Amer. Math. Soc. Transl.
(2) Vol. 162, 1994

Time Series Analysis and Grassmannians

V. M. Buchstaber

Introduction

We consider the well-known problem of constructing a model (approximation) for a time series $f = (f_1, \dots, f_N)$ in the class of functions of the form

$$g(t) = \sum_{k=1}^{K} a_k(t) e^{\lambda_k t} \sin(\omega_k t + \varphi_k), \tag{1}$$

where the $a_k(t)$ are polynomials. Attempts to solve this problem using the least square method meet substantial difficulties [1]. The only methods that really work in practical application are those that extensively use additional information about the form and the number of terms in the sum in (1) (methods of polynomial approximation, spectral analysis, etc.). One of the first important examples of a realization of such an approach is the solution of the problem about the modeling of the gas expansion laws using sums of damping exponents. This solution was found by Gaspard Riche de Prony back in 1795. His method essentially uses the fact that exponents are eigenfunctions of the shift operator $t \to t + \Delta t$.

In the present paper we develop a method of multidimensional unfolding that enables us to use statistical analysis of the original time series in order to estimate the contribution of various terms in (1), and, most important, their number, prior to applying complicated approximation algorithms.

A multidimensional (n-dimensional) unfolding of the time series $f = (f_1, \dots, f_N)$ is a piecewise-linear curve X_f in the n-dimensional Euclidean space $\mathbb{R}^n$ that consecutively joins the vectors $X_q \in \mathbb{R}^n$, $q = 1, \dots, p = N - n + 1$, determined by length n segments $X'_q = (f_q, \dots, f_{q+n-1})$ of the original series. By going from the time series f to its n-dimensional unfolding we can use the geometry of unfolding for the analysis of f.

For example, the analysis of one- and two-dimensional projections of the curve X_f using the projection pursuit method [2] enables us to get an idea about the character of distinctions between projections of the unfolding of the original series and projections of unfoldings of various components of the discrete model expression (1).

We say that the n-rank of the time series f does not exceed r if there exists an r-dimensional plane $L \subset \mathbb{R}^n$ such that $X_f \subset L$. Below we prove the following results.

0065-9290/94/$1.00 + $.25 per page

- Let $N_1 \geq N + n$. The series $f = (f_1, \ldots, f_N)$ can be completed to a series $\widetilde{f} = (f_1, \ldots, f_N, \widetilde{f}_{N+1}, \ldots, \widetilde{f}_{N_1})$ whose n-rank does not exceed $r < n$ if and only if

$$f_m = \sum_{s=1}^{l} c_{l-s+1} f_{m-s}, \qquad l \leq r,$$

for all $m = l + 1, \ldots, N$.

- Let g be the series $g = (g_1, \ldots, g_N)$, where $g_q = g((q-1)\Delta t)$ for a function $g(t)$ of the form (1). Then the n-rank of g does not exceed r for all $n > r$, N, and Δt if and only if $g(t)$ is a solution of an ordinary differential equation

$$\sum_{q=0}^{r} b_q \frac{d^q}{dt^q} g(t) = c \tag{2}$$

with constant coefficients $b_0, \ldots, b_r, c$.

Let L be an r-dimensional plane in $\mathbb{R}^n$. For a given time series f denote by $X_f(L)$ the orthogonal projection of the unfolding X_f to L. It is clear that $X_f(L)$ is a piecewise linear curve in $\mathbb{R}^n$ with nodes $(X_1(L), \ldots, X_p(L))$. Denote

$$\rho_n(f, L) = \|X_f - X_f(L)\|^2 = \frac{1}{p} \sum_{q=1}^{p} \|X_q - X_q(L)\|^2,$$

where $\|\cdot\|$ is the standard Euclidean metric in $\mathbb{R}^n$. Let us consider the Grassmann manifold $G'(r, n)$ of all r-dimensional planes in $\mathbb{R}^n$. Using the n-dimensional unfolding X_f, we obtain the function

$$\rho_n(f)\colon G'(r, n) \to \mathbb{R}^1, \quad \rho_n(f)(L) = \rho_n(f, L).$$

Denote

$$r_n(f) = \inf_{L \in G'(r,n)} \rho_n(f, L).$$

Below we use the principal component method **[2]** to compute $r_n(f)$ and to describe the set $\mathscr{L}_n(f, r)$ of all r-dimensional planes where the function $\rho_n(f)$ takes the value exactly $r_n(f)$. We suggest an algorithm for the construction of a time series $f(L, r)$ for each piecewise linear curve $X_f(L)$ and prove that for a time series $f_*(r) = f(L_*, r)$, $L_* \in \mathscr{L}_n(f, r)$, the following theorem holds.

If $g(t)$ is a solution of the equation (2), *then*

$$\|f - g\|_\mu^2 \geq \|f - f_*(r)\|_\mu^2,$$

where μ is a natural Euclidean metric (to be defined below) in the space T of all time series T with N samples (or marks).

Therefore, $\|f - f_*(r)\|_\mu^2 = \delta_r$ gives a *lower* bound to the quality of the approximation of the series f in the class of functions of the form (1) satisfying equations of the form (2) without applying algorithms for estimation of the parameters λ_k, ω_k, φ_k, and the coefficients of the polynomials $a_k(t)$. In particular, we obtain a lower bound to the order of the equation (2) if we need an a priori approximation error of order less than a given threshold δ.

Additional possibilities occur in the visual analysis of deviations of values of the series f from the values of the series $f(r)$, thus allowing us to select terms in the

model (1). This is especially important, since in general parameters in this model are dependent.

The theory described in this paper is realized as a software package that is used to solve the following problems of ecological monitoring:

(1) Modeling of time series (construction of a model of a time series using sampling data).

(2) Description of the dynamics of sampling anomalies (deviation from model times series).

(3) Detection of structural changes of a time series (dynamics of parameters of model times series)

Practical results are described in [3–5]. Some of these results are presented in Figures 1–4 to illustrate the main ideas of the paper.

