БИОИНФОРМАТИКА И СИСТЕМНАЯ БИОЛОГИЯ
Оригинальное исследование / Original article
УДК 519.246.87:517.968.233

Вавиловский журнал генетики и селекции. 2019;23(8):1032-1036
DOI 10.18699/VJ19.584

Principal component analysis and its generalizations for any type of sequence (PCA-Seq)

V.M. Efimov[1, 2, 3, 4] ✉, K.V. Efimov[5], V.Y. Kovaleva[2]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia
[3] Novosibirsk State University, Novosibirsk, Russia
[4] Tomsk State University, Tomsk, Russia
[5] Moscow Institute of Physics and Technology (State University), Moscow, Russia
✉ e-mail: efimov@bionet.nsc.ru

In the 1940s, Karhunen and Loève proposed a method for processing a one-dimensional numeric time series by converting it into multidimensional by shifts. In fact, a one-dimensional number series was decomposed into several orthogonal time series. This method has many times been independently developed and applied in practice under various names (EOF, SSA, Caterpillar, etc.). Nowadays, the name 'SSA' (Singular Spectral Analysis) is the most often used. It turned out that it is universal, applicable to any time series without requiring stationary assumptions, automatically decomposes time series into a trend, cyclic components and noise. By the beginning of the 1980s, Takens had shown that for a dynamical system such a method makes it possible to obtain an attractor from observing only one of these variables, thereby bringing the method to a powerful theoretical basis. In the same years, the practical benefits of phase portraits became clear. In particular, it was used in the analysis and forecast of animal abundance dynamics. In this paper we propose to extend SSA to a one-dimensional sequence of any type of elements, including numbers, symbols, figures, etc., and, as a special case, to a molecular sequence. Technically, the problem is solved using an algorithm like SSA. The sequence is cut by a sliding window into fragments of a given length. Between all fragments, the matrix of Euclidean distances is calculated. This is always possible. For example, the square root of the Hamming distance between fragments is a Euclidean distance. For the resulting matrix, the principal components are calculated by the principal-coordinate method (PCo). Instead of a distance matrix, one can use a matrix of any similarity/dissimilarity indexes and apply methods of multidimensional scaling (MDS). The result will always be PCs in some Euclidean space. We called this method 'PCA-Seq'. It is certainly an exploratory method, as is its particular case SSA. For any sequence, including molecular, PCA-Seq without any additional assumptions allows presenting its principal components in a numerical form and visualizing them in the form of phase portraits. A long history of SSA application for numerical data gives all reason to believe that PCA-Seq will be not less useful in the analysis of non-numerical data, especially in hypothesizing. PCA-Seq is implemented in the freely distributed Jacobi 4 package (http://jacobi4.ru/).
Key words: time series; SVD; PCA; PCo; MDS; SSA; molecular sequences; *p*-distance.

For citation: Efimov V.M., Efimov K.V., Kovaleva V.Y. Principal component analysis and its generalizations for any type of sequence (PCA-Seq). Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2019;23(8):1032-1036. DOI 10.18699/VJ19.584

Метод главных компонент и его обобщения для последовательности любого типа (PCA-Seq)

В.М. Ефимов[1, 2, 3, 4] ✉, К.В. Ефимов[5], В.Ю. Ковалева[2]

[1] Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, Россия
[2] Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, Россия
[3] Новосибирский государственный университет, Новосибирск, Россия
[4] Томский государственный университет, Томск, Россия
[5] Московский физико-технический институт (государственный университет), Москва, Россия
✉ e-mail: efimov@bionet.nsc.ru

В 1940-х гг. К. Карунен и М. Лоев предложили метод обработки одномерного числового временного ряда через его преобразование в многомерный путем сдвига несколько раз подряд и разложения на несколько ортогональных временных рядов методом главных компонент (PCA). Предложенный метод ранее независимо возникал и применялся на практике под разными названиями (EOF, SSA, Гусеница и т. д.). Оказалось, что он универсальный, применим к любому временному ряду и, не требуя предположения стационарности, автоматически разлагает его на тренд, циклические составляющие и шум. В наши дни чаще всего используется название SSA (сингулярный спектральный анализ). В начале 1980-х гг. Ф. Такенс показал, что для динамической системы сдвиги только одной наблюдаемой переменной позволяют построить аттрактор всей системы, и тем самым подвел под SSA мощную теоретическую базу. Тогда же выяснилась практическая польза фазовых портретов, что было применено, в частности, при анализе и прогнозе динамики численности животных. В настоящей работе предлагается распространить SSA на одномерную последовательность элементов любого типа, включая числа, символы, фигуры и т. д., и в качестве частного случая –

