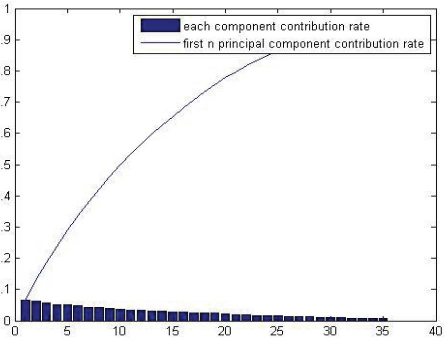

Figure 3. The first 35 principal component contribution rate

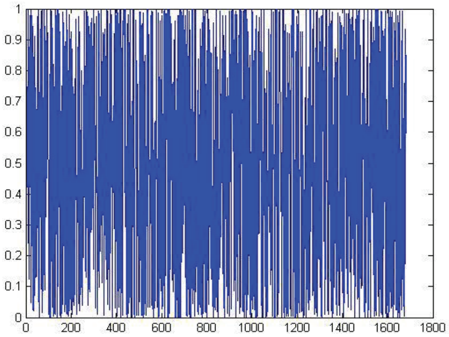

Figure 4. Sequence diagram of the refactoring sequence

of the original sequence and the sequence of reconstruction. The test results are shown in Figure 6 and Figure 7, respectively.

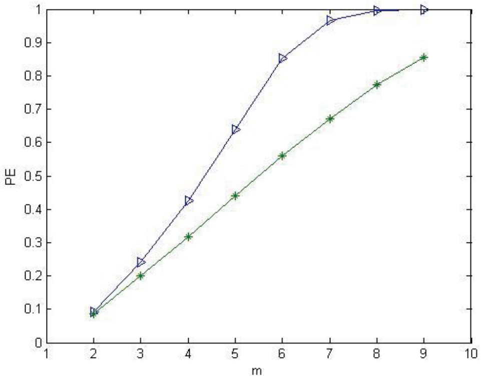

Figure 5. The complexity of the two kinds of sequences

We can see from Figure 5, that the complexity of the reconstructed chaotic sequence is significantly higher than the original. In the same way, we take the sequences with lengths of 5000 and 10000 points respectively to test the complexity of the original sequence and the sequence of reconstruction. The test results are shown in Figure 6 and Figure 7, respectively.

We can see from Figure 6 and Figure 7 that the complexity of the reconstructed sequence is higher than that of the original chaotic sequence, namely the randomness of the reconstructed sequence is stronger. It is more suitable to apply chaotic sequences in information security and secure communications. The main reason is that the PCA converts multiple variables into a few variables based on statistical analysis of data. In the study of multivariate data, due to the number of many variables, which have a certain relationship each other, there exists information overlap to a certain extent. PCA is through a dimension reduction method to simplify the data from several variables to find fewer variables, these new variables as maximum as possible to reflect the information of the original variables, and the new variables are unrelated to each other.

We finally tested the correlation of the original sequence and the reconstructed

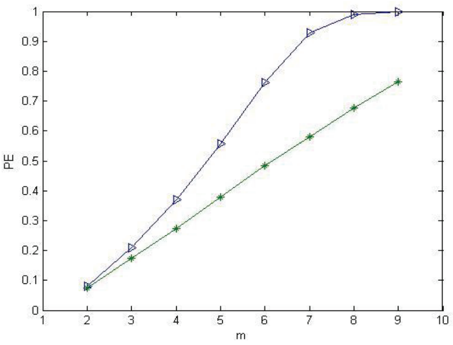

Figure 6. The complexity of the two kinds of sequences (5000)

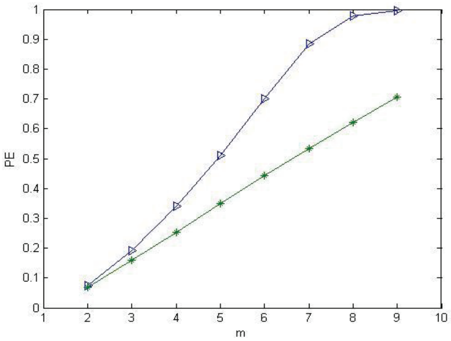

Figure 7. The complexity of the two kinds of sequences (10000)

sequence, with the test results shown in Figure 8 and Figure 9.

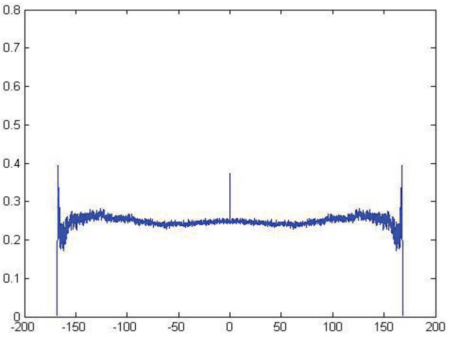

Figure 8. The correlation test of the original sequence

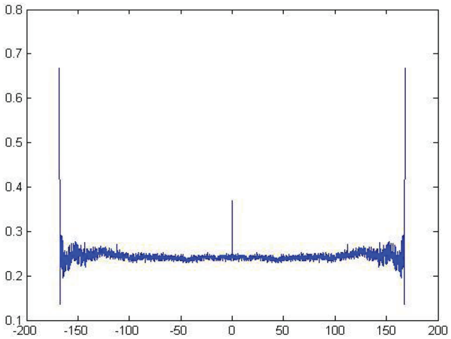

Figure 9. The correlation test of the reconstructed sequence

We can see from the graphs that the correlation of reconstructed sequences is better than the original sequence. It is mainly due to the following three aspects of the role of PCA for dimension reduction:

(1) After dimension reduction, all of the principal components become orthogonal to each other, which form an orthogonal basis of the data space, and therefore has no redundant information;

(2) The order of the principal component is according to the variances in descending order, which is always the main component of the maximum variance first appeared in the resulting sequence;

(3) Small variance contributions in the component of the principal components were deleted, usually by ignoring components mainly containing noise. The data dimension reduction is the most important, benefiting noise reduction. Usually, the first several principal components' variances will be more than the sum of the original data with 90% of the total variance, so the PCA can reduce the dimensions of the variables without significant loss of information. After PCA dimension reduction, multidimensional data will transform from a high-dimension space to a low-dimension space, which not only reduce the amount of calculation, but also reveal the original data features.

5. Conclusion

PCA is an effective method in the statistical analysis of data, and it is widely used in signal processing and neural network calculation. It is the best transformation in data compression in the sense of minimum mean square error, decreasing relevance and highlighting the role of difference. This paper applies a feature extraction algorithm based on PCA to the classic Logistic chaotic map for signal feature extraction, and apply the PE algorithm to test the complexity of the data sequences. Through simulation experiment, we can see that by using the PCA to reconstruct the sequences, the complexity has been greatly improved. This shows that the refactored sequence has better randomness, therefore is more suitable for the application of chaotic sequences in information security and secure communications.

Acknowledgments

This paper is supported by Science and Technology Research Projects of Heilongjiang Province Education Department (No.12541647). The Youth Science Foundation of Heilongjiang University (No.QL201305). The Specialized Research Fund for the Doctoral Program of Higher Education, Ministry of Education (No.20132301-110004). Innovated Team Project of "Modern Sensing Technology" in Colleges and Universities of Heilongjiang Province (No.2012TD007).

References

[1] C. Bandt and B. Pompe, *Permulation entropy a natural complexity measure for time series*, Physical Review Letters, 88(17)2002, 174102-1-174102-4.

[2] X. Chen, Z. Li and B. Bai, *A new complexity metric of chaotic pseudorandom sequences based on fuzzy entropy*, Journal of Electronics & Information Technology, 33(5)(2011), 1198-1203.

[3] A. Chen, *Identify some feature extraction method in face recognition research*, [Ph.D. Thesis]. Nanjing: Nanjing University of Science and Technology, 2006.

[4] L. Huang and Q. Yin, *A chaos synchronization secure communication system based on output control*, Journal of Electronics & Information Technology, 31(10)(2009), 2402-2405.

[5] D. Hu, Z. Zhao and Y. Zheng, *Based on principal component analysis of sensor fault detection and diagnosis,* Instrumentation Technology, (6)2005, 30-32.

[6] J.E. Jackson, *A User's Guide To Principal Components*, New York: John Wiley, 1991.

[7] A N. Kolmogrov, *Three approaches to the quantitative definition of information,* Problem in Information Transmission, 1(1)(1965), 1-7.

[8] A. Lempel and J. Ziv, *On the complexity of finite sequences*, IEEE Transactions on Information Theory, 22(1)(1976), 75-81.

[9] H.A. Larrondo, C.M. Gonzalez, M.T. Martin, et al., *Intensive statistical complexity measure of pseudorandom number generators*, Physica A, 356(2005), 133-138.

[10] U. Parlitz and S. Ergezinger, *Robust communication based chaotic spreading sequences*, Physics Letters A, 188(2)(1994), 146-150.

[11] S.M. Pincus, *Approximate entropy as a measure of system complexity*, Proc Natl Acad Sci, 88(1991), 2297-2301.

[12] R. Ren and H. Wang, *Multivariate Statistical Data Analysis Theory, Method, Instance,* Beijing: National Defence Industry Press, 1997.

[13] F. Xiao, G. Yan and Y. Han, *A symbolic dynamics approach for the complexity analysis of chaotic pseudorandom sequences,* Acta Physica Sinica, 53(9)(2004), 2877-2880.

[14] W. Xu, Q. Ding and X. Zhang, *Detection complexity of chaotic sequence*, Information Technology Journal, 12(20)(2013), 5487-5491.

[15] S. Yang, D. Wu and H. Su, *Control chart is out of control mode based on PCA and SVM intelligent identification method*, Journal of System Simulation, 17(5)2006, 1314-1318.

[16] Y. Zheng, J. Pan, Y. Song, H. Cheng and Q. Ding, *Research on the quantifications of chaotic random number generator*, International Journal of Sensor Networks, 15(1)(2014), 139-143.

[17] G. Zhao and J. Fang, *Modern information safety and advances in application research of chaos-based security communication*, Progress in Physics, 23(2)(2003), 212-252.