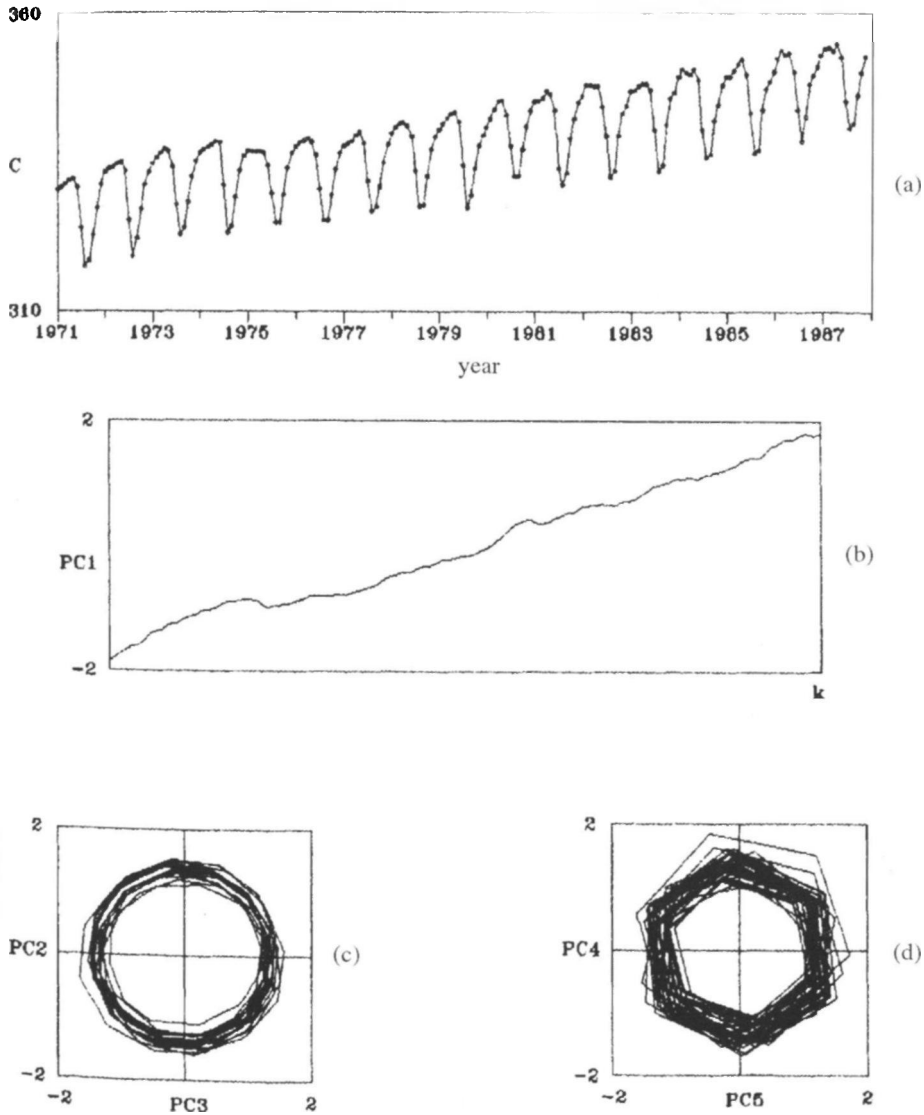

FIGURE 1

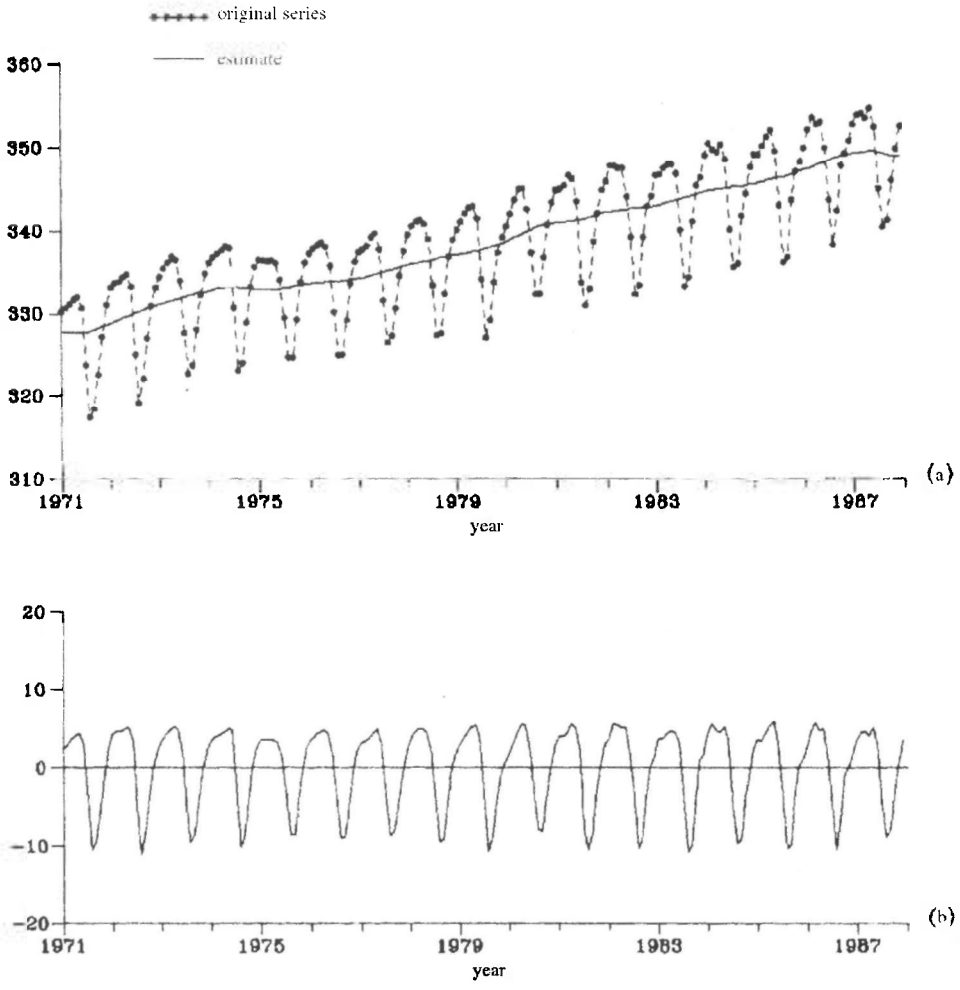

FIGURE 2

§1. Main definitions and general results

Let $f(t)$ be the time series under investigation, $f(t) = (f_1, \dots, f_N)$, where $f_q = f((q-1)\Delta t)$, $q = 1, \dots, N$.

DEFINITION 1. A piecewise linear curve $X_f \subset \mathbb{R}^n$ with nodes $X_1, \dots X_p$, where $X_q' = (f_q, \dots, f_{q+n-1})$, is called the *n-dimensional unfolding* of the time series $f(t)$.

Here and later by a vector $X \in \mathbb{R}^n$ we mean a column vector, and by X' we denote the corresponding row vector.

First, let us consider n-dimensional unfoldings of model time series. Without loss of generality below we assume that $\Delta t = 1$. 1. $g(t) = a^t$, *where* $a > 0$. We have $a^{t+\tau} = a^t a^\tau$, so that

$$X_1' = (1, \dots, a^{n-1}), \dots, X_q' = (a^{q-1}, \dots, a^{q+n-2}) = a^{q-1} X_1'.$$

