

Tensor decomposition and forecasting for multivariate time series

Kirill Semkin^{1*} and Vadim Strijov^{2*}

¹*Forecsys LLC Moscow Russia.

²*Forecsys LLC Moscow Russia.

*Corresponding author(s). E-mail(s): semkin.ki32@gmail.com;
strijov@forecsys.ru;

Keywords: time series; decomposition; forecasting; singular spectrum analysis; tensor decomposition; inertial measurement unit

Processing of multidimensional time series is associated with an additional task of determining dependencies between signals. Its inclusion in models boosts the quality of forecasts. On the other hand, taking this dependency into account makes models more complex and less interpretable. The paper proposes a non-parametric method based on tensor data representation and Singular Spectrum Analysis (SSA). It derives a time series decomposition technique and an explicit forecast model. Finally, it applies an elaborated theory to electricity consumption and inertial measurement unit datasets. It compares the obtained quality of forecast with mSSA, VAR, and RNN models.

1 Introduction

The main object of the paper is multivariate time series. It is a set of m time series $\{x_i(t)\}_{i=1}^m$ observed in discrete time $t \in 1, \dots, N$. We state two problems concerning them. First, to make a *forecast* is to estimate future values $\{x_i(T)\}_{i=1}^m$ of the time series at time $T > N$. Second, to build an additive *decomposition* is to represent each signal in a set as a sum of several components: $x_i(t) = c_1(t) + \dots + c_s(t)$, $\forall i \in 1, \dots, m$.

Analyzing multivariate time series allows to understand complex systems where multiple factors interact over time. It provides a more accurate understanding of the systems. For example, functional magnetic resonance imaging (fMRI) [1] aims to discover correlations with the specific cognitive states, such as memory and recognition, induced in the subject [2]. It is usually represented and processed as multivariate time series. When multiple variables are analyzed together, it leverages the relationships between them to make better forecast. For instance, in weather forecasting, considering temperature, humidity, and wind speed together yields more accurate predictions than analyzing them individually [3].

1.

For the decomposition in a single-variate time series case, the papers [4–6] introduce seasonal-trend-cycle techniques. For the forecast, the authors in [7] propose autoregressive methods as well as works [4, 8] suggest exponential smoothing, regression models, and neural networks.

However, these methods can not be transferred to the multivariate time series straightforwardly if they are interdependent. Indeed, each time series in the set is in general expressed as

$$x_i(t) = g_i(t, x_1(t), \dots, x_i(t), \dots, x_m(t)),$$

where $g_i(\cdot)$ is a $\mathbb{R}^{m+1} \rightarrow \mathbb{R}$ mapping. However, single-variate methods approximate it with $\hat{g}_i : \mathbb{R}^2 \rightarrow \mathbb{R}$ as

1.

$$x_i(t) = \hat{g}_i(t, x_i(t)),$$

which may not be sufficient. For example, the paper [9] models the size of prey-and-predator populations as a coupled system of differential equations. The dynamics of one population depends directly on the other's and vice-versa. Therefore, the forecast for populations can not be made separately.

The decomposition of multivariate time series reveals the structure of the interdependency. For example, multidimensional Fourier transform (MFT) [10] gives a time series' spectrum which describes decomposition on elementary harmonics. A few applications of MFT are image compressing [11] and partial differential equation solving [12]. The decomposition also allows denoising to be performed by separating noise from the actual time series. If the noise components of the multivariate time series are cross-correlated, it is clear that single-variate denoising methods will fail in this case. Several multivariate approaches include wavelet-based multivariate denoising [13] and multivariate empirical mode decomposition [14].

Some methods both take into account multivariate time series interdependence and make the forecast. First, *recurrent neural network* (RNN) [15] connects time series and their past values through several layers of nonlinear transformations. Information from the past values is encapsulated in the hidden state vector at each time step. It is used to forecast future values [16].

