MULTIFACTOR ANALYSIS OF TIME SERIES[1]

Contents

[1]This description is implemented by V.N. Troyan and Yu.V. Kiselev, St.Petersburg State University, 2003.

List of Figures

1 Introduction

Last twenty years many publications devoted to investigation and application of the multifactor analysis (or singular spectrum analysis [3]) to time series analysis have appeared in various areas of scientific exploration: climatology, meteorology, physics, signal processing.

The birth of such algorithms is usually associated with the first papers of Broomhead and King [1], [2]. At present the list of publications is more than hundred. Around sixty references can be found in the book [3] together with many examples of application of such approach to analysis of the time series and description of various theoretical and practical tasks.

We call the represented in section 2 algorithm as multifactor analysis of the time series (MAS) [4], since the algorithm has likeness to one of variants of the factor analysis, because in MAS the time series are represented as the sample values of a random vector (of given length M) with the subsequent application of singular value decomposition (SVD) to a sample correlation matrix of such vector and evaluation of the principal components. On the other hand, it is possible to interpret MAS as decomposition of original time series on a system of basis functions constructed on the basis of the original time series with the help of its sample correlation matrix. The order of the correlation matrix and, consequently, a number of the basis functions is a free parameter of the problem and it should be determined by the interpreter according to a current task.

Such algorithm is essentially model-free technique; it is more an exploratory, model building tool than a confirmatory procedure. The goal of MAS is the decomposition of the original series into a sum of a small number of interpretable components such as a slow varying trend, oscillatory components and a 'structureless' noise.

2 Description of algorithm

We consider a time series

$$\{x_i\}_{i=1}^N \Rightarrow x_i = f((i-1)\Delta t), \quad i = 1, 2, \ldots, N,$$

where Δt is a time slice and N is a number of samples.

At the first step we evolvent of 1-D time series to multidimensional one by creation of matrix

$$X = (x_{ij})_{i,j=1}^{k,M} = \begin{bmatrix} x_1 & x_2 & x_3 & \ldots & x_M \\ x_2 & x_3 & x_4 & \ldots & x_{M+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_k & x_{k+1} & x_{k+2} & \ldots & x_N \end{bmatrix},$$

where M is a length of the row i ($i = 1, \ldots, k$, $k = N - M + 1$).

At the second step the centering and normalization of the matrix X are implemented:

$$x_{ij}^* = \frac{(x_{ij} - \bar{x}_j)}{s_j}, \quad i = 1, \ldots, k; \quad j = 1, \ldots, M$$

where

$$\bar{x}_j = \frac{1}{k}\sum_{i=1}^{k} x_{i+j-1} \quad \text{and} \quad s_j = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_{i+j-1} - \bar{x}_j)^2} \quad (j = 1, \ldots, M)$$

are the sample mathematical expectation and sample standard deviation respectively.

At the third step with the use of the singular value decomposition (SVD) of sample correlation matrix

$$R = \frac{1}{k} X^{*T} X^* \tag{1}$$

we get decomposition:

$$R = P \Lambda P^T,$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_M \end{bmatrix} \quad \text{and} \quad P = (\vec{p}_1, \vec{p}_2, \dots, \vec{p}_M) = \begin{bmatrix} p_{11} & p_{21} & \dots & p_{M1} \\ p_{12} & p_{22} & \dots & p_{M2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1M} & p_{2M} & \dots & p_{MM} \end{bmatrix}$$

are the matrices of eigenvalues and eigenvectors of sample correlation matrix R (1) respectively. For the matrices P and Λ the next properties are valid

$$P^T = P^{-1}, \quad P^T P = P P^T = I_M, \quad \Lambda = P^T R P, \quad \sum_{i=1}^{M} \lambda_i = M, \quad \prod_{i=1}^{M} \lambda_i = \det R.$$

At the fourth step we introduce the matrix of principal components

$$Y = (\vec{y}_1, \vec{y}_2, \dots, \vec{y}_M) = X^* P \tag{2}$$

which possesses the next properties

$$Y^* = Y \Lambda^{-1/2}, \quad {Y^*}^T Y^* = I_M.$$

The transform (2) can be interpreted as linear filtering of the initial time series

$$y_j(l) = \sum_{q=1}^{M} X^*_{lq} p_{jq} = \sum_{q=1}^{M} (x_{l+q-1} - \bar{x}_q) \frac{p_{jq}}{s_q} = \sum_{q=1}^{M} x_{l+q-1} \frac{p_{jq}}{s_q} - \sum_{q=1}^{M} \bar{x}_q \frac{p_{jq}}{s_q}.$$

At the last step we recover the matrix

$$X^* = Y P^T$$

with the full matrix Y or with the use of only one or more than one principal components. In this case, after denormalization and decentering of X^*, we obtain the initial matrix X and consequently the initial time series or some extraction from it.