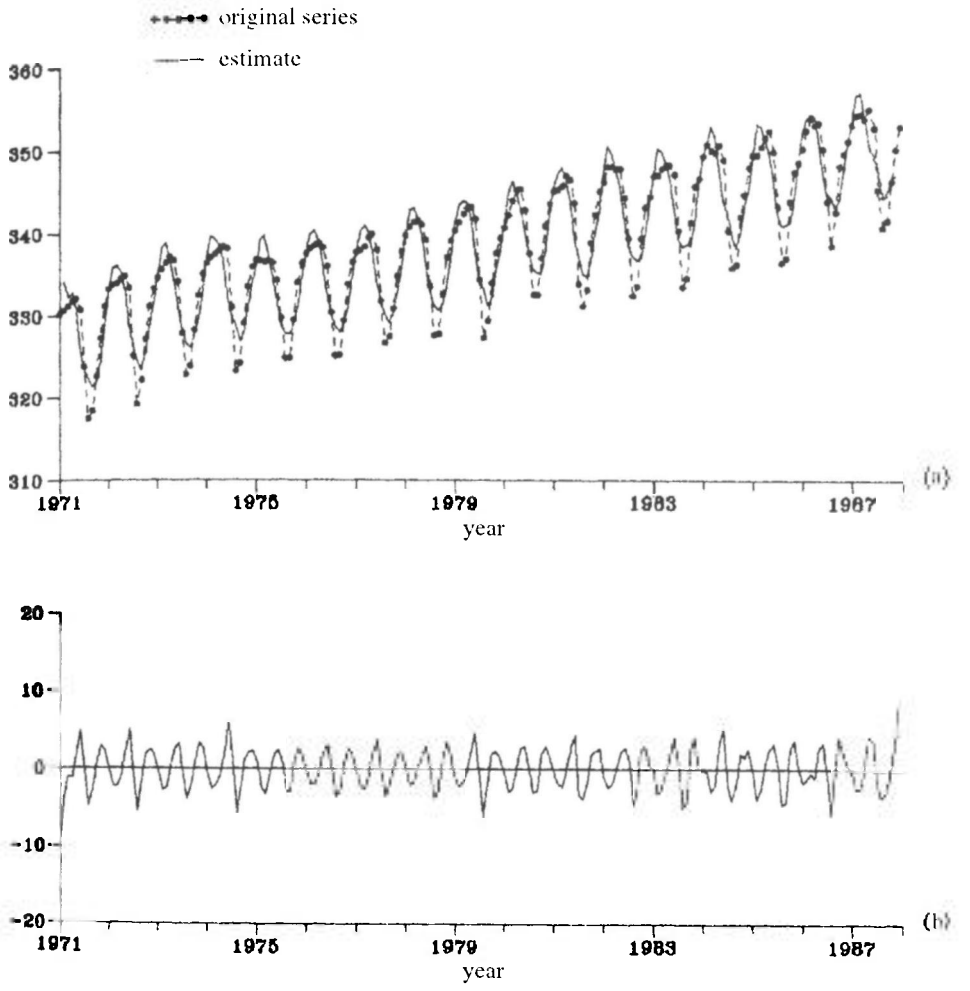

FIGURE 3

Therefore for all n and N we have $X_{g(t)} \subset L \subset \mathbb{R}^n$, where L is the line in $\mathbb{R}^n$ with the director vector $X_1/\|X_1\|$.

2. $g(t) = \sin(\omega t + \varphi)$, *where* ω *and* φ *are parameters.* We have

$$g(t+\tau) = g(t)\cos(\omega\tau) + g(t+\pi/(2\omega))\sin(\omega\tau).$$

Denoting

$$Y_1' = (1, \cos\omega, \ldots, \cos\omega(n-1)), \qquad Y_2' = (0, \sin\omega, \ldots, \sin\omega(n-1))$$

we have

$$X_q = \sin(\omega(q-1)+\varphi)\,Y_1 + \cos(\omega(q-1)+\varphi)\,Y_2, \qquad q = 1, \ldots, p.$$

Therefore $X_{g(t)} \subset L \subset \mathbb{R}^n$, where L is the two-dimensional plane spanned by the vectors Y_1 and Y_2.

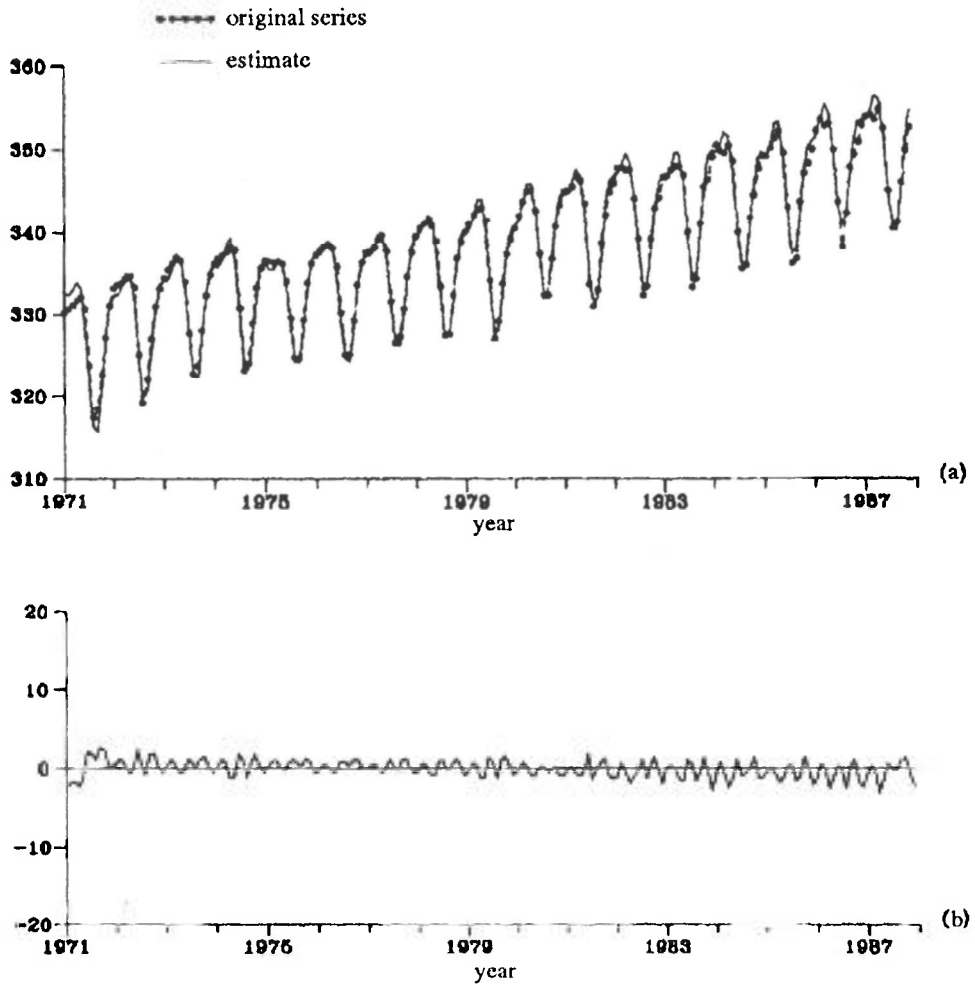

FIGURE 4

Let us note that

$$\langle Y_1, Y_2 \rangle = \sum_{m=0}^{n-1} \cos(m\omega)\sin(m\omega)$$

$$= \frac{1}{2}\sum_{m=0}^{n-1} \sin(2m\omega) = \tfrac{1}{2}\csc(\omega)\sin(n\omega)\sin((n-1)\omega).$$

In particular, in the case $\omega = 2\pi/n$ the vectors Y_1 and Y_2 are orthogonal and for $N \geq 2n + 1$ the unfolding X_g is the right n-gon in the plane L.

In Figure 1 projections of the 12-dimensional unfolding of the actual time series are presented. Figure 1c indicates the existence of a cycle with $\omega = 2\pi/12$, and

Figure 1d indicates the existence of a cycle with $\omega = 2\pi/6$. Deviation of the projection of the curve from the right 12-gon and the right 6-gon allows us to judge anomalies of sampling data.

