The PCA-seq method applied to analyze of the dynamics of COVID-19 epidemic indicators

V M Efimov[1,2,3,5], D A Polunin[2], V Y Kovaleva[3] and K V Efimov[4]
[1]Institute of Cytology and Genetics SB RAS, 10 Ac. Lavrentieva Ave., Novosibirsk 630090, Russia
[2]Novosibirsk State National Research University, 1 Pirogova St., Novosibirsk 630090, Russia
[3]Institute of Systematics and Ecology SB RAS, 11 Frunze St., Novosibirsk 630091, Russia
[4]Higher School of Economics – National Research University, 20 Myasnitskaya St.,
Moscow 101000, Russia
[5]National Research Tomsk State University, 36 Lenina Ave., Tomsk 634050, Russia

E-mail: efimov@bionet.nsc.ru

Abstract. In time series analysis using the SSA method, a univariate series is converted into the multivariate one by shifts. The resulting trajectory matrix is subjected to principal component analysis (PCA). However, the principal components can also be computed using the PCA-Seq method if segments of the original series are selected as objects. The matrix of Euclidean distances between the objects can be obtained using any method, which offers additional opportunities for time series analysis compared to the conventional SSA. In this study, the PCA-Seq method was used to analyze the dynamics of COVID-19 epidemic indicators.

1. Introduction.

When does new knowledge get integrated into the science? Does it occur once this piece of knowledge has been published, or when it has become known to the general public, or when it starts to be used to obtain further knowledge?

Let us take the Principal Component Analysis (PCA) as an example. According to its most generally used definition, this method finds the linear combinations of the variables with maximum variance. Almost half a century has passed between the idea and its practical application. The original idea (the plot) was presented by Galton (1886) [1], and Pearson's article (1901) [2] was the first publication focusing on this method. PCA has become widely known and started to be practically applied after the Hotelling's article (1933) [3] had been published. The main approach involved computing the correlation matrix and its eigenvectors. This very publication has immediately made Hotelling famous, wherease it was not until the late 20th century that the world learned about the Pearson's priority and Galton's contribution. It is worth noting that Galton's work has paved the way for biometrics, which later on became mathematical statistics. However, this very plot remained ununderstood at that time (probably for Galton as well), and no citations to the Pearson's study (including autocitations) were made for almost a century.

Principal coordinate (PCo) analysis consists in finding mutual arrangement of the objects in a multivariate Euclidean space using the matrix of Euclidean distances (EDM) between them. Torgerson's paper [4] was the first publication focusing on principal coordinate analysis. Gower [5] proposed an algorithm and proved that PCo and PCA are *equivalent*. To date, the Gower's work has been cited more than 4,000 times. It would be wrong to deny that it is widely known.

However, PCo is almost never used in practice. The reason for that is rather simple. In full compliance with the aforementioned definition, researchers do not perceive PCA separately from variables and the correlation/covariance matrices. From this perspective, it is difficult to compute PCA

in a situation when there are many variables (or extremely difficult if their number is very large). In the case when there are no variables, using PCA is out of the question. And if there are no variables, how can the results be interpreted?

PCo is still perceived only as a multidimensional scaling (MDS) method. The experts simply are unaware that PCo is another technique for computing PCA and it is enough to know the EDM between the objects. The objects themselves can even be non-numeric [6].

This is true for time series analysis using SSA [7]. When analyzing time series using the SSA method, a univariate series is converted into a multivariate one by shifts. The resulting "object–variable" trajectory matrix is subjected to PCA. Although the objects and variables in this matrix are formally equitable, SSA views the trajectory matrix solely as a combination of variables, while the objects are not even regarded as independent entities.

However, the principal components can also be computed using the PCA-Seq method if segments of the original series are selected as objects. Any Euclidean distances between the segments can be chosen in time series analysis [8], thus offering additional opportunities compared to the conventional SSA method.

In this study, the PCA-Seq method was used to analyze the dynamics of COVID-19 epidemic characterisics.