3 Multifactor analysis of simple functions

The goal of this section is to acquaint of the interpreter with features of MAS on examples multifactor analysis of the simple functions:

- unit sample function;
- unit step function;
- roof function;
- smooth finite function;
- sine function;
- sine function with linear trend;
- sine function and unit sample function.

We guess, that the analysis of unit sample function, sine function and unit sample function, unit step function, roof function can be considered as model of various outliers or unsufficient frequent quantization of time series. The analysis of sine function and sine function with linear trend are examples of eduction periodic and slowly varying component. The goal of these examples also to show a necessity of experience acquisition on the analysis of the model functions before the analysis of experimental datasets.

On horizontal axes of all figures of this section the numbers of samples are marked.

3.1 Unit sample function

The unit sample function can be considered as a model of outlier. The recovery of the unit sample function with the use of 3 principal components is represented at Fig. 1. The total number of principal components (see Fig. 2) for decomposition is 5. Explicit recovery of the original time series can be gained with use of 4 principal components.

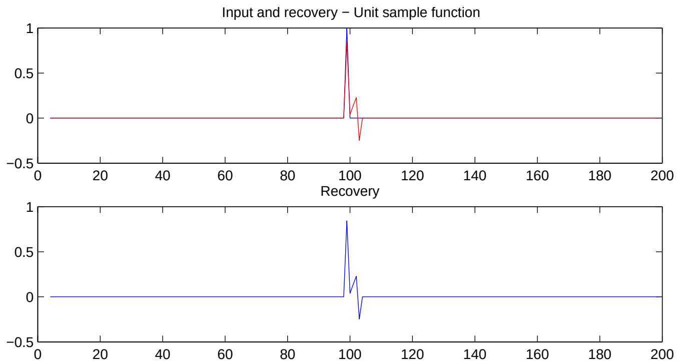

Figure 1. Recovery of the unit sample function with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 5 (horizontal axis – samples, here and further).

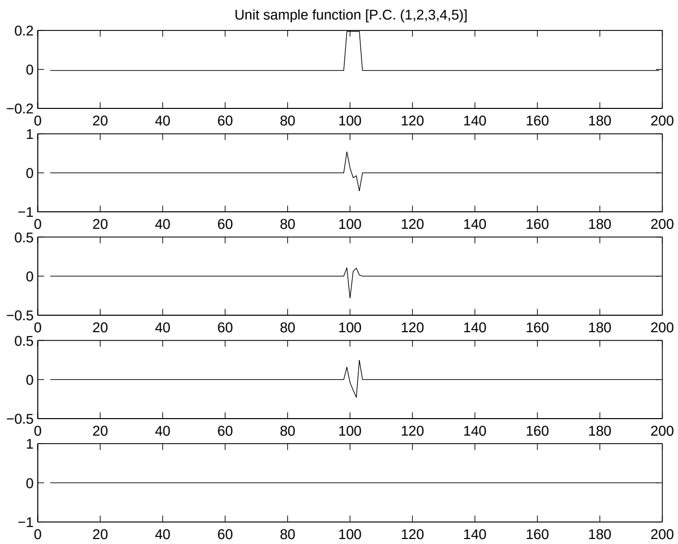

Figure 2. Five normalized principal components for the unit sample function. Total number of the principal component for decomposition is 5.

3.2 Unit step function

The unit step function can be considered, for example, as a model of unsufficient frequent quantization. The recovery of the unit step function with the use of 3 principal components is represented at Fig. 3. The total number of principal components (see Fig. 4) for decomposition is 5.