3. $g(t) = \sum_{k=0}^{m} b_k t^k$. We have

$$g(t+\tau) = \sum_{k=0}^{m} g^{(k)}(t)\frac{\tau^k}{k!},$$

where

$$g^{(k)}(t) = \frac{d}{dt}g^{(k-1)}(t), \qquad k \geq 1.$$

Denote

$$Y_1' = (1, 1, \ldots, 1),$$
$$Y_k' = (0, 1/(k-1)!, \ldots, (n-1)^{k-1}/(k-1)!), \qquad k = 2, \ldots, m+1.$$

Since $g^{(m)}(t) = m!b_m = \text{const}$, we get

$$X_q = \sum_{k=0}^{m-1} g^{(k)}(q-1)Y_{k+1} + m!b_m Y_{m+1},$$

so that for $n \geq m$ we have $X_{g(t)} \subset L \subset \mathbb{R}^n$, where L is an m-dimensional subspace containing the vectors $Y_1, \ldots, Y_{m+1}$.

A common feature of all these model time series is that the corresponding n-dimensional unfoldings lie in some subspaces L whose dimension does not depend on n and N. These model time series belong to the class of functions that possess an addition formula of the form

$$g(t+\tau) = \sum_{k=0}^{m} \varphi_k(t)\psi_k(\tau) \tag{3}$$

for some m and some functions $\varphi_k(t)$, $\psi_k(t)$. Denote

$$Y_k' = (\psi_k(0), \ldots, \psi_k(n-1)).$$

Then the addition formula for $g(t)$ shows that the nodes of the n-dimensional unfolding $X_{g(t)}$ can be written in the form

$$X_q = \sum_{k=0}^{m} \varphi_k(q-1)Y_k,$$

so that if one of the functions $\varphi_k(t)$ is a constant, say $\varphi_0(t) \equiv 1$, then $X_{g(t)} \subset L \subset \mathbb{R}^n$, where L is a subspace in $\mathbb{R}^n$ of dimension not exceeding m.

Let us consider now the case when a function $g(t)$ with the addition formula (3) has m derivatives and the functions $\psi_k(t)$, $k = 1, \ldots, m$, also have m derivatives.

Differentiating (3) l times with respect to τ and substituting $\tau = 0$, we get

$$g^{(l)}(t) = \sum_{k=0}^{m} \varphi_k(t)\psi_k^{(l)}(0).$$

Therefore, under the condition $\varphi_0(t) \equiv 1$ formula (3) shows that the $m+1$ functions $g(t) - \psi_0(0)$, $g'(t) - \psi_0'(0), \ldots, g^{(m)}(t) - \psi_0^{(m)}(0)$ belong to the m-dimensional linear

space generated by the functions $\varphi_k(t)$, $k = 1, \ldots, m$, i.e., are linearly dependent. Therefore, there exist constants $b_0, \ldots, b_m$ such that

$$\sum_{q=0}^{m} b_q (g^{(q)}(t) - \psi_0^{(q)}(0)) \equiv 0$$

Denoting

$$\mathfrak{D} = \sum_{q=0}^{m} b_q \frac{d^q}{dt^q},$$

we get $\mathfrak{D}g(t) = c$, where $c = \sum_{q=0}^{m} b_q \psi_0^{(q)}(0)$. The theory of ordinary differential equations shows that $g(t)$ must be of the form (1).

Therefore the class of functions of the form (1) coincides with the class of functions that have sufficiently many derivatives and possess the addition formula (3).

An arbitrary piecewise linear curve Y in $\mathbb{R}^n$ with nodes $Y_1, \ldots, Y_p$ is described by a table

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1p} \\ y_{21} & \cdots & y_{2p} \\ \cdots & \cdots & \cdots \\ y_{n1} & \cdots & y_{np} \end{pmatrix}$$

where $Y_q' = (y_{1q}, \ldots, y_{pq}) \in \mathbb{R}^n$.

DEFINITION 2. We say that *the rank of a curve Y does not exceed r* if there exists an r-dimensional plane $L \subset \mathbb{R}^n$ such that $Y_q \in L$ for all $q = 1, \ldots, p$.

We note that the definition of the rank of a curve does not depend on the choice of a basis in $\mathbb{R}^n$.

Let us consider the scattering matrix $T(Y)$ of the curve Y:

$$T(Y) = \sum_{q=1}^{p} (Y_q - \overline{Y})(Y_q - \overline{Y})',$$

where

$$\overline{Y} = \frac{1}{p} \sum_{q=1}^{p} Y_q.$$

The principal components theory [**2**] immediately implies the following result.

LEMMA 1. *The rank of the curve Y does not exceed r ($\operatorname{rk} Y \leq r$) if and only if $\lambda_s = 0$ for all $s > r$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues of the scattering matrix $T(Y)$.*

DEFINITION 3. We say that *the n-rank of the time series $f(t)$ does not exceed r* if $\operatorname{rk} X_f \leq r$, where X_f is the n-dimensional unfolding of $f(t)$.

To the unfolding X_f of a time series f there corresponds the table $X_d = (x_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq p$, with $x_{ij} = f_{i+j-1}$, so that X_f is a Hankel matrix [**6**]. Evidently, this special form of the table X_f is not preserved under coordinate changes. We describe an approach to the analysis of properties of the unfolding $X_f \subset \mathbb{R}^n$, hence also of the series f, that essentially uses a fixed ordering of coordinates in the Euclidean space $\mathbb{R}^n$.

We need the notion of the Schubert symbol, which plays the central role in the construction of a cell decomposition of Grassmann manifolds. We will follow the book [7]. Consider in $\mathbb{R}^n$ the chain of subspaces $\mathbb{R}^0 \subset \mathbb{R}^1 \subset \cdots \subset \mathbb{R}^k \subset \cdots \subset \mathbb{R}^n$, where $\mathbb{R}^k$ consists of all vectors of the form $(y_1, \ldots, y_k, 0, \ldots, 0)$. Each l-dimensional plane $W \subset \mathbb{R}^n$ passing through the origin define a sequence of numbers

$$0 \leq \dim(W \cap \mathbb{R}^1) \leq \dim(W \cap \mathbb{R}^2) \leq \cdots \leq \dim(W \cap \mathbb{R}^n) = l.$$

Evidently, two consecutive numbers in this sequence differ by at most 1. Therefore, the nondecreasing function

$$\varphi(k) = \dim(W \cap \mathbb{R}^k), \qquad k = 0, \ldots, n,$$

has exactly l discontinuity points. Therefore, for a given plane W we get a sequence of numbers $1 \leq \sigma_1 < \sigma_2 < \cdots < \sigma_l \leq n$, where σ_i is uniquely determined from the conditions

$$\varphi(\sigma_i) = i, \quad \varphi(\sigma_i - 1) = i - 1.$$

This sequence is called the *Schubert symbol* of the plane $W \subset \mathbb{R}^n$. The following general result holds.

An l-dimensional plane $W \subset \mathbb{R}^n$ has a Schubert symbol $\sigma_1 < \sigma_2 < \cdots < \sigma_l$ if and only if it possesses an orthogonal basis $W_1, \ldots, W_l$ such that $W_k \in \mathbb{R}^{\sigma_k}$ and the σ_kth coordinate of the vector W_k equals -1 for all $k = 1, \ldots, l$.