на молекулярную последовательность. Технически проблема решается практически тем же алгоритмом, что и SSA. Последовательность режется скользящим окном на фрагменты заданной длины. Между всеми фрагментами вычисляется матрица евклидовых расстояний. Это всегда возможно. Например, квадратный корень из *p*-дистанции (дистанции Хэмминга) является евклидовым расстоянием. Для полученной матрицы методом главных координат (PCo) вычисляются главные компоненты. Вместо расстояний можно использовать любые индексы сходства/различия и применить методы многомерного шкалирования (MDS). В итоге все равно будут получены главные компоненты в некотором евклидовом пространстве. Мы назвали этот метод PCA-Seq. Это, безусловно, разведочный метод, как и его частный случай SSA. Для любой последовательности, в том числе молекулярной, PCA-Seq без всяких дополнительных предположений позволяет получить ее главные компоненты в числовом виде и визуализировать их в виде графиков и фазовых портретов. Многолетний опыт применения SSA для числовых данных дает все основания полагать, что PCA-Seq окажется не менее полезным при анализе нечисловых данных, особенно при выдвижении гипотез. PCA-Seq реализован в свободно распространяемом пакете Jacobi 4 (http://jacobi4.ru/).
Ключевые слова: временные ряды; SVD; PCA; PCo; MDS; SSA; молекулярные последовательности; *p*-дистанция.

Introduction

In the 1940s, Karhunen and Loève proposed a method for processing a one-dimensional numerical time series by shifting it several times and decomposing into several orthogonal time series by a multidimensional method of principal components (PCA) (Karhunen, 1947; Loève, 1948). In the 1980s, Takens showed for a dynamic system, that shifts of only one observed variable allow constructing an attractor of the entire system, thereby bringing the method to a powerful theoretical basis (Takens, 1981).

The method was independently developed and applied in practice under various names (EOF, SSA, Caterpillar, etc.), including by us for the analysis of animals abundance dynamics (Efimov, Galaktionov, 1983; Efimov et al., 1988, 2003), and for other topics (Golyandina et al., 2001, 2018; Golyandina, Zhigljavsky, 2013). Today the name 'SSA' (Singular Spectral Analysis) is the most often used. The method can be extended for a sequence of any type of elements, including numbers, symbols, figures, etc. and, as a special case, for a molecular sequence (Efimov et al., 2018). This is the point of this article.

Material and methods

Algorithm. Let there be a sequence $Q = \{q_1, q_2, \ldots, q_N\}$ of any type of elements. Choose a lag L, $N > L > 1$. Denote by Q_i the fragment Q of length L terminated by the element q_i, $Q_i = (q_{i-L+1}, q_{i-L+2}, \ldots, q_{i-1}, q_i)$, $i = L, \ldots, N$. Compute the matrix of Euclidean distances $D = (d_{ij} = d(Q_i, Q_j))$ between all fragments (this is always possible, for example, using the number of unmatched elements, but not only that). Apply the method of principal coordinates (PCo) to D and obtain its principal components PCs (Gower, 1966). Call this method 'PCA-Seq'.