2. Material and methods.

The original data was taken from the website https://ourworldindata.org/coronavirus and further verified using the website https://www.worldometer.info/coronavirus/. The dynamics of the daily new_cases_smoothed and new_deaths_smoothed in the USA are analyzed. December 31, 2019 (the day the epidemic began in Wuhan, China) is considered the first day of the global pandemic, so for the United States, the first curve starts from the 22nd day, the second from the 62nd day.

Both series were logarithmized and processed using SSA and PCA-Seq.

The PCA-Seq algorithm [8]. Let there be a sequence $Q = \{q_1, q_2, ..., q_N\}$ 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, $Q_i = (q_{i-L+1}, q_{i-L+2}, ..., q_{i-1}, q_i)$, $i = L, ..., N$. Compute the matrix of any Euclidean distances $D = (d_{ij} = d(Q_i, Q_j))$ between all fragments. Apply the method of principal coordinates (PCo) to D and obtain its principal components PCs [5]. If there is a numerical series and the squared Euclidean distances are ordinary sums of squared differences, then PCA-Seq is equivalent to SSA. In general, SSA is a special case of PCA-Seq.

The outbreak typically involves several stages: the onset that goes almost unnoticed and is difficult to detect; the rise (usually exponential); the plateau (fluctuations around or near a certain level); and the decline (either slow or fast). Sometimes the plateau stage is suddenly followed by one or more uplift waves. As a rule, this indicates the territorial disunity of the population.

In order to adequately take this into account, we calculate the distance between the fragments as follows. The behavior patterns of the fragments corresponding to each stage are supposed to be more similar. They can actually be regarded as small time series. For time series, the behavioral similarity regardless of their scale is usually assigned by correlation coefficient r. If the time series are pre-standardized, then $r = \sum x_i y_i$. Standardization is subtracting the mean and dividing by the standard deviation. After the standardization, all the series are of unit length; the correlation coefficient is the cosine of the angle between two vectors. Angle is a locally unit spherical Euclidean distance of Cavalli-Sforza & Edwards; $d = \arccos(r)$ [9, 10]. Therefore, in order to compute PCs, it is sufficient to build a correlation matrix for all the segments of a time series (standardization is automatically performed for the correlation coefficient), transform it to EDM through arccosine, and apply PCo to it. We will denote them arcPCs in order to distinguish them from PCs obtained using regular SSA (ssaPCs).

The computations were carried out using the PAST, Statistica, and Jacobi4 packages [11].

3. Results.

It is known that there are 7-day fluctuations in the series under study, caused both by a delay in the registration of illness and death on weekends, and by a real rise in morbidity due to an increase in the number of contacts on weekends. On the other hand, when processing time series using a sliding window, there is always a small induced cyclicity with a period equal to the length of the window (Slutsky effect). To avoid interference of these periods lag L is chosen equal to 11.

Fig. 1a shows the dynamics of the daily new_cases_smoothed (Ncs) in USA, the 11-day mean, and ssaPC1. It can be seen that the behavior of ssaPC1 almost completely coincides with the dynamics of the mean (r=0.99). This is not surprising, since it accounts for almost 100% of the total variance. Undoubtedly, this is caused by the peculiarities of the series itself: a fast and powerful rise, even on a logarithmic scale, and then a slight decline, interrupted by small rises. Against the background of the initial rise, all these fluctuations seem small, but the number 4, around which these fluctuations occur, means 10^4, that is, about a ten thousand cases per day.

The dynamics of arcPC1 shows that the positions of the 11-day segments on the arcPC1 axis are different during the up and down phases, but they are close in each phase. The position of the trajectory above zero coincides with the rise of the epidemic curve. This graph and the phase portrait (Fig. 1b) demonstrate quite clear and regular cyclical fluctuations with a period of about 100 days, two full and the beginning of the third. The nature of these fluctuations is unclear. A possible explanation could be the following consideration. The United States consists of several dozen fairly autonomous populations. The central government does not have the ability to quarantine the whole country, state governors have enough authority to conduct an independent policy. Therefore, epidemics in them occur at different times and at different scales; as a result, the epidemic curve actually consists of the sum of the curves and stretches over time. However, fluctuations look too regular for such an explanation.