The indicated basis in the plane W is determined uniquely, and we will call vectors W_k of this basis the *Schubert vectors* of the plane W.

Now let us consider a time series f whose n-rank does not exceed $r \leq n - 1$. Then there exists a r-dimensional plane L such that $X_q \in L$ for all $q = 1, \ldots, p = N - n + 1$. Any vector $X \in L$ can be uniquely represented in the form $X = \widetilde{X} + \xi$, where $\widetilde{X}$ is the orthogonal projection of X to the r-dimensional plane $\widetilde{L}$ which is parallel to L and passes through the origin $0 \in \mathbb{R}^n$, and ξ is a vector that is orthogonal to L and does not depend on X. Denote by $L^\perp$ the $(n - r)$-dimensional subspace that is orthogonal to $\widetilde{L}$. Associate to $L^\perp$ its Schubert symbol $\sigma_1 < \sigma_2 < \cdots < \sigma_{n-r}$ and the corresponding orthogonal basis $W_1, \ldots, W_{n-r}$ of Schubert vectors. By construction, for any Schubert vector $W_k = (w_{k1}, \ldots, w_{k,\sigma_k - 1}, -1, 0, \ldots, 0)$ we have

$$\langle X_q, W_k \rangle = \sum_{j=1}^{\sigma_k - 1} w_{kj} f_{q+j-1} - f_{q+\sigma_k - 1} + w_{k0} = 0,$$

where $w_{k0} = -\langle \xi, W_k \rangle$. Denote $q + \sigma_k - 1 = m$, $q + j - 1 = m - s$. We obtain

$$f_m = \sum_{s=1}^{\sigma_k - 1} w_{k,\sigma_k - s} f_{m-s} + w_{k0}, \qquad m = \sigma_k, \ldots, p_k, \quad k = 1, \ldots, n - r, \tag{7}$$

where $p_k = N - n + \sigma_k$. For the Schubert symbol we have

$$n - r - 1 \leq \sum_{k=1}^{n-r-1} (\sigma_{k+1} - \sigma_k) = \sigma_{n-r} - \sigma_1 \leq n - \sigma_1,$$

so that $\sigma_1 \leq r + 1$. Therefore, we have proved that if the n-rank of a time series f

does not exceed $r \leq n-1$, then there exist constants $c_0, c_1, \ldots, c_l$, $l \leq r$, such that

$$f_m = \sum_{s=1}^{l} c_{l-s+1} f_{m-s} + c_0, \qquad m = l+1, \ldots, N-n+l+1. \tag{8}$$

Indeed, it suffices to take $(c_1, \ldots, c_l, -1, 0, \ldots, 0) = W_1$ and $l = \sigma_1 - 1$ for any r-dimensional plane L that contains the unfolding X_f.

Formula (8) allows us to compute the numbers $\widehat{f}_m$ for $m = N-n+l+2, \ldots, N$. In general, this values should not coincide with the corresponding terms of the series f_m. The following simple example allows a better understanding of the situation.

Consider the series $(0, 1, 2, 3, 4, 2)$. Its three-dimensional unfolding has the nodes

$$X_1 = (0,1,2), \quad X_2 = (1,2,3), \quad X_3 = (2,3,4), \quad X_4 = (3,4,2).$$

Simple computations show that the 3-rank of this series equals 2, the Schubert symbol of the line in R^3 which is orthogonal to the two-dimensional support L of the unfolding is $\sigma_1 = 2$, and the corresponding Schubert vector is $(1, -1, 0)$. According to (8), we have

$$f_m = f_{m-1} + 1 \quad \text{for } m = 2,3,4,5, \qquad f_6 \neq \widehat{f}_6 = f_5 + 1.$$

The time series $\widetilde{f} = (\widetilde{f}_m)$ of length $\widetilde{N} > N$ is called *the extension of the series* $f = (f_m)$ *of length* N if $\widetilde{f}_m = f_m$ for all $m \leq N$.

In the assumptions and notation of (8) one can easily see that if

$$f_m \neq \sum_{s=1}^{l} c_{l-s+1} f_{m-s} + c_0$$

for $m = N-n+l+2 \leq N$, then the n-rank of *any* extension $\widetilde{f}$ of the series f is greater than r. The possibility of extending a time series without increasing the n-rank is described in the following lemma.

LEMMA 2. *Let $r \leq n-1$ and $N_1 > N+n$. A given series $f = (f_1, \ldots, f_N)$ admits an extension $\widetilde{f} = (\widetilde{f}_1, \ldots, \widetilde{f}_{N_1})$ with the n-rank not exceeding r if and only if there exist constants $c_0, c_1, \ldots c_l$, $l \leq r$, such that*

$$f_m = \sum_{s=1}^{l} c_{l-s+1} f_{m-s} + c_0$$

for all $m = l+1, \ldots, N$.

PROOF. Let $\widetilde{f}$ be an extension of the series f with the n-rank of $\widetilde{f}$ not exceeding r. Then there exist constants $c_0, \ldots, c_l$ that provide the representation (8) for the series $\widetilde{f}$. As was shown earlier, this formula holds for all m such that $l+1 \leq m \leq N_1 - n + l + 1$. In particular, it holds for $l+1 \leq m \leq N$, since, by the conditions of the lemma, $N < N_1 - n + l + 1$. Conversely, let the required formula hold for all $m = l+1, \ldots, N$. Using this formula, we construct an extension $\widehat{f}$ of length $\widehat{N}$ and prove that for each $\widehat{N} > N$ the n-rank of this extension does not exceed r.

Consider the n-dimensional unfolding $X_{\tilde{f}} = (X_1, \ldots, X_p, \ldots)$. Formula (8) implies the formula

$$X_q = \sum_{s=1}^{l} c_{l-s+1} X_{q-s} + c_0, \qquad q \geq l+1. \tag{9}$$

We have

$$X_{q+1} - X_q = c_l X_q + \sum_{s=1}^{l-1} (c_{l-s} - c_{l-s+1}) X_{q-s} - c_1 X_{q-l}.$$

Therefore,

$$X_{q+1} = \sum_{k=0}^{l} \lambda_k X_{q-k},$$

where $\lambda_0 = 1 + c_l$, $\lambda_k = (c_{l-k} - c_{l-k+1})$ for $1 \leq k \leq l-1$, and $\lambda_l = -c_1$. Since $\sum_{k=0}^{l} \lambda_k = 1$, for any $q \geq l+1$ the node X_{q+1} of the unfolding $X_{\tilde{f}}$ lies in the l-dimensional subspace spanned by the first $l+1$ nodes $X_1, \ldots, X_{l+1}$. Since $l \leq r$, the n-rank of $\tilde{f}$ does not exceed r. The lemma is proved. □

For the above examples of model time series

$$a^t, \quad \sin(\omega t + \varphi), \quad P_s(t) = \sum_{l=0}^{s} b_l t^l$$

we have for arbitrary Δt, n, and N

$$\operatorname{rk}_n a^t \leq 1 \qquad \text{for any } a > 0,$$

$$\operatorname{rk}_n \sin(\omega t + \varphi) \leq 2 \qquad \text{for any } \omega \text{ and } \varphi,$$

$$\operatorname{rk}_n P_s(t) \leq s \qquad \text{for any } b_0, \ldots, b_s.$$

DEFINITION 4. We say that the *absolute rank* of a continuous time signal f does not exceed r if for any Δt, n, and N the rank of the corresponding time series does not exceed r.