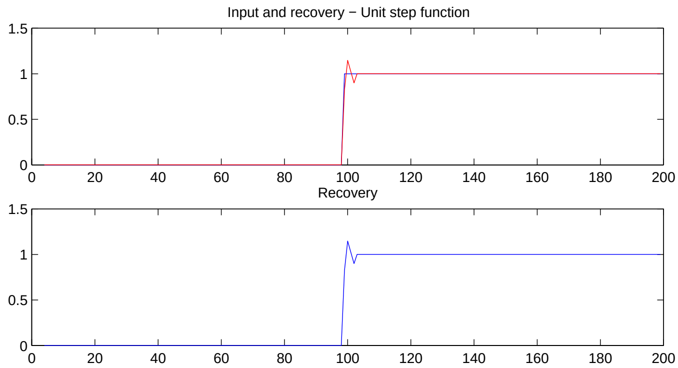

Figure 3. Recovery of the unit step function with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 5.

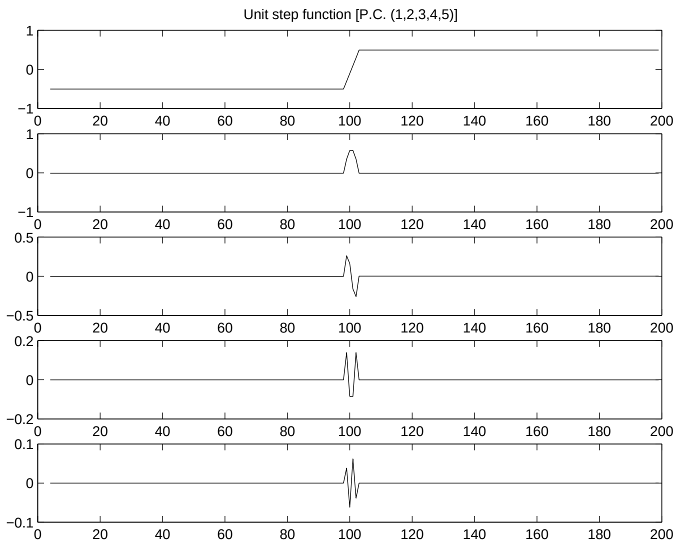

Figure 4. Five normalized principal components for the unit step function. Total number of the principal component for decomposition is 5.

3.3 Roof function

The roof function can be considered, for example, as a model of unsufficient frequent quantization. The recovery of the roof function with the use of 3 principal components is represented at Fig. 5. The total number of principal components (see Fig. 6) for decomposition is 5.

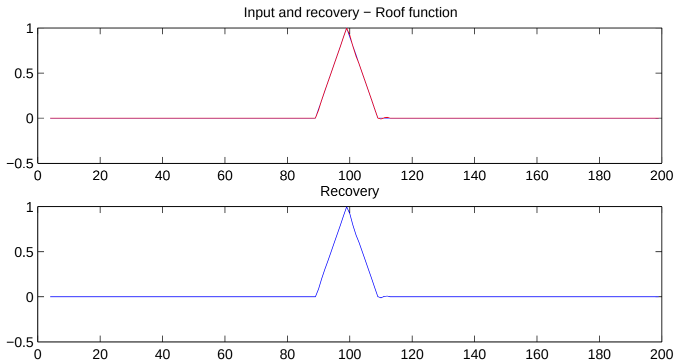

Figure 5. Recovery of the roof function with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 5.

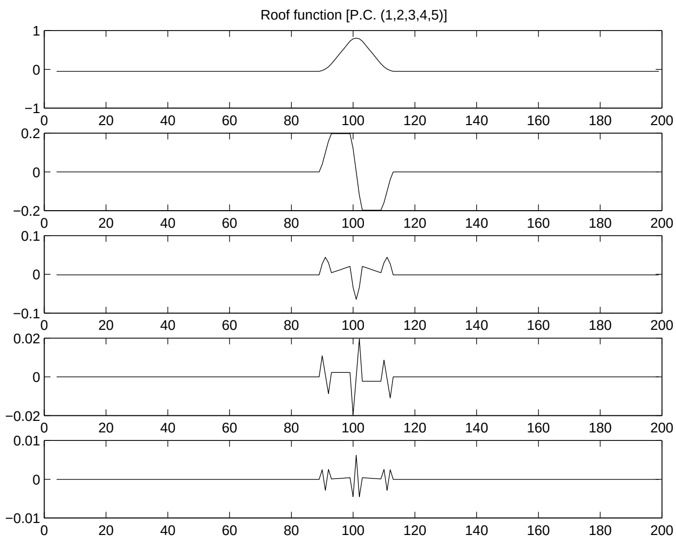

Figure 6. Five normalized principal components for roof function. Total number of the principal component for decomposition is 5.