The usual method of finding principal components consists in the following (Jolliffe, Cadima, 2016). Let X be a centered matrix of objects' coordinates in a certain Euclidean space. We can apply to X the singular value decomposition (SVD): $X = PSV^T$, where P, V^T are orthogonal matrices, and S is the diagonal matrix of X singular values. It is possible to apply SVD to a symmetric matrix XX^T: $XX^T = P\Lambda P^T$, where P is the same orthogonal matrix as for X, and Λ is a diagonal matrix of the matrix XX^T singular values. But $XX^T = PSV^T \times VSP^T = PSSP^T = PS^2P^T$. Consequently, $S^2 = \Lambda$ and $S = \Lambda^{1/2}$. That is, the matrix of singular values of the matrix XX^T is the matrix of eigenvalues of the matrix X. Therefore, it is necessary to calculate the principal components by the formula $U = P\Lambda^{1/2}$. This is very useful in practice if the number of objects is significantly lower than the number of traits that are becoming more common in biological research, especially molecular ones.

More than half a century ago, Gower (Gower, 1966) found that if we calculate the matrix D of Euclidean distances between rows X, square these distances, double center and multiply them by –1/2, then we obtain the XX^T matrix. Applying SVD to it, we obtain principal components. For this reason, Gower called this method the 'principal coordinates (PCo) analysis'. However, it follows from the results of Gower that the matrix X itself is not needed and may not even exist in numerical form. To calculate the principal components of a certain set of objects, it is enough to have a matrix of Euclidean distances between them obtained no matter which way. If we calculate the Euclidean matrix of distances between the rows of the matrix of principal components, then it will coincide with the initial matrix of Euclidean distances D. This property can be used to verify the calculations.

PCo is quite often used for dissimilarity matrices, for which it is unknown whether they are Euclidean distances between objects or not. In the case of non-Euclidean distances, some diagonal values of the matrix Λ will be negative. Small negative diagonal values can sometimes arise due to the accumulation of computational errors. All such "components", as well as zero ones, should be excluded from consideration.

Instead of a distance matrix, one can use a matrix of any coefficients of similarity/dissimilarity. In this case, it is necessary to apply methods of multidimensional scaling (MDS). The results will always be the PCs in some Euclidean space (Gower, 1966). PCo has another name: metric multidimensional scaling abbreviated as MMDS or simply MDS. It is more correct to call all the multidimensional scaling methods 'MDS', and apply 'MMDS' to PCo only.

Data. The amino acid sequence of the *Homo sapiens Cytb* gene was used (Q0ZFD6_HUMAN, Swiss Model repository) (Table 1). The sequence (length N = 380) contains two chains at positions 19−204, 259−359 and nine transmembrane helices at positions 30−56, 77−98, 113−133, 140−158, 178−200, 229−246, 288−308, 320−339, and 345−372 (https://swissmodel.expasy.org/repository/uniprot/Q0ZFD6).

Processing. Denote $N-L+1$ by N_L. For L = 2, … , 24, the sequence Q0ZFD6 represented as Seq_L matrix of size $N_L \times L$ (Table 2 with L = 8, as an example). For each Seq_L, the matrix A_L of size $N_L \times 20$ – each amino acid content in the fragment and the H_L vector of length N_L – the fraction of fragment positions coinciding with transmembrane helices are additionally calculated. For all matrices Seq_L, matrices

Table 1. The amino acid sequence of the *Homo sapiens Cytb* gene
(Q0ZFD6_HUMAN, Swiss Model repository, https://swissmodel.expasy.org/repository/uniprot/Q0ZFD6)

>tr|Q0ZFD6|Q0ZFD6_HUMAN Cytochrome b OS=Homo sapiens GN=CYB

MTPMRKTNPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDASTAFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGS
FLYSETWNIGIILLLATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVQWIWGGYSVDSPTLTRFFTFHFILPFIIAALATLHLLFLHETGSNNPLGITSH
SDKITFHPYYTIKDALGLLLFLLSLMTLTLFSPDLLGDPDNYTLANPLNTPPHIKPEWYFLFAYTILRSVPNKLGGVLALLLSILILAMIPILHMSKQQSMMFRPLSQ
SLYWLLAADLLILTWIGGQPVSYPFTIIGQVASVLYFTTILILMPTISLI**ENKMLKWA**

Note: The top line is the sequence identifier, the lower line is the sequence itself. The first and last fragments of length 8 are highlighted in bold type (see Table 2).