As we have shown earlier, the absolute rank of the continuous time signal $f(t)$ does not exceed r if and only if $f(t)$ is a solution of an ordinary differential equation

$$\sum_{k=0}^{r} b_k \frac{d^k f(t)}{dt^k} = C,$$

where C and b_k, $k = 1, \ldots, r$, are some constants.

Using the notion of the absolute rank of a time series $f(t)$ we can introduce, for fixed Δt, n, N, a filtration in the set $\mathfrak{MD}$ of real solutions of ordinary differential equations with constant coefficients as follows:

$$\mathfrak{MD}_1 \subset \mathfrak{MD}_2 \subset \cdots \subset \mathfrak{MD}_n = \mathfrak{MD},$$

where $\mathfrak{MD}_r$ is the set of time signals with absolute rank not exceeding r.

§2. Nonparametric modeling of time series from projections of their unfoldings

Consider the set $\mathfrak{M}(p,n) = \mathfrak{M}$ of all piecewise linear curves with p nodes in the space $\mathbb{R}^n$. Introduce in $\mathfrak{M}$ the structure of the Euclidean space $\mathbb{R}^{pn}$ in such a way that if $Y_1 = (Y_{11}, \dots, Y_{1p})$, and $Y_2 = (Y_{21}, \dots, Y_{2p})$ are two curves, then

$$\|Y_1 - Y_2\|^2 = \frac{1}{p}\sum_{q=1}^{p} \|Y_{1q} - Y_{2q}\|^2,$$

where $\|\cdot\|^2$ is the standard Euclidean distance in $\mathbb{R}^n$.

In the space $\mathfrak{M} \sim \mathbb{R}^{pn}$ we have the filtration $\mathfrak{M}_1 \subset \mathfrak{M}_2 \subset \cdots \subset \mathfrak{M}_n = \mathfrak{M}$, where $\mathfrak{M}_r$ is the set of curves with rank not exceeding r.

DEFINITION 5. The curve $X(r) \in \mathfrak{M}_r$ is called a *projection* of a given curve $X \in \mathfrak{M}$ to $\mathfrak{M}_r$ if

$$\|X - X(r)\|^2 = \min_{Y \in \mathfrak{M}_r} \|X - Y\|^2.$$

The theory of principal components [**2**] gives the following theorem.

THEOREM 1. *For any curve $X \in \mathfrak{M}$ we have*

$$X(r) = \overline{X} + V(r)V'(r)X,$$

where $\overline{X}$ is a constant curve $(\overline{X}, \dots, \overline{X})$ and $V(r)$ is the matrix composed of column vectors $V_1, \dots, V_r$, where $V_s \in \mathbb{R}^n$ are the eigenvectors of the scattering matrix $T(X)$ with eigenvalues λ_s ordered in such a way that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq \cdots \geq \lambda_n \geq 0$.

PROOF. Denote

$$\rho(X, \mathfrak{M}_r) = \min_{Y \in \mathfrak{M}_r} \|X - Y\|^2.$$

Consider the following function $\rho(X, L)$ on the Grassmannian $G'(r,n)$ of all r-dimensional planes in $\mathbb{R}^n$:

$$\rho(X, L) = \|X - X(L)\|^2 = \frac{1}{p}\sum_{q=1}^{p} \|X_q - X_q(L)\|^2,$$

where $X_q(L)$ is the orthogonal projection of the node X_q of the curve X to the plane L and $X(L)$ is the curve with nodes $(X_1(L), \dots, X_p(L))$. It is clear that

$$\rho(X, \mathfrak{M}_r) = \min_{L \in G'(r,n)} \rho(X, L).$$

Let $G(r,n)$ be the Grassmannian of all r-dimensional planes in $\mathbb{R}^n$ passing through the origin 0. Recall that $G(r,n)$ is a compact smooth manifold of dimension $r(n-r)$. There exists a canonical map $G'(r,n) \to G(r,n)$, which sends the plane L to the unique plane $\widetilde{L}$ that is parallel to L and passes through the origin. Using this map we can identify $G'(r,n)$ with the manifold of pairs $(\widetilde{L}, \xi) \in G(r,n) \times \mathbb{R}^n$, where $\xi \in \mathbb{R}^n$ is

a vector that is orthogonal to $\tilde{L}$. For $L = (\tilde{L}, \xi)$ denote by $L^{\perp}$ the plane which is orthogonal to $\tilde{L}$. We have

$$X_q(L) = X_q(\tilde{L}) + \xi, \qquad \overline{X} = \frac{1}{p}\sum_{q=1}^{p} X_q = \overline{X}(\tilde{L}) + \overline{X}(L^{\perp}).$$

Therefore

$$(X_q - X_q(L)) = [(X_q - \overline{X}) - (X_q - \overline{X})(\tilde{L})] + (\overline{X}(L^{\perp}) - \xi). \tag{10}$$

Substituting (10) into the formula for $\rho(X, L)$, we obtain

$$\rho(X, L) = \rho(X - \overline{X}, \tilde{L}) + \|\overline{X}(L^{\perp}) - \xi\|^2.$$

Let us remark that by $\overline{X}$ we denote both the average vector of the curve X and the constant curve $(\overline{X}, \ldots, \overline{X})$. Therefore,

$$\rho(X, (\tilde{L}, \overline{X}(L^{\perp})) = \rho(X - \overline{X}, \tilde{L}) \leq \rho(X, (\tilde{L}, \xi))$$

for all curves X, all r-dimensional planes $\tilde{L}$, and all vectors $\xi \in \tilde{L}$. Therefore,

$$\rho(X, \mathfrak{M}_r) = \min_{\tilde{L} \in G(r,n)} \rho(X - \overline{X}, \tilde{L}).$$

Choose an orthonormal basis in the subspace $L^{\perp} \subset \mathbb{R}^n$ and form the matrix $P(\tilde{L})$ whose columns are vectors of this basis. By construction,

$$\|(X_q - \overline{X}) - (X_q - \overline{X})(\tilde{L})\|^2 = \|P(\tilde{L})'(X_q - \overline{X})\|^2.$$

Using the identity

$$\|P(\tilde{L})'(X_q - \overline{X})\|^2 = \operatorname{tr}(P(\tilde{L})'(X_q - \overline{X})(X_q - \overline{X})'P(\tilde{L})),$$

where $\operatorname{tr}(\cdot)$ is the trace of the matrix, we see that

$$\rho(X - \overline{X}, \tilde{L}) = \operatorname{tr} P(\tilde{L})'T(X)P(\tilde{L}). \tag{12}$$

The scattering matrix $T(X)$ is symmetric and nonnegative definite. In $\mathbb{R}^n$ we consider an orthonormal basis formed by eigenvectors $V_1, \ldots, V_n$ of the matrix $T(X)$ ordered by the decreasing of the corresponding eigenvalues.