3.4 Smooth finite function

The smooth finite function can be considered as a model of a part of the time series, which stands out against of the other part of the time series. The recovery of the smooth function (Ricker wavelet) with the use of 3 principal components is represented at Fig. 7. The total number of principal components (see Fig. 8) for decomposition is 5. The Ricker wavelet $f(t)$ in spectral domain ($F(\omega)$) is given by the formula

$$F(\omega) = \left(\frac{\omega}{\omega_0}\right)^2 \exp\left\{-\left(\frac{\omega}{\omega_0}\right)^2\right\} \exp\left\{-\frac{2\pi i\omega}{\omega_0}\right\},$$

where ω is angular frequency; ω_0 is apparent angular frequency of the wavelet.

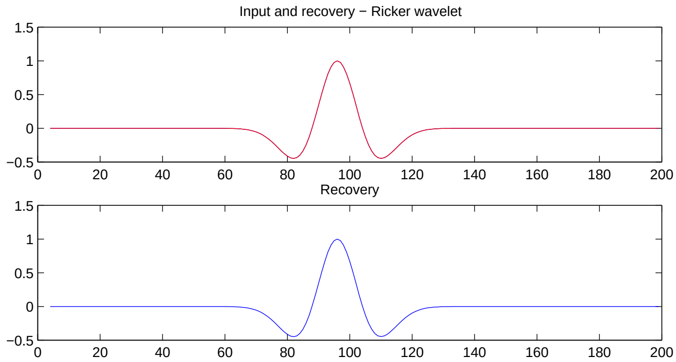

Figure 7. Recovery of smooth finite function (Ricker wavelet) with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 5.

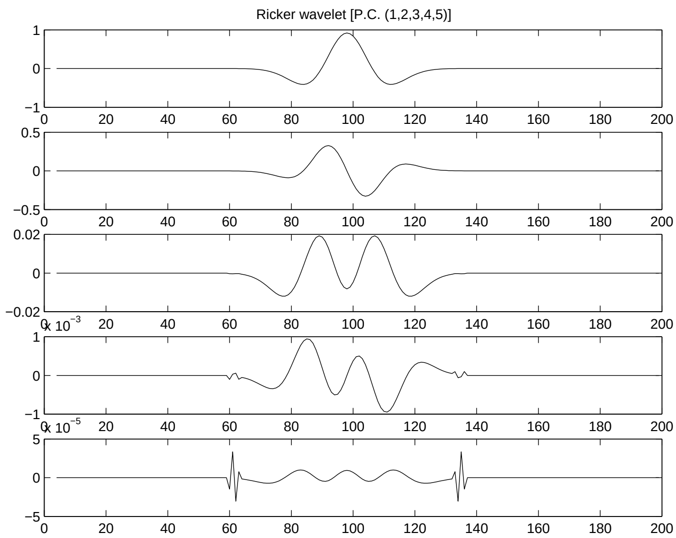

Figure 8. Five normalized principal components for the smooth finite function (Ricker wavelet). Total number of the principal component for decomposition is 5.

3.5 Sine function

The sine function can be considered, as a model of oscillatory component of the time series. The recovery of the roof function with the use of 3 principal components is represented at Fig. 9. The total number of principal components (see Fig. 10) for decomposition is 5. At Fig. 11 the diagram for two first normalized principal components is represented. In the case of extraction of the oscillatory component from the time series the number of principal component for decomposition should be equal or greater than a number of samples per period of the oscillatory component.

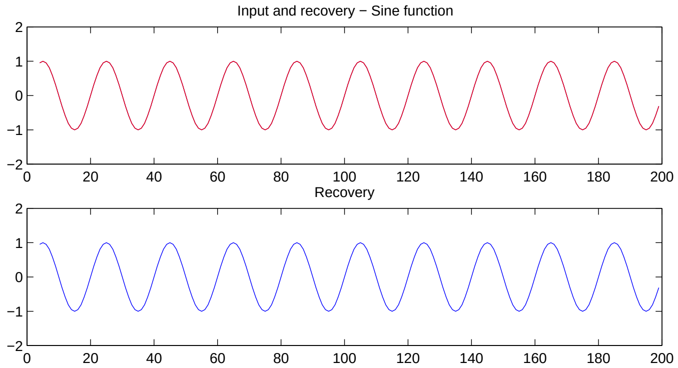

Figure 9. Recovery of the sine function with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 5.