Second, the *vector autoregression* (VAR) [17, 18] is a linear stochastic model for multivariate time series. Denote vector $\mathbf{x}_t = (x_1(t) \dots x_m(t))^\top$ as a time series realizations at time t , their further dynamics

$$\mathbf{x}_t = \boldsymbol{\mu} + \sum_{i=1}^p A_i \mathbf{x}_{t-i} + \mathbf{u}_t.$$

Here μ is a constant vector, A_i are $m \times m$ matrices, \mathbf{u}_t is a random vector (e.g. white noise $\text{WN}(t)$). The time series interdependence is defined by A_i matrices, so each time series depends on each other and their past linearly.

The RNN and VAR methods are able to make the forecast, but, firstly, have a large number of learnable and structural parameters. That means their underlying models have an extensive complexity. Their architecture needs to be picked over and tuned. Secondly, the methods' structure does not contain an explicit way of building time series decomposition.

To rectify the listed disadvantages, we have developed a new method called the *tensor SSA* (tSSA). It has only two adjustable parameters. The tSSA extends the SSA method [19] for the multivariate time series and uses dynamical systems theory. The Takens theorem [20] is the theoretical foundation of the tSSA. *Hankel* matrices and their properties are involved in the decomposition problem. It is a matrix with equal elements on each anti-diagonal. Such matrices are tightly connected with Toeplitz and circulant matrices.

The key task of the tSSA is to find the shared *phase subspace* [21, 22] for all observed time series. This is a low-dimensional linear subspace where all system's trajectories can be mapped into *diffeomorphically*. Finding it enables us to obtain phase representation of the observed time series. For illustration, on the Fig. 1 (left) we use single time series $s(t)$ to build phase space and the specific trajectory \mathbf{x}_t of the dynamical system that generated the given time series, Fig. 1 (right). The forecast then is found upon the fact that the continuation of the trajectory lies in the very same phase space. Partitioning basis vectors of this space into several groups builds the decomposition of the time series. The paper develops named conceptions in the next sections.

Our method captures autocorrelation and multilinearity of the time series via tensor data representation. *Canonical polyadic decomposition* (CPD) is used to build the

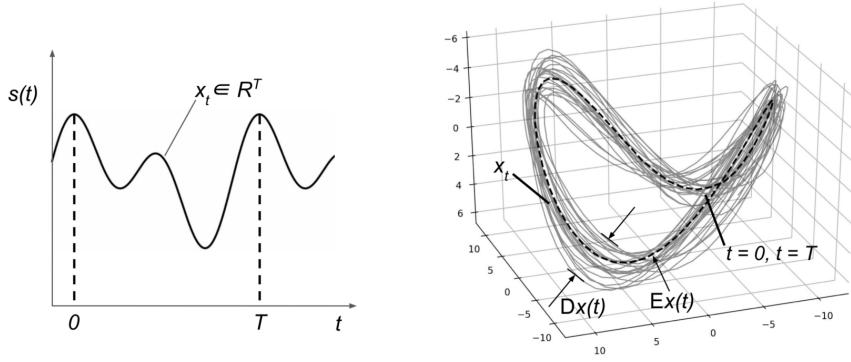


Figure 1 Time series's phase space visualization. Single time series $s(t)$ builds phase trajectory \mathbf{x}_t in a 3-dimensional phase space

phase space of the time series. Another solution is to use matrix data representation and matrix decompositions. Such method is called matrix SSA (mSSA) [23]. It is also an extension of the SSA method for the forecast and the decomposition of the multivariate time series. The tSSA is compared with the mSSA in the computational experiment section.

The main contributions of the paper are a mathematically rigid problem statement and corresponding derivation of the decomposition and forecast techniques, analysis of the optimal decomposition problem, autoregressive interpretation of the resulting forecasting model, tSSA comparison with other methods on multivariate time series datasets from diverse domains.

The rest of the paper covers the theory and application of our method to real data. Firstly, the dynamical systems model is introduced and the problem of basis search in the phase space is stated and resolved. These results enable us to propose a way of decomposing the time series and to make the forecast. Simultaneously, features of obtained solutions are examined and two associated theorems are formulated. Finally, tSSA and mentioned methods are applied to three datasets: electricity consumption, accelerometer/gyroscope observations, weather parameters. We obtain their forecast and additive decompositions. For the latter, a special quality metric is introduced.