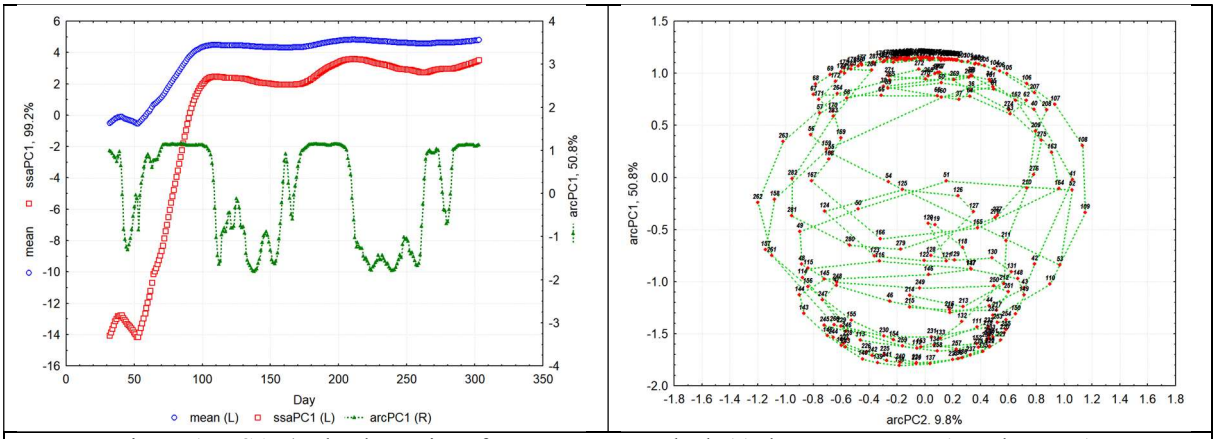

Figure 1. *USA*. a) The dynamics of new cases smoothed: 11-day mean, ssaPC1, and arcPC1;
b) The phase portrait for the trajectory of new cases dynamics in the projection onto first two arcPCs plane.

The new deaths dynamic and its indicators (Fig 2a, b), naturally, are several days behind the new cases dynamic and the scale is much smaller, but the nature of the fluctuations is the same. The only difference is that the new cases dynamic is not going to decline, and the new deaths dynamic is going down a little bit. Time will tell whether this trend will continue. In any case, it is too early to talk about the end of the pandemic.

Why are SSA and PCA-Seq useful in epidemic curve analysis and how are they different? The ssaPC1 practically repeats the moving average and practically does not react to its small deviations in one direction or another. The arcPC1, on the other hand, begins to fluctuate sharply if the constancy

(plateau) has been broken (even slightly). There is nothing mysterious about this. The arcPC1 plot, obtained through correlations between fragments, shows the direction of changes in the moving average, not the absolute level it reached. Since the arcPC1 value was calculated using all the values within the fragment, it is as computationally robust as the moving average. However, the moving average cannot distinguish between the rising and falling stages, since in both cases it passes through the same values.