Table 2. Takens embedding transformation
of the amino acid sequence from Table 1

i	q_{i-7}	q_{i-6}	q_{i-5}	q_{i-4}	q_{i-3}	q_{i-2}	q_{i-1}	q_i
8	**M**	**T**	**P**	**M**	**R**	**K**	**T**	**N**
9	T	P	M	R	K	T	N	P
10	P	M	R	K	T	N	P	L
11	M	R	K	T	N	P	L	M
12	R	K	T	N	P	L	M	K
13	K	T	N	P	L	M	K	L
14	T	N	P	L	M	K	L	I
15	N	P	L	M	K	L	I	N
16	P	L	M	K	L	I	N	H
17	L	M	K	L	I	N	H	S
18	M	K	L	I	N	H	S	F
19	K	L	I	N	H	S	F	I
20	L	I	N	H	S	F	I	D
21	I	N	H	S	F	I	D	L
22	N	H	S	F	I	D	L	P
23	H	S	F	I	D	L	P	T
24	S	F	I	D	L	P	T	P
25	F	I	D	L	P	T	P	S
26	I	D	L	P	T	P	S	N
27	D	L	P	T	P	S	N	I
28	L	P	T	P	S	N	I	S
29	P	T	P	S	N	I	S	A
30	T	P	S	N	I	S	A	W
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
379	I	E	N	K	M	L	K	W
380	**E**	**N**	**K**	**M**	**L**	**K**	**W**	**A**

Note: The first and last rows are the first and last fragments of the sequence, the same as in Table 1. The number of a fragment is defined by the number of its last amino acid, therefore the table begins with row 8.

of Euclidean distances between the fragments are calculated (square root of the *p*-distance (Hamming distance) between a couple of fragments is used as the Euclidean distance (Efimov et al., 2013)).

For all matrices of Euclidean distances, its PCs are calculated by the method of PCo (Gower, 1966). The matrices PC-Seq_L, A_L and the vector H_L were combined into one matrix, and for it, the matrix of Pearson correlation coefficients was calculated between all columns. Only the correlation coefficients exceeding a threshold of 0.316 in absolute magnitude ($r^2 \sim 0.1$, i. d. 10 %; $p < 10^{-8}$) were considered. Jacobi 4 package was used for calculations (Polunin et al., 2014).

Results

For the first principal component PC1-Seq_L, the correlation exceeding the threshold was found with a fraction of helix positions ($0.370 \le r \le 0.547$ in the range $4 \le L \le 18$; $r_{max} = 0.547$ for L = 12), leucine content in the fragment ($0.95 \le r$ in the range $2 \le L \le 17$; $r_{12} = 0.974$), proline content ($0.331 \le r \le 0.364$ in the range $9 \le L \le 14$) and tyrosine content ($0.318 \le r \le 0.351$ in the range $14 \le L \le 18$), what is more, correlations PC1-Seq_L with the contents of proline and tyrosine have inverse signs in relation of correlations with the fraction of helix positions and the leucine content in the fragments. The graph of PC1-Seq_L against the background of the fraction of helix positions is shown in Fig. 1, the dynamics of the leucine content against the same background is in Fig. 2, the scatterplot of PC1-Seq_L vs the leucine content in the fragment is shown in Fig. 3 (all for L = 12).

Discussion

The good correlation of the first principal component with a fraction of helix positions means that the similarity of the fragments depends on how much the fragments intersect with α-helices. It is known that hydrophobic amino acids are most often found in α-helices, and hydrophilic ones are outside them. Leucine is a hydrophobic amino acid, and indeed it appears in α-helices of humans more often than other amino acids. This explains the high correlation between the first principal component and the leucine content in the fragments. Note that we did not specifically look for any information about the amino acids content in the fragments or about the secondary structure of the sequence. The only thing that we investigated was how much the fragments coincided with each other by amino acids in total for all L positions. If we had set another measure of similarity, then perhaps we would have discovered some other regularity. In this case, this one is found.

PCA-Seq is certainly an exploratory method, as is its particular case SSA. For any sequence, including molecular, PCA-Seq without any additional assumptions allows obtaining its principal components in a numerical form and visualizing them in the form of phase portraits. Today SSA for numerical series is a huge scientific field with applications in various sciences. There is no doubt that the analysis of non-numeric sequences will become a scientific field, no less in scope than SSA.