2 Problem statement

Suppose a *dynamical system* defined on X , with the evolution function f and the initial state \mathbf{y}_0 . Time space is discrete

$$\mathbf{y}(t+1) = f(\mathbf{y}(t)), \quad t \in \mathbb{N},$$

$$\mathbf{y}(0) = \mathbf{y}_0.$$

Generally, X is a high-dimensional smooth manifold. Then unknown mapping $\phi : X \rightarrow \mathbb{R}^m$ takes each trajectory point $\mathbf{y}(t)$ to the multivariate time series realization \mathbf{x}_t

$$\phi(\mathbf{y}(t)) = \mathbf{x}_t. \text{ Coordinately: } \begin{cases} \phi_1(\mathbf{y}(t)) = x_1(t), \\ \dots \\ \phi_m(\mathbf{y}(t)) = x_m(t). \end{cases}$$

The Fig. 2 visualizes the given relations. On the left, it depicts the Lorenz attractor and its dynamical system in the form of differential equations. The attractor is a special set of the system's trajectories. Each point on the attractor is projected on some axis which gives us observed time series on the right. So we have the trivial case here when $\phi(\mathbf{y}(t)) \equiv \mathbf{y}(t)$.

4.

Assume that the trajectories $\mathbf{y}(t)$ belong to the smaller dimension manifold $M \subset X$. Now the problem is to find an embedding of the M into \mathbb{R}^L for some L . The second problem is to build a basis in the image of the embedding. Having done that, the initial dynamical system will be described in terms of the standard linear space. The same will hold to all time series $x_i(t)$.

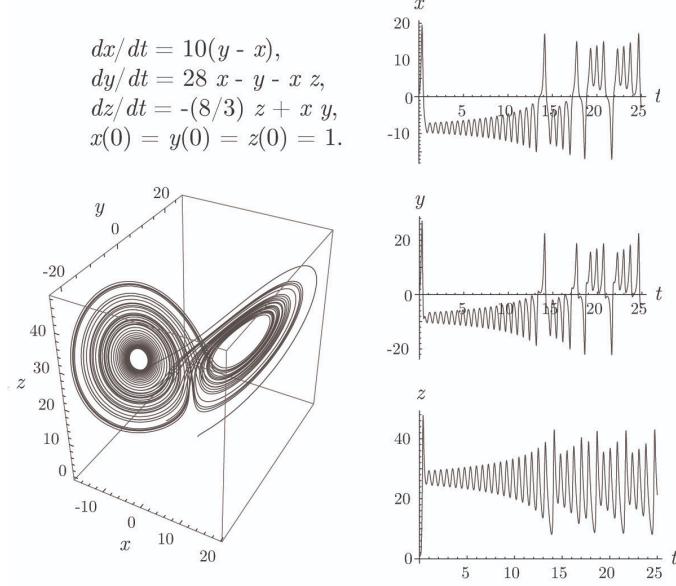


Figure 2 The Lorenz system and the Lorenz attractor (left). Attractor's projections on each axis (right).

3 Single-variate time series case

The tSSA relies on the SSA method and the Takens theorem. The theorem provides the needed embedding in the single-variate time series case. Any point $\mathbf{y}(t) \in M$ corresponds to the following vector:

$$[\phi \circ f^{t-L+1}(\mathbf{y}(t)), \dots, \phi \circ f(\mathbf{y}(t)), \phi \circ \mathbf{y}(t)]^\top = [x(t-L+1), \dots, x(t-1), x(t)]^\top.$$

It is called the *delay vector* at time t and is denoted as $\overleftarrow{\mathbf{x}}_t$. The vector's dimension L must satisfy condition $L > 2 \cdot \dim(M)$. The function $\phi(\cdot)$ must satisfy several regularity conditions. They are omitted here and considered fulfilled.

The Fig. 2 shows an application of the Takens theorem to the Lorenz attractor.

The attractor (on the left) is mapped to the observed time series $x(t)$ (at the bottom).

The 3-dimensional delay vectors based on the $x(t)$ are then constructed (on the right).