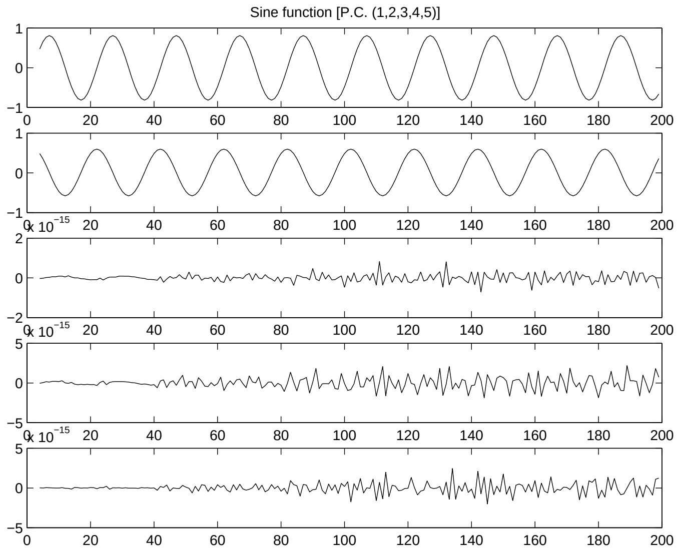

Figure 10. Five normalized principal components for the sine function. Total number of the principal component for decomposition is 5.

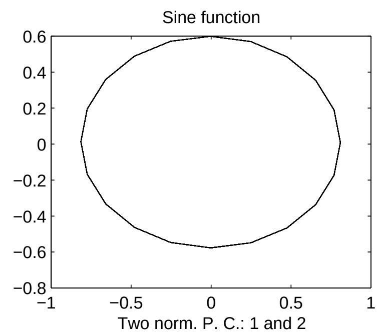

Figure 11. The diagram for two first normalized principal components of the sine funcion. Total number of the principal component for decomposition is 5.

3.6 Sine function with linear trend

The sum of the sine and linear functions

$$f(t) = 0.01t + \sin(2\pi t/20) + \cos(2\pi t/20)$$
$$(t = 0, 1, \ldots, 200)$$

can be considered as a model of extraction of the oscillatory component and slow varying trend. The recovery of $f(t)$ with the use of 3 principal components is represented at Fig. 12 (Fig. 14 – with addition of noise). The total number of principal components (see Fig. 13 and Fig. 15 with addition of noise) for decomposition is 40.

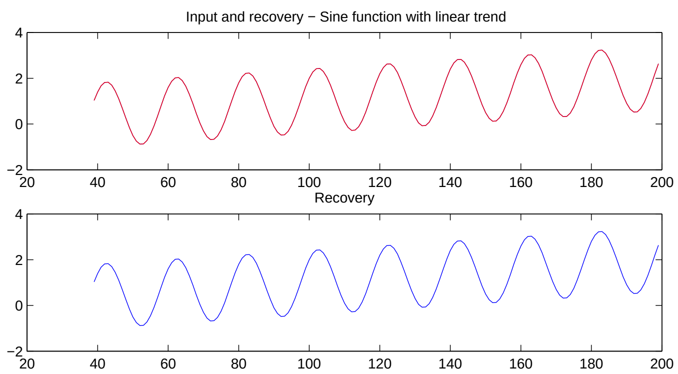

Figure 12. Recovery of the sine functions and the linear trend with the use of 3 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 40.