It should be noted that the approach of calculation is used in the standard SSA through a covariation (correlation) matrix

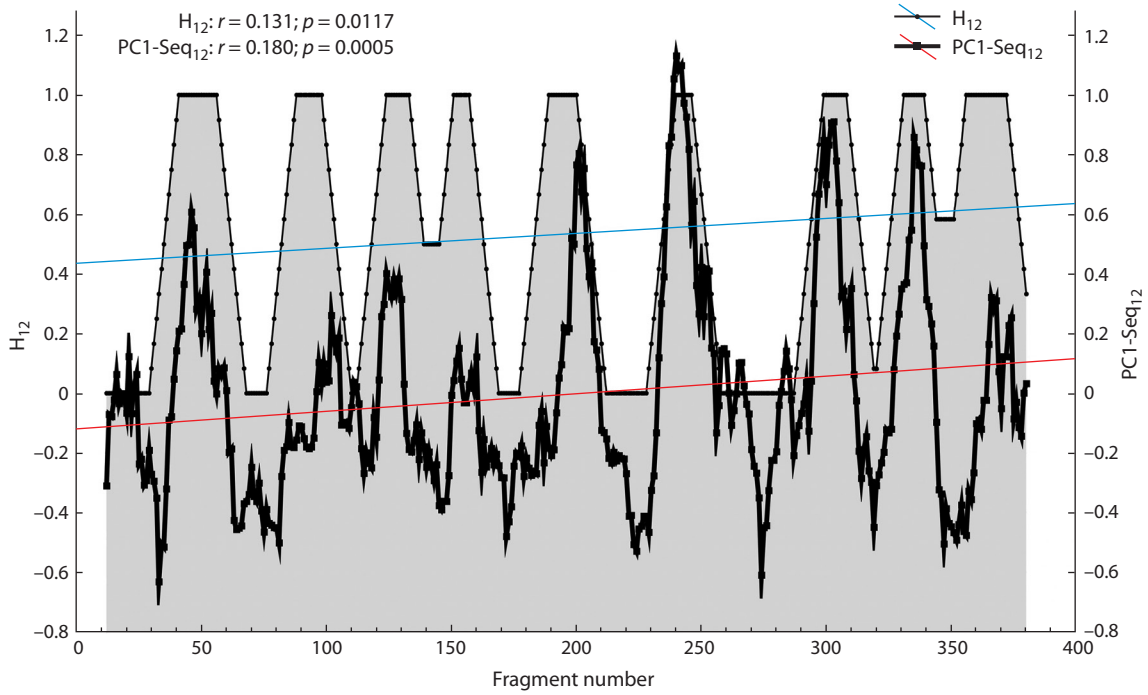

Fig. 1. PC1-Seq$_{12}$ is first principal component of the *Homo sapiens Cytb* amino acid sequence (Q0ZFD6_HUMAN, Swiss Model repository) and H_{12} is the fraction of helix position in sequence fragments of length L = 12 ($r = 0.547$, N = 369, $p < 10^{-9}$).