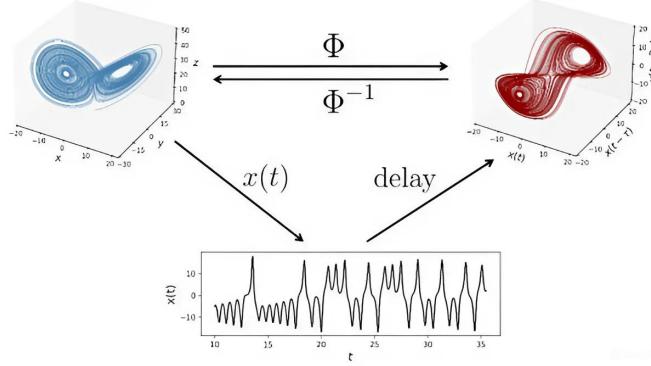


Figure 3 Visualization of the Takens theorem applied to the Lorenz attractor.

By the Takens theorem the resulting attractor is diffeomorphic to the initial one. The diffeomorphism is denoted here as Φ . 4.

So the time series $x(t)$ with the length N gives $N - L + 1$ delay vectors. Therefore image of the embedding is the linear span of these vectors $\text{Lin}(\{\overleftarrow{x}_t\})$. It represents the individual phase space of the time series and must be low dimensional. This means $\text{Lin}(\{\overleftarrow{x}_t\}) \subset \mathbb{R}^L$. Finally, the chosen orthonormal basis in the phase space is the U -component of the Singular value decomposition (SVD) on the *trajectory matrix* H_x . It is a stacking of the delay vectors

$$H_x = [\overleftarrow{x}_1 \dots \overleftarrow{x}_{N-L+1}].$$

4 Multivariate time series case and tSSA method

Now there given multivariate time series. Let m be their number. Hence, the $\phi(\cdot)$ is a multidimensional map. This means that instead of the single delay vector we have m of them at any time t . Let \overleftarrow{X}_t be the matrix $[\overleftarrow{x}_{1t} \dots \overleftarrow{x}_{mt}]$ called a *delay matrix*. Then, the embedding image is a linear span of all \overleftarrow{X}_t . It is also called a *shared* phase space of the multivariate time series.

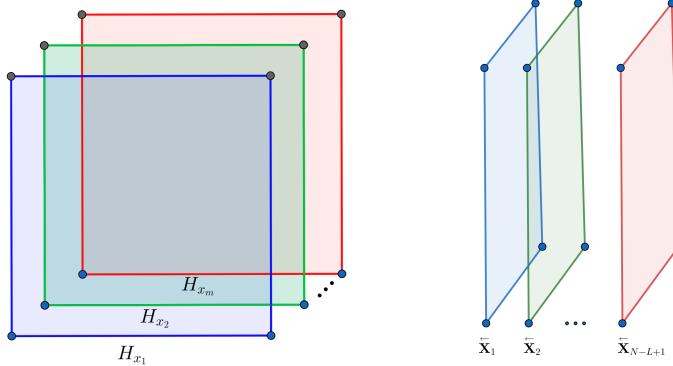


Figure 4 Two views on trajectory tensor. The left is in terms of signals' trajectory matrices $\{x_i(t)\}_{i=1}^m$. The right is in terms of delay matrices

Similarly to the trajectory matrix introduce a *trajectory tensor* denoted by \mathbf{T} . It stacks delay matrices along the second index of the tensor, Fig. 4 (left). From the definition, it follows that each tensor slice along the second index is a trajectory matrix. Then, the \mathbf{T} is also the trajectory matrices H_{x_i} stacked along the third index, Fig. 4 (right). Therefore, the shared phase space of the time series is equal to a linear span of all H_{x_i} . As discussed in the Sec 3, each trajectory matrix also defines the individual phase space for the corresponding time series $x_i(t)$. Finally, we give

Definition 1. The multivariate time series are said to be *interdependent* if they share the common phase space with the same basis.