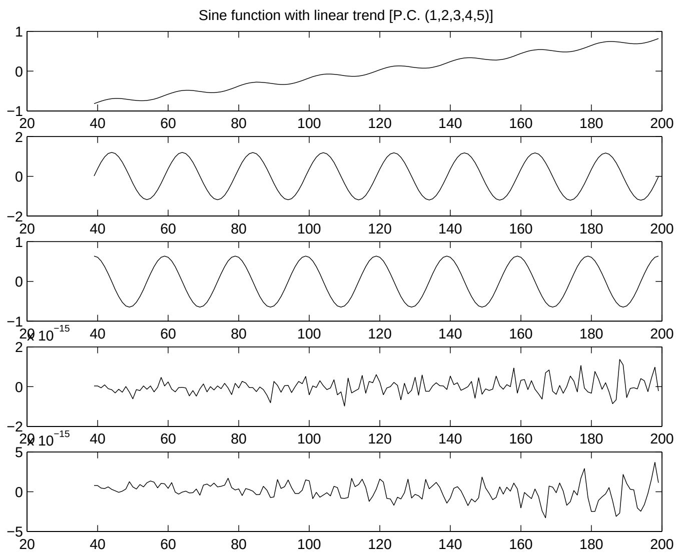

Figure 13. Five normalized principal components for the sine functions with linear trend. Total number of the principal component for decomposition is 40.

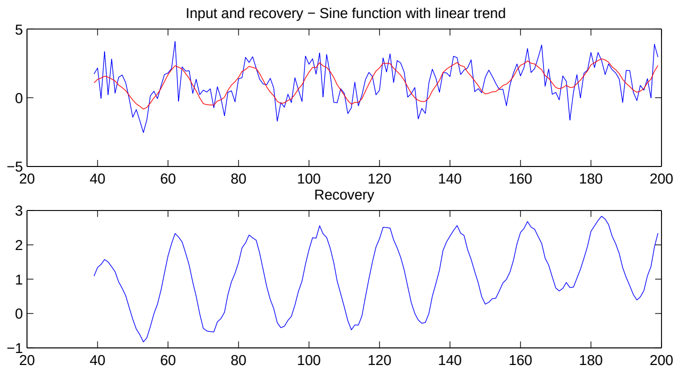

Figure 14. Recovery of the sine functions and the linear trend with the use of 3 first principal components. Gaussian uncorrelated noise with standard deviation 1 is added. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 40.

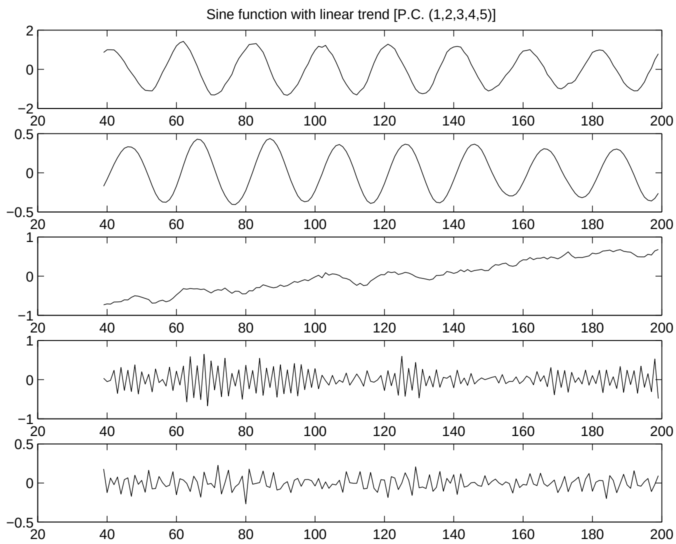

Figure 15. Five normalized principal components for the sine functions with linear trend. Gaussian uncorrelated noise with standard deviation 1.0 is added. Total number of the principal component for decomposition is 40.

3.7 Sine function and unit sample function

The sum of the sine function and unit sample function can be considered, as a model of oscillating component and outlier. The recovery of this function with the use of 5 principal components is represented at Fig. 16. The total number of principal components (see Fig. 17) for decomposition is 10.