Let $\tilde{L}_*$ be the r-dimensional plane with the basis formed by vectors $V_1, \ldots, V_r$. Then the vectors $V_{r+1}, \ldots, V_n$ form a basis in $L_*^{\perp}$. Denote by $V(r)$ the matrix formed by the column vectors $V_1, \ldots, V_r$, and by $V(n-r)$ the matrix $P(\tilde{L}_*)$ formed by the columns $V_{r+1}, \ldots, V_n$. By construction,

$$T(X)P(\tilde{L}_*) = P(\tilde{L}_*)\Lambda(n-r),$$

where $\Lambda(n-r)$ is the diagonal matrix with entries $\lambda_{r+1}, \ldots, \lambda_n$. Therefore,

$$\rho(X - \overline{X}, \tilde{L}_*) = \sum_{k=1}^{n-r} \lambda_{r+k}.$$

Now standard methods of linear algebra using formula (12) written in the basis $V_1, \ldots, V_n$ easily show that

$$\min_{\tilde{L} \in G(r,n)} \rho(X - \overline{X}, \tilde{L}) = \rho(X - \overline{X}, \tilde{L}_*).$$

The theorem is proved. □

Denote $\sum_{k=1}^{n-r} \lambda_{k+r} = r_n(X)$, and consider the set of planes

$$\mathscr{L}_n(X;r) = \{\widetilde{L} \in G(r,n) : \rho(X - \overline{X}, \widetilde{L}) = r_n(X)\}.$$

If $\lambda_r > \lambda_{r+1}$, then $\mathscr{L}_n(X;r)$ consists of a single plane $\widetilde{L}_*$. If $\lambda_r = \lambda_{r+1}$, we can find $k_1 \geq 0$ and $k_2 \leq n$ such that $k_1 < r < k_2$ and

$$\lambda_{k_1} \neq \lambda_r, \quad \lambda_{k_1+1} = \lambda_r, \quad \lambda_{k_2} = \lambda_r, \quad \lambda_{k_2+1} \neq \lambda_r.$$

Let $W_1 \subset \mathbb{R}^n$ be the k_1-dimensional subspace with the basis $V_1, \ldots, V_{k_1}$, and $W_2 \subset \mathbb{R}^n$ the k_2-dimensional subspace with the basis $V_1, \ldots, V_{k_2}$. Then

The set $\mathscr{L}_n(X;r)$ consists of all r-dimensional subspaces in $\mathbb{R}^n$ that contain the subspace $W_1 \subset W_2$ and are contained in W_2.

Therefore, the projection of a curve X to $\mathfrak{M}_r$ is determined uniquely if and only if $\lambda_r > \lambda_{r+1}$. If $\lambda_r = \lambda_{r+1}$, the role of the projection can be played by any curve of the form $\overline{X} + X(\widetilde{L})$ with $\widetilde{L} \in \mathscr{L}_n(X;r)$.

Let us note that for $X = X_f$ we gave the description of the set $\mathscr{L}_n(f;r)$ mentioned in the Introduction.

We identify the space of all time series with N samples with a linear space of dimension N given with a fixed basis.

DEFINITION 6. A time series $g_Y = (g_1, \ldots, g_N)$ is called a *projection* of a given curve $Y \in \mathfrak{M}$ to the space of time series T if

$$\|Y - X_{g_Y}\|^2 = \min_{f \in T} \|Y - X_f\|^2,$$

where X_{g_Y} and X_f are the n-dimensional unfoldings of time series g_Y and f.

THEOREM 2. (a) *For a curve $Y \in \mathfrak{M}$, its projection g_Y to the space of time series T is of the form $g_Y = (g_1, \ldots, g_N)$, where*

$$g_s = \begin{cases} \dfrac{1}{s} \displaystyle\sum_{l=1}^{s} y_{l,s-l+1}, & 1 \leq s \leq n, \\ \dfrac{1}{n} \displaystyle\sum_{l=1}^{n} y_{l,s-l+1}, & n \leq s \leq p, \\ \dfrac{1}{N-s+1} \displaystyle\sum_{l=1}^{N-s+1} y_{l+s-p,p-l+1}, & p \leq s \leq N. \end{cases}$$

Recall that $p = N - n + 1$.

(b) *Let X_f be the n-dimensional unfolding of a time series f. Then for any curve $Y \in \mathfrak{M}$ we have*

$$\|X_f - Y\|^2 = \|X_f - X_{g_Y}\|^2 + \|X_{g_Y} - Y\|^2.$$

Now we consider the following distance in the space of time series T. Let $f = (f_1, \ldots, f_N)$ and $g = (g_1, \ldots, g_N)$ be two time series. For a fixed n set

$$\|f - g\|_\mu^2 = \frac{1}{N} \sum_{s=1}^{N} \mu_s (f_s - g_s)^2,$$

where

$$\mu_s = \begin{cases} s/n, & 1 \leq s \leq n, \\ 1, & n \leq s \leq p, \\ (N-s+1)/n, & p \leq s \leq N \end{cases}$$

Let us comment on the choice of the distance in the space of time series T. Passing from a series f to its n-dimensional unfolding X_f, we get a linear map $i: T \to \mathfrak{M}$, $i(f) = X_f$. Choosing the standard Euclidean metric $\|Y_1 - Y_2\|$ in $\mathfrak{M} \sim \mathbb{R}^{pn}$, we define a metric in T in such a way that the imbedding i preserves the distance up to a factor. In this case the form of the weight function is uniquely determined by the condition

$$\|f - g\|_\mu^2 = (p/nN)\|X_f - X_g\|^2.$$

To conclude this comment, we present a construction of a class of metrics in the space T of time series that satisfy all principal results of this paper. We choose systems of positive weights $\alpha = (\alpha_1, \ldots, \alpha_n)$, $\beta = (\beta_1, \ldots, \beta_p)$, and define the following Euclidean metrics in $\mathbb{R}^n$ and in $\mathfrak{M}$:

$$\|Y_1 - Y_2\|_\alpha^2 = \sum_{k=1}^{n} \alpha_k (y_{k1} - y_{k2})^2, \qquad Y_1, Y_2 \in \mathbb{R}^n,$$

$$\|X - Y\|_{\alpha,\beta}^2 = \sum_{q=1}^{p} \beta_q \|X_q - Y_q\|_\alpha^2,$$

where $X = (X_q)$, $Y = (Y_q) \in \mathfrak{M}$. Consider the characteristic polynomials of weight systems α and β

$$\alpha(z) = \sum_{k=1}^{n} \alpha_k z^{k-1}, \qquad \beta(z) = \sum_{q=1}^{p} \beta_q z^{q-1}$$

and define the weight system $\gamma = (\gamma_1, \ldots, \gamma_N)$ corresponding to the polynomial $\gamma(z) = c\alpha(z)\beta(z)$, where $c = \text{const}$. We recall that $p = N - n + 1$, so that $(n-1) + (p-1) = N - 1$. Introducing in the space T of length N time series the metric

$$\|f - g\|_\gamma^2 = \sum_{m=1}^{N} \gamma_m (f_m - g_m)^2,$$

we see by a straightforward computation that

$$\|f - g\|_\gamma^2 = c\|X_f - X_g\|_{\alpha,\beta}^2.$$

DEFINITION 6′. A time series $g_Y = (g_1, \ldots, g_N)$ is called an (α, β)-*projection* of a curve $Y \in \mathfrak{M}$ to the time series space T if

$$\|Y - X_{g_Y}\|_{\alpha,\beta}^2 = \min_{f \in T} \|Y - X_f\|_{\alpha,\beta}^2.$$

THEOREM 2′. (a) *Let $g_Y = (g_1, \ldots, g_N)$ be an (α, β)-projection of a curve $Y \in \mathfrak{M}$. Then*

$$g_m = \frac{1}{\gamma_m} \sum \alpha_k \beta_q y_{kq}, \qquad m = 1, \ldots, N,$$

where the summation is over all k, q such that $k + q = m + 1$, $1 \le k \le n$, $1 \le q \le N - n + 1$.