Assuming the observed time series interdependence, it follows that each H_{x_i} has a matrix factorization with the same set of factors. Since the trajectory matrices constructs the trajectory tensor, it has a low-dimensional representation. Using these two corollaries, we apply the CP-decomposition to the \mathbf{T} and obtain

$$\mathbf{T} = \sum_{i=1}^r \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i. \quad (1)$$

The CPD is defined by the tensor rank r and a set of vectors. The vectors are usually composed in full-rank matrices

$$A = [\mathbf{a}_1 \dots \mathbf{a}_r], B = [\mathbf{b}_1 \dots \mathbf{b}_r], C = [\mathbf{c}_1 \dots \mathbf{c}_r]. \quad (2)$$

Denote by σ_{x_k} the k -th row of the C . Finally, with respect to the third-dimension slices of the \mathbf{T} , CPD goes as

$$\left\{ \begin{array}{l} H_{x_1} = \sum_{i=1}^r \sigma_{x_1}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top, \\ H_{x_2} = \sum_{i=1}^r \sigma_{x_2}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top, \\ \dots \\ H_{x_m} = \sum_{i=1}^r \sigma_{x_m}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top. \end{array} \right. \quad (3)$$

The CPD exists for any tensor. In case of the trajectory tensor of the interdependent time series three properties must hold.

Theorem 1. *Suppose that the given multivariate time series are interdependent (Def. 1). Then*

1. Any σ_{x_k} vector does not have zero elements.
2. The tensor rank $r \leq L$. It is equal to the dimension of the shared phase space.
3. Transposition of the trajectory matrix is a trajectory matrix. The shared phase space for the transposed matrices is the $\text{Lin}(\{\mathbf{b}_i\})$.

Proof. 1. Suppose some vector has a zero element at the position j and the other is non-zero in this position. Then two trajectory matrices factorize with a different set of factors. That violates the definition of the time series interdependence. The other case is that all σ_{x_k} have a zero element at the position j . Then the matrix C has a zero row and is not full-ranked. It contradicts to the CPD definition.

2. Since delay vectors are L -dimensional, each time series has its phase space dimension $\leq L$. All delay vectors of every time series also lie in the shared phase space $\text{Lin}(\{\mathbf{a}_i\})$. Therefore, dimensionality of the shared space can not be greater than L . Then, the set of $\{\mathbf{a}_i\}_{i=1}^r$ vectors is linear-independent as follows from the CPD definition. So the dimension of the shared phase space equals r .
3. Rows of the trajectory matrices are the delay vectors of size $N - L + 1$. Transpose all equalities in (3). Now it is clear that rows of any H_{x_i} belongs to the $\text{Lin}(\{\mathbf{b}_i\})$. This is the shared phase space for the new delay vectors.

□

5 Time series decomposition

The time series are decomposed into additive components using the following conception: *factorization of the trajectory matrix H_{x_k} defines the decomposition of the related time series*. This idea comes from the SSA method. Since all H_{x_k} have a similar factorization, each time series is decomposed separately.

Trajectory matrices of the time series are *hankel* by their definition. That means equal elements along each anti-diagonal of the matrix. It is trivial that any time series of length N bijectionally corresponds to a hankel matrix of size $L \times (N - L + 1)$.

Describe the decomposition algorithm. First, the factors of the H_{x_k} are arbitrary partitioned into s groups. The s is also a chosen number. Then all factors within each partition are summed up. As a result we obtain matrices C_1, \dots, C_s . If they are all hankel then each C matrix corresponds to the time series component of the $x_i(t)$. However, it is almost infeasible even with the trivial time series (see the [19]). Therefore every C matrix is additionally *hankelized*. This operator averages each matrix's anti-diagonals so the matrix becomes hankel. Denote it as $\text{Hankel}(\cdot)$. Finally, the algorithm

can be written as a chain of identical transformations

$$H_{x_k} = \quad (4)$$

$$\sum_{i=1}^r \boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top = \quad (5)$$

$$\sum_{i \in \mathbb{I}_1} \boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top + \dots + \sum_{i \in \mathbb{I}_s} \boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top = \quad (6)$$

$$C_1 + \dots + C_s = \quad (7)$$

$$\text{Hankel}(C_1) + \dots + \text{Hankel}(C_s).$$

Final expression corresponds to the $x_k(t) = c_1(t) + \dots + c_s(t)$.