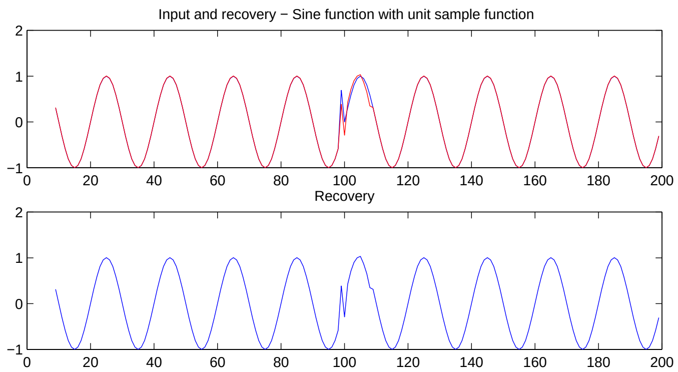

Figure 16. Recovery of the sine function and unit sample function with the use of 5 first principal components. Upper part - original function (blue curve) and its recovery (red curve); lower part - recovery. Total number of the principal component for decomposition is 10.

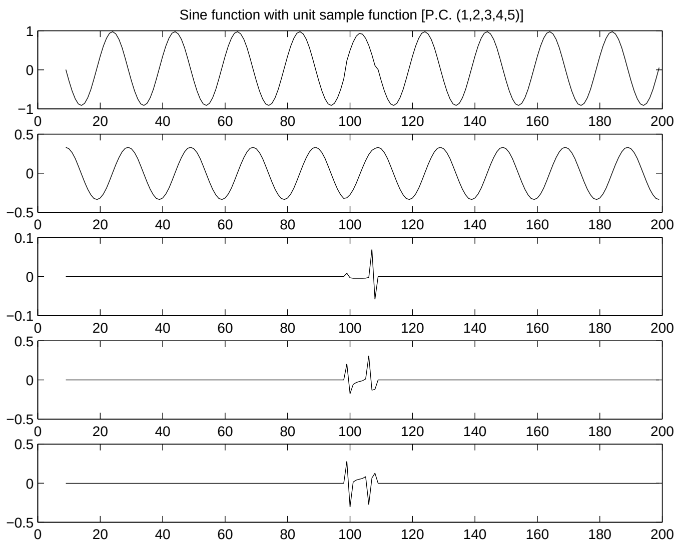

Figure 17. Five normalized principal components for the sine function and unit sample function. Total number of the principal component for decomposition is 10.

4 Examples of analysis of temperature time series and river run-off.

The multifactor analysis was implemented to 38 time series, recorded at the north meteorological stations, and run-off of 7 Siberia rivers (North Dvina, Enisei, Indigirka, Kolyma, Lena, Ob', Pechora) [4]. The main attention is given to comparison of results of processing annual and monthly data. At Fig. 18 the results of extraction of slow varying component from time series (meteorological station Bergen/Fredriksberg) are shown. The slow vary-

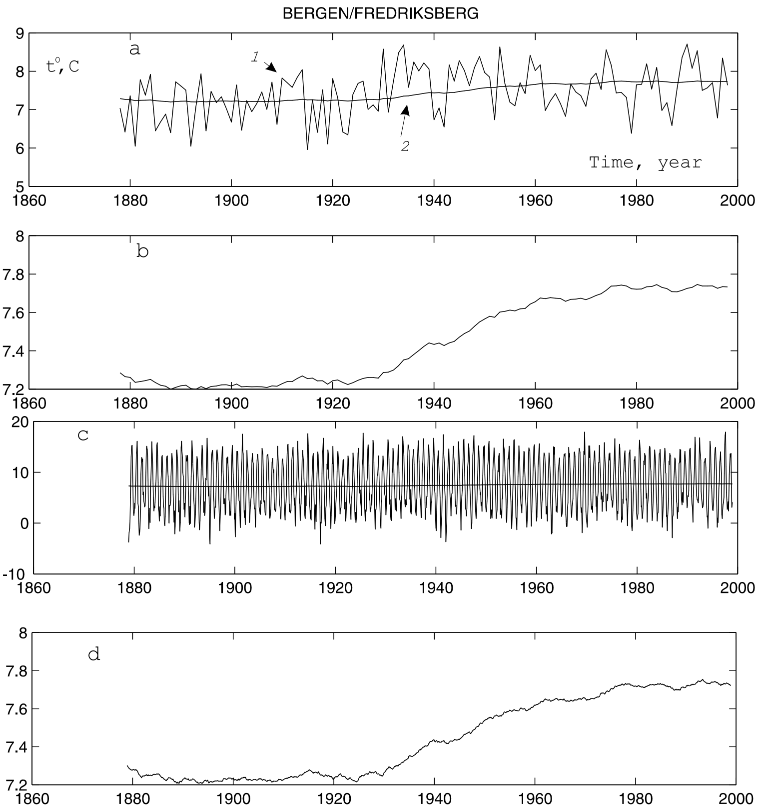

Figure 18. Example of extraction of slow varying component (temperature time series). a – mean annual temperature (1) and slow varying component (2); b – slow varying component extracted from annual data (60 principal components are used); c and d – mean monthly temperature and its slow varying component correspondingly (720 principal components are used).