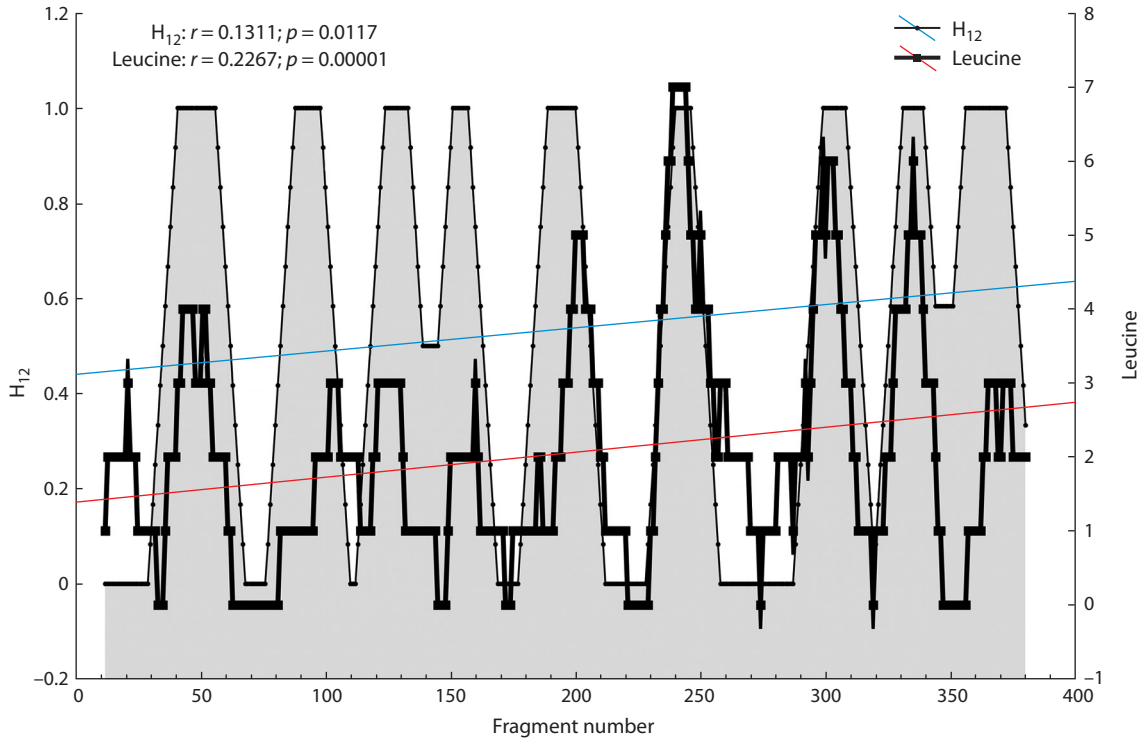

Fig. 2. Content of leucine in *Homo sapiens Cytb* amino acid sequence fragments (Q0ZFD6_HUMAN, Swiss Model repository) and H_{12} is the fraction of helix position in sequence fragments of length L = 12 ($r = 0.499$, N = 369, $p < 10^{-9}$).

only, and the MDS methods, including PCo, despite more than half a century-long history, are almost unknown. This gives reason to hope that PCA-Seq can be useful in the analysis of real data, especially in hypothesizing. PCA-Seq is a particular case of the geometric approach, in which any similarity/dissimilarity relations between objects are modeled by the distance between points in a certain Euclidean space. In this case, the objects are fragments of any sequence by length L, including non-numeric, in particular molecular one. Orthogonal rotations of the entire set of points leave the distances

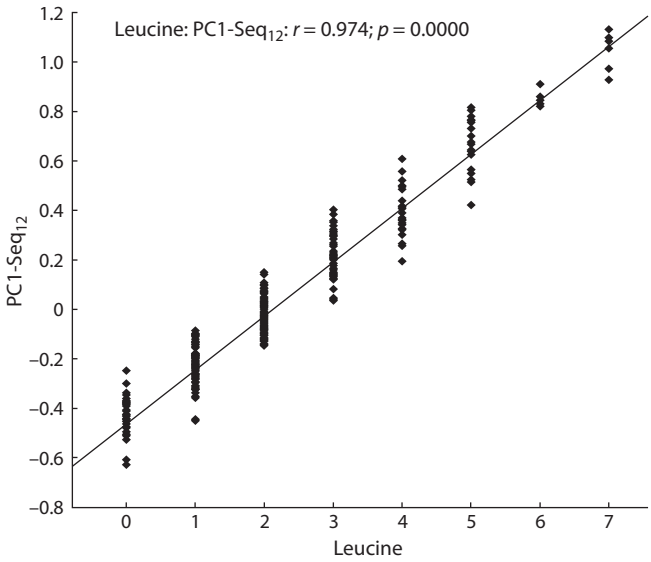

Fig. 3. Content of leucine in *Homo sapiens Cytb* amino acid sequence fragments (Q0ZFD6_HUMAN, Swiss Model repository) vs first principal component PC1-Seq$_{12}$ (L = 12, r = 0.974, N = 369, $p < 10^{-9}$).

between them unchanged. This allows calculating such axes, in the projection on which the maximum variances of the set of points are reached. These axes always exist, they are exactly the principal components. By construction, they are not statistically correlated with each other.