Here $\mathbb{I}_1 \sqcup \dots \sqcup \mathbb{I}_s = \{1, \dots, r\}$ are chosen groups of indices. $c_1(t), \dots, c_s(t)$ are the time series components extracted from the $\text{Hankel}(C_1), \dots, \text{Hankel}(C_s)$ (see the second paragraph of this section). The (7) step requires a proof. But first, consider a hankel operator's property:

Lemma 1. *The $\text{Hankel}(\cdot)$ is a linear operator.*

Proof. Having matrices A and B with the same size, take their elements from any anti-diagonal. Denote them by a_1, \dots, a_n and b_1, \dots, b_n . Then this anti-diagonal in $\text{Hankel}(A + B)$ is

$$\frac{1}{n} \sum_{i=1}^n (a_i + b_i) = \frac{1}{n} \sum_{i=1}^n a_i + \frac{1}{n} \sum_{i=1}^n b_i.$$

It exactly results in a sum of the $\text{Hankel}(A)$'s and $\text{Hankel}(B)$'s anti-diagonals.

Secondly, having the scalar α , it is trivial that

$$\frac{1}{n} \sum_{i=1}^n \alpha \cdot a_i = \alpha \frac{1}{n} \sum_{i=1}^n a_i.$$

Hence, we derive $\text{Hankel}(\alpha A) = \alpha \text{Hankel}(A)$. □

Now recall that the H_{x_k} is hankel. Combining it with the previous lemma, apply the Hankel(\cdot) to the $H_{x_k} = C_1 + \dots + C_s$ from the algorithm. This proofs (7). Therefore we have shown the correctness of the decomposition algorithm.

Note that the algorithm does not guarantee any properties of the built components in general. For example, take the SSA method. Its decomposition algorithm is also a partitioning of the factor matrices. Let the input time series be a sum of $\cos(t)$ and $\exp(t)$ time series. Then it is known that the algorithm cannot separate these components. However, it is possible to obtain similar in some sense components, see [19]. So it is not always possible to perfectly separate trend and seasonability with the tSSA but it is possible to obtain close approximation of these components. Another drawback of the algorithm is ambiguity in the choice of partitioning. One possible solution is proposed and analyzed in the next section. In practice, one may take a random partition or define custom score function to be optimized w.r.t. choice of partition.

5.

6 Optimal decomposition problem

To eliminate ambiguity in the partitioning of the H_{x_k} factors, we restrict C_j matrices to be hankel. Thus, the (7) step of the decomposition algorithm becomes unnecessary. In addition, such partitioning completely fulfills the decomposition concept.

Formulate the decomposition problem as a system of matrix equations. Take (5) equality and apply the hankel operator to both sides. Recall that the H_{x_k} is hankel. Also take the very same equality as it is. We obtain

$$\begin{cases} H_{x_k} = \sum_{i=1}^r \boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top, \\ H_{x_k} = \sum_{i=1}^r \text{Hankel}(\boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^\top). \end{cases}$$

Then, subtract the second from the first and introduce

$$H_i = \boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^T - \text{Hankel}(\boldsymbol{\sigma}_{x_k}(i) \cdot \mathbf{a}_i \mathbf{b}_i^T) \quad (8)$$

It is called the *hankel residual matrix* of the i -th factor. Therefore we get

$$H_1 + \dots + H_r = 0. \quad \text{Or equally } H_r = - \sum_{j=1}^{r-1} H_j. \quad (9)$$

The condition on the C_i to be hankel is reformulated as the residual matrices sum up to zero within each partition group. The problem is to find the non-intersecting sets $\mathbb{I}_1, \dots, \mathbb{I}_s$ such that

$$\sum_{k \in \mathbb{I}_i} H_k = 0, \quad i \in 1, \dots, s, \quad (10)$$

$$\mathbb{I}_1 \sqcup \dots \sqcup \mathbb{I}_s = \{1, \dots, r\}.$$

Without loss of generality, we set s equals two. Only $\mathbb{I}_1, \mathbb{I}_2$ must be found. If (10) is solved in such setting, then \mathbb{I}_1 and \mathbb{I}_2 can be further partitioned separately. Hence, any number of the time series components can be obtained. Moreover, for any H_{x_k} its residual matrices are linearly dependent as follows from (9). Thus, the partition group for the H_r is uniquely defined if all other residuals are already partitioned. This matrix can be discarded. We also assume all H_j to be non-zero matrices. Otherwise, the zero residual matrix alone and all the rest make the desired partition on two groups. This assumption implies each of the groups to have at least two residual matrices. Finally, for any H_j matrix introduce an indicator-variable $\beta_j \in \{0, 1\}$. It shows what group the residual matrix belongs to. Denote by $\boldsymbol{\beta}$ a vector $[\beta_1 \dots \beta_{r-1}]^T$. The decomposition