ing components represented in Fig. 18b and Fig. 18d extracted respectively from annual data (Fig. 18a) and monthly data (Fig. 18c) are very close to each other. In the case of annual data slow varying component is produced by the first principal component and in the case of monthly data it is produced by the fifth principal component.

The smoothing of time series is also can be stable relative to annual and monthly data. At Fig. 19 is represented North Dvina run-off (Fig. 19a – annual data, Fig. 19c – monthly data). The smoothing is implemented using five principal components for both annual (Fig. 19b, with the use of 1 – 5 principal components) and monthly data (Fig. 19d, with the use of 12 – 16 principal components). We observed similar stability on other surveyed data.

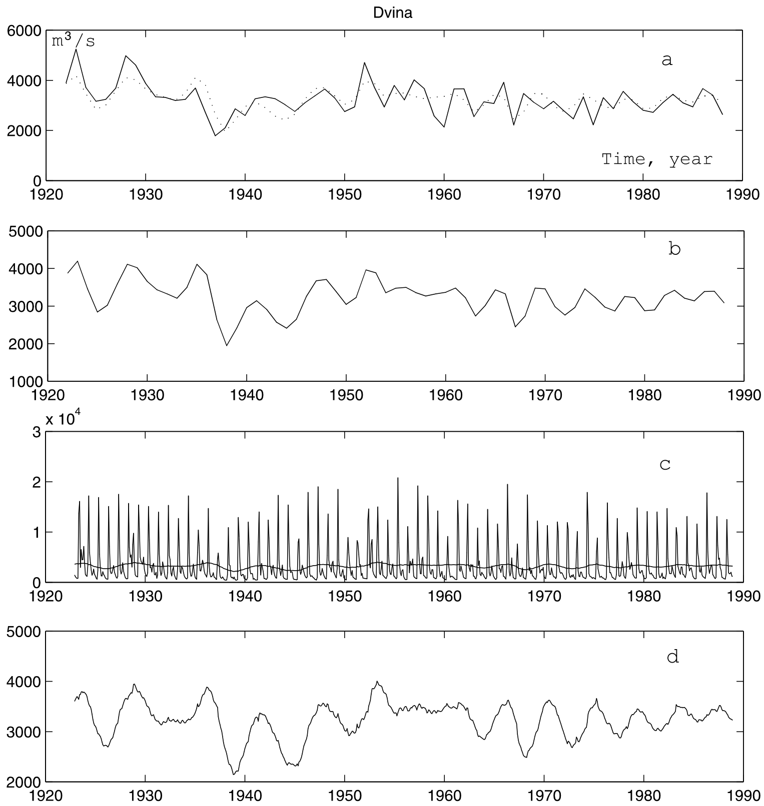

Figure 19. Example of smoothing of river run-off (North Dvina). a – mean annual run-off (solid) and and its smoothing using 5 principal components (dotted); b – smoothing using annual data (22 principal components are used); c and d – mean monthly run-off and its smoothing (264 principal components are used).

References

[1] Broomhead, D. S. and G. P. King (1986) Extracting qulitative dynamics from experimental data, *Physica D*, **20**, 217-236.

[2] Broomhead, D. S. and G. P. King (1986) On the qualitative analysis of experimental dynamical system, *In S. Sarkar (Ed.), Nonlinear Phenomena and Chaose*, pp. 113-144, Adam Higer, Bristol.

[3] Golyandina, N., V. Nekrutkin, A. Zhigljavsky, (2001) *Analysis of time series structure*, Chapman & Hall/CRC, Boca Raton, London.

[4] Troayn, V. N. and Yu. V. Kiselev (2002) *Application of the multifactor analysis to processing climatic time series,* Proc. of the International conference Problems of Geocosmos, St. Petersburg, June 3-8, 2002, pp.297-300, St.Petersburg University, St.Petersburg.