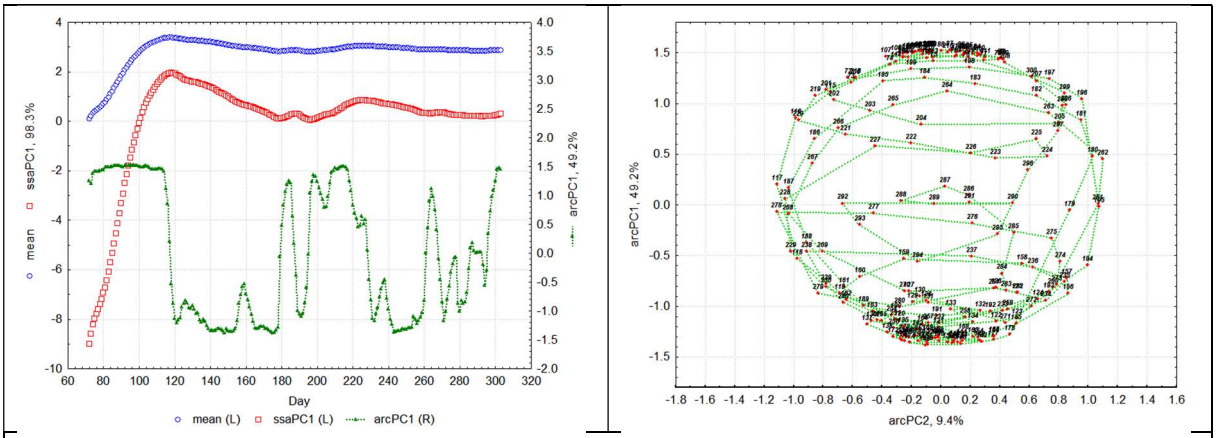

Figure 2. *USA*. a) The dynamics of new deaths smoothed: 11-day mean, ssaPC1, and arcPC1; b) The phase portrait for the trajectory of new deaths dynamics in the projection onto first two arcPCs plane

The direction of changes in the number of new cases is positive at the growth stage, fluctuates around zero at the plateau, and negative at the decline stage. The arcPC1 provides additional information regarding the phase of the epidemic and changes that will occur in the near future. Of course, we can and get this information if we take the derivative of the moving average. But for this we need a few more points. But arcPC1 provides this information now, based on the same fragment, which was used to calculate the moving average.

One can try using the following principal components for this purpose, ssaPC2 and ssaPC3 (Fig 3a). If we calculate the first and second derivatives of the dynamics new deaths and look at their phase portrait, then we will see a significant similarity (Fig. 3b; Table 1). This is not an accidental similarity, but a well-known property of orthogonal decomposition of a time series [11]). By Bartlett's theorem for stationary random time series, all odd derivatives are in the aggregate orthogonal to even. One of the components is usually a derivative of the other component. The sine–cosine pair is an example. The correlation coefficient for them is zero.

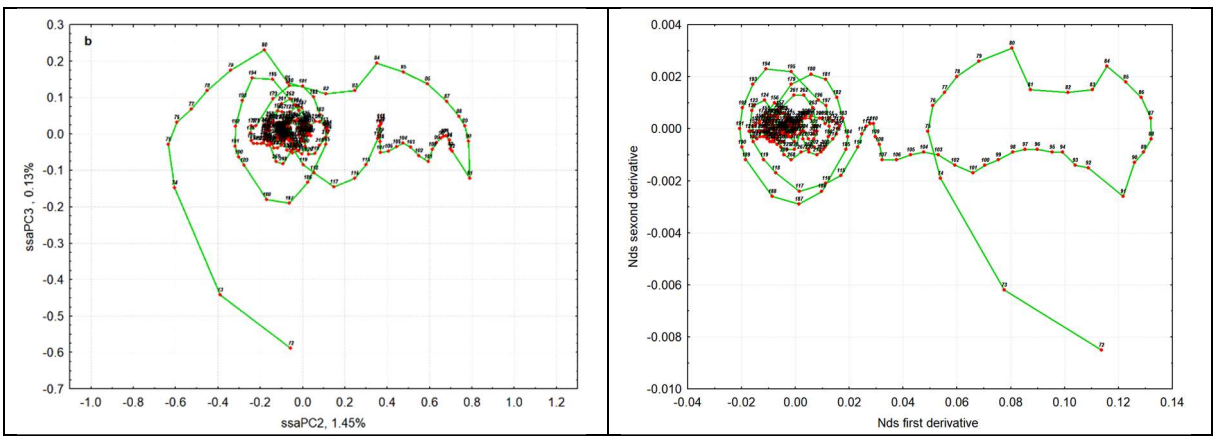

Figure 3. a) The phase portrait for the trajectory of new deaths dynamics in the projection onto ssaPC2–ssaPC3 plane; b) The phase portrait for the trajectory of new deaths dynamics in the projection onto the plane of first & second derivatives.

If so, why then apply PCA-seq to a time series at all? Derivatives are also sensitive to process scaling (Fig. 3b; Table 1). In fact, they only react to noticeable deviations from the current level.