problem is to find β such that

$$\begin{cases} \sum_{j=1}^{r-1} \beta_j H_j = 0, \\ \beta_j \in \{0, 1\} \forall j \in 1 \dots r, \\ \sum_{i=1}^{r-1} \beta_i \geq 2. \end{cases} \quad (11)$$

The [19] shows that even for trivial time series such problem is infeasible. Now we transform it into a feasible one. First, vectorize each H_i and stack resulting vectors into the matrix denoted by Λ . Then the first equation in (11) is equivalent to the $\Lambda\beta = 0$. Second, change the problem of solving the system of equations to an optimization problem. Let's find such β that delivers a minimum to the norm of the $\Lambda\beta$. The changed problem is

$$\begin{cases} \|\Lambda\beta\| \rightarrow \min_{\beta}, \\ \beta_j \in \{0, 1\} \forall j \in 1 \dots r, \\ \sum_{i=1}^{r-1} \beta_i \geq 2. \end{cases} \quad (12)$$

Call solution of (12) an *optimal time series decomposition*. The problem itself is an Integer Least Squares (ILS) problem, which is proved to be NP-hard [24]. Therefore, *the optimal time series decomposition is an NP-hard problem*.

In spite of the complexity, methods and heuristics exist to reduce the ILS to computationally effective problems, see [25].

7 Time series forecasting

Using the build shared phase space each time series is forecast individually. We make a one-step forecast to estimate $x(N+1)$. Phase trajectory of the time series is a sequence of the delay vectors $\{\overleftarrow{\mathbf{x}}_t\}$. The whole trajectory belongs to the phase space $\text{Lin}(\{\mathbf{a}_i\})$. Continuation of the trajectory is the $\overleftarrow{\mathbf{x}}_{N+1}$. Therefore $\overleftarrow{\mathbf{x}}_{N+1} \in \text{Lin}(\{\mathbf{a}_i\})$. At the same

time $x(N + 1)$ is the last component of the $\overleftarrow{\mathbf{x}}_{N+1}$. Now, recall the A matrix from (2).

The system of linear equation to find $x(N + 1)$ is

$$\overleftarrow{\mathbf{x}}_{N+1} = A\boldsymbol{\lambda} \text{ or equally } \begin{cases} \mathbf{x} = \tilde{A}\boldsymbol{\lambda} \\ x(N + 1) = \mathbf{a}^\top \boldsymbol{\lambda} \end{cases}, \text{ where} \quad (13)$$

$$A = \begin{pmatrix} \tilde{A} \\ \mathbf{a}^\top \end{pmatrix}, \quad (14)$$

$$\overleftarrow{\mathbf{x}}_{N+1} = (\mathbf{x} \ x(N + 1))^\top. \quad (15)$$

The delay vector $\overleftarrow{\mathbf{x}}_{N+1}$ and the matrix A are split into two blocks, see (14) and (15). The blocks are associated with the known and predicted part of the time series. The forecast comes from the second equation in (13). Therefore, the vector $\boldsymbol{\lambda} \in \mathbb{R}^r$ is to be found from the first equation in (13). It is a system of linear equations with the matrix \tilde{A} . The matrix has a rank at least $r - 1$ since A matrix is full-ranked. However, we assume \tilde{A} to have rank r , too. As experiments show this is not burdensome. In addition, the linear system is overdetermined since $r \leq L$, see Sec 4. Therefore its solution is given by the least-squares formula $\boldsymbol{\lambda} = (\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top \mathbf{x}$. After replacing it in (13) we obtain the forecast

$$x(N + 1) = \mathbf{a}^\top (\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top \mathbf{x}. \quad (16)$$

Denote by \mathbf{d} the vector $\mathbf{a}^\top (\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top$. Once computed, it is used to make the forecast for further steps. Rewrite (16) in terms of $x(