(b) *Let $f = (f_1, \ldots, f_N)$ be the n-dimensional unfolding of a time series. Then for any curve $Y \in \mathfrak{M}$ we have*

$$\|X_f - Y\|_{\alpha,\beta}^2 = \|X_f - X_{g_Y}\|_{\alpha,\beta}^2 + \|X_{g_Y} - Y\|_{\alpha,\beta}^2.$$

Here g_Y is the (α, β)-projection of the curve Y described in (a).

The proof of Theorem 2′ is based on the following identity, which admits a direct verification. Let $\{x_j\}$ and $\{\lambda_j\}$, $j = 1, \ldots, J$, be two number sequences with $\sum_{j=1}^{J} \lambda_j = s_\lambda \neq 0$. Then for any number y we have

$$\sum_{j=1}^{J} \lambda_j (x_j - y)^2 = \sum_{j=1}^{J} \lambda_j (x_j - x_\lambda)^2 + s_\lambda (x_\lambda - y)^2,$$

where

$$x_\lambda = \frac{1}{s_\lambda} \sum_{j=1}^{J} \lambda_j x_j.$$

Let us note that, together with our main metric μ with the characteristic polynomial of the system of weights equal to $(1/N)\mu(z) = c\alpha(z)\beta(z)$, where

$$c = \frac{p}{nN}, \qquad \alpha(z) = \frac{1 - z^n}{1 - z}, \qquad \beta(z) = \frac{1}{p} \frac{1 - z^p}{1 - z},$$

a practically important metric is the metric γ with the characteristic polynomial $\gamma(z) = \alpha(z)\beta(z)$, where

$$\alpha(z) = \left(\frac{1 + z}{2}\right)^{n-1}, \qquad \beta(z) = \left(\frac{1 + z}{2}\right)^{p-1}.$$

Now we continue the exposition for the case of the metric μ.

For a fixed r and our time series f, denote by X_* the projection of the n-dimensional unfolding X_f to $\mathfrak{M}_r$ and by f_* the projection of X_* to T (see Definitions 1, 5, 6). Theorems 1 and 2 imply that the time series f_*, as a nonparametric model of the series f, has the following extermal properties.

COROLLARY 1. *For any curve* $Y \in \mathfrak{M}_r$,

$$\|f - f_*\|_\mu^2 + (p/nN)\|X_{f_*} - X_*\|^2 \leq \|f - g_Y\|_\mu^2 + (p/nN)\|X_{g_Y} - Y\|^2$$

COROLLARY 2. *If* g *is a time series of rank not exceeding* r, *then*

$$\|f - g\|_\mu^2 \geq \|f - f_*\|_\mu^2 + (p/nN)\|X_{f_*} - X_*\|^2.$$

The proof follows from the fact that if $Y = X_g$, then $g_Y = g$.

§3. Main results

THEOREM 3. *If the rank of a time series* g *does not exceed* r, *then for our time series* f *we have*

$$\|f - g\|_\mu^2 \geq \|f - f_*\|_\mu^2.$$

Therefore we have an algorithm for the construction of nonparametric approximations $f_*(r)$, $r = 1, \ldots, n - 1$, of a time series f (see Theorems 1 and 2), and can compute the errors

$$\delta_r = \|f - f_*(r)\|_\mu^2.$$

Theorem 3 implies that if $\delta_r > \delta$, where δ is some threshold, then for the approximation of the time series f by functions of the form

$$g(t) = \sum a_k(t) e^{\lambda_k t} \sin(\omega_k t + \varphi_k)$$

with $\operatorname{rk} g(t) \leq r$ we get

$$\|f - g\|_\mu^2 > \delta.$$

Hence, we have a lower bound for the approximation error even before applying algorithms for the estimation of the parameters λ_k, ω_k, φ_k and the coefficients of the polynomials $a_k(t)$, which in general is a hard nonlinear problem that can have a nonunique solution (see [**1,** Chapter 11]).

Let us note in conclusion that the bounds δ_r can be improved using iterations of the algorithm. Denote $f = f_{(0)}$, $f_{(1)} = f_*, \ldots, f_{(m)} = f_{(m-1)*}, \ldots$, Then we have the following improved estimate.

THEOREM 4. *If the rank of a time series g does not exceed r, then for the series f and for an arbitrary $m \geq 1$ we have*

$$\|f - g\|_\mu^2 \geq \|f - f_{(1)}\|_\mu^2 + \|f_{(1)} - f_{(2)}\|_\mu^2 + \cdots + \|f_{(m-1)} - f_{(m)}\|_\mu^2.$$

COROLLARY 3. $\|f_{(m-1)} - f_{(m)}\|_\mu \to 0$ *as* $m \to \infty$.

The proof follows from the fact that the sequence $f_{(m)}$, $m = 0, 1, \ldots$, does not depend on g, so that the number sequence

$$\sum_{k=1}^{m} \|f_{(k-1)} - f_{(k)}\|_\mu^2,$$

is bounded from above (by Theorem 4) and monotone increasing.

References

1. S. L. Marple, Jr., *Digital spectral analysis with applications*, Prentice Hall, Englwood Cliffs, NJ, 1987.
2. S. A. Aĭvazyan, V. M. Buchstaber, I. S. Yenyukov, and L. D. Meshalkin, *Applied statistics, classification, and reduction of dimensionality*, "Finansy i Statistika", Moscow, 1989. (Russian)
3. M. Ya. Antonovskiĭ, V. M. Buchstaber, and L. S. Veksler, *Application of multivariant statictical analysis for the detection of structural changes in the series of monitoring data*, Working paper WP-91-31, IIASA, Laxenburg, 1991.
4. ———, *Application of multivariant statictical analysis for the detection of structural changes in the time series of ecological monitoring data*, Problems of Ecological Monitoring and Modelling of Ecological Systems, vol. 15, Gidrometeoizdat, St-Petersburg, 1993. (Russian)
5. ———, *Detection of changes in the properties of monitoring time series by the methods of multivariant statictical analysis*, Problems of Measurements of Parameters of Hydroacoustical and Geophysical Fields and Information Processing, NPO "VNIIFTRI", Moscow, 1992. (Russian)
6. F. R. Gantmakher, *Theory of matrices*, 2nd ed., "Nauka", Moscow, 1966; English transl. of 1st ed., Chealsea, New York, 1959.
7. J. W. Milnor and J. D. Stasheff, *Characteristic classes*, Princeton Univ. Press, Princeton, NJ, 1974.