This does not mean at all that they are meaningfully independent. In particular, for a time series, it is a general rule that their PCs, despite being uncorrelated, break up into couples in which one component is derivative of another. When one component shifts from another by a quarter of a period, the correlation appears again. In successful cases, this allows predicting the future values of one component from the already known values of another and thus predicting, to some extent, the initial series. The sine/cosine couple is a telling example. Moreover, it is possible that there is a third component, as a rule, a trend, which, despite the lack of correlation with the first two, modulates their amplitude. Thus, it is a part of a general interconnected complex. In 3D phase space such components form a funnel.

PCA-Seq is implemented in the freely distributed Jacobi 4 package (http://jacobi4.ru/).

Conclusion

PCA-Seq is promising for processing molecular sequences – and then some.

References

Efimov V.M., Galaktionov Y.K. On the possibility of predicting cyclic changes in the abundance of mammals. Zhurnal Obshchey Biologii = Journal of General Biology. 1983;3:343-352. (in Russian)

Efimov V.M., Galaktionov Y.K., Galaktionova T.A. Reconstruction and prognosis of water vole population dynamics on the basis of tularemia morbidity among Novosibirsk oblast residents. Doklady. Biological Sciences. 2003;388(1/6):59-61.

Efimov V.M., Galaktionov Y.K., Shushpanova N.F. Analysis and Prediction of Time Series by the Principal Component Method. Novosibirsk: Nauka Publ., 1988. (in Russian)

Efimov V.M., Kovaleva V.Y., Efimov K.V. Principal Component Analysis for any type Sequences (PCA-Seq). In: Mathematical Modeling and High-Performance Computing in Bioinformatics, Biomedicine and Biotechnology (MM-HPC-BBB-2018): Proc. of the 3rd Int. Symp. Novosibirsk, 21–24 Aug 2018. Novosibirsk, 2018;20.

Efimov V.M., Melchakova M.A., Kovaleva V.Y. Geometric properties of evolutionary distances. Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov Journal of Genetics and Breeding. 2013;17(4/1):714-723. (in Russian)

Golyandina N., Korobeynikov A., Zhigljavsky A. Singular Spectrum Analysis with R. (Ser. Use R!) Berlin; Heidelberg: Springer Verlag, 2018.

Golyandina N., Nekrutkin V., Zhigljavsky A.A. Analysis of Time Series Structure: SSA and Related Techniques. Chapman and Hall/CRC, 2001.

Golyandina N., Zhigljavsky A. Singular Spectrum Analysis for Time Series. Springer Science & Business Media, 2013.

Gower J.C. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53(3/4):325-338.

Jolliffe I.T., Cadima J. Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A. 2016;374:20150202.

Karhunen K. Über lineare methoden in der wahrscheinlich-keitsrechnung. Ann. Acad. Sci. Fennicea. 1947;Ser. A137.

Loève M. Fonctions Aléatoires de second order. In: Lévy P. (Ed.). Processus Stochastiques et Movement Brownien. Paris: Hermann, 1948.

Polunin D.A., Shtaiger I.A., Efimov V.M. Development of software system JACOBI 4 for multivariate analysis of microarray data. Vestnik Novosibirskogo Gosudarstvennogo Universiteta. Seriya Informatsyonnye Tekhnologii = Vestnik NSU. Information Technology. 2014;12(2):90-98. (in Russian)

Takens F. Detecting strange attractors in turbulence. In: Dynamical Systems and Turbulence. Warwick, 1980. Berlin; Heidelberg: Springer, 1981;366-381.

Acknowledgements. Supported by Russian Foundation for Basic Research (# 19-07-00658). The authors are grateful to D.A. Afonnikov, P.N. Menshanov and two anonymous reviewers for useful discussion and constructive comments.

Conflict of interest. The authors declare no conflict of interest.

Received September 19, 2018. Revised June 28, 2019. Accepted July 2, 2019.