In PCA-Seq, the situation is different. We choose the distance ourselves and can choose it more suitable for the situation. For example, in Fig. 1b, 2b phase portraits are practically a circles. The difference with the previous graphs, Fig 1a, 2a, is that another distance was chosen between the fragments. The correlation coefficient removes not only the mean, but also the variance. In fact, the angular distance of Cavalli-Sforza adjusts itself on any part of the curve.

Table 1. Correlation coefficients between indicators of COVID-19 dynamics

USA	Mean	StDev	FirstDev	SecondDev	ssaPC1	ssaPC2	ssaPC3	arcPC1	arcPC2
Mean	1	-0.723	-0.717	0.028	0.991	-0.016	0	0.438	0.061
StDev	-0.723	1	0.958	-0.185	-0.733	-0.619	-0.029	-0.546	0.04
FirstDev	-0.717	0.958	1	-0.167	-0.728	-0.685	0.011	-0.706	0.033
SecondDev	0.028	-0.185	-0.167	1	0.032	0.225	0.97	0.109	-0.657
ssaPC1	0.991	-0.733	-0.728	0.032	1	0	0	0.447	0.059
ssaPC2	-0.016	-0.619	-0.685	0.225	0	1	0	0.555	-0.123
ssaPC3	0	-0.029	0.011	0.97	0	0	1	-0.038	-0.628
arcPC1	0.438	-0.546	-0.706	0.109	0.447	0.555	-0.038	1	0
arcPC2	0.061	0.04	0.033	-0.657	0.059	-0.123	-0.628	0	1

4. Discussion.

From the point of view of using the mathematical apparatus, both SSA and its generalization PCA-Seq help to more comprehensively imagine the nature of the phenomenon under study and, possibly, will be useful in developing forecast methods. Another possible application is the verification of mathematical models expressed as systems of differential equations. The processing of the solutions obtained in them using orthogonal decompositions and comparison with the results of processing the observed data can help to verify the models themselves.

The options for choosing Euclidean distances in PCA-Seq are certainly not limited to the spherical Euclidean distance of Cavalli-Sforza & Edwards. In particular, very promising are the distances obtained from ecological indices [13], for example, the Jaccard-Steinhaus distance obtained from the Jaccard index, which is widely used now.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (project no. 19 07 00658 a) and the Budget Project of the Institute of Cytology and Genetics, SB RAS (project no. 0324 2019 0040 C 01).

References

[1] Gal1ton F 886 Regression towards mediocrity in hereditary stature *J. Ant. Inst. Great Britain and Ireland* **15** 246-263

[2] Pearson K 1901 On lines and planes of closest fit to systems of points in space *Phil. Mag. J. Sci.* **2** 559-572

[3] Hotelling H 1933 Analysis of a complex of statistical variables into principal components *J. Edu. Psyc.* **24** 417
[4] Torgerson W S 1952 *Psychometrika* **17** 401-419
[5] Gower J C 1966 *Biometrika* **53** 325-338
[6] Efimov V M, Efimov K V and Kovaleva V Y 2019 Principal component analysis and its generalizations for any type sequence (PCA-Seq) *Vavilov J. Gen. and Breeding* **23** pp 1032-1036
[7] Golyandina N, Nekrutkin V and Zhigljavsky A A 2001 Analysis of time series structure: SSA and related techniques (CRC press)
[8] Deza M M and Deza E 2009 Encyclopedia of distances (Springer, Berlin, Heidelberg) pp 1-583
[9] Burago D, Burago I D, Burago Y, Ivanov S A and Ivanov S 2001 A course in metric geometry **33** (Am. Math. Soc.)
[10] Cavalli-Sforza L L, Edwards A W 1967 Phylogenetic analysis. Models and estimation procedures *American journal of human genetics* **19** 233
[11] Polunin D A, Shtaiger I A and Efimov V M 2019 JACOBI4 software for multivariate analysis of biological data *bioRxiv* 803684
[12] Bartlett M S 1978 An introduction to stochastic processes: with special reference to methods and applications. 3rd Ed (Cambridge: Cambridge University Press)
[13] Gower J C and Legendre P 1986 Metric and Euclidean properties of dissimilarity coefficients *Journal of classification* **3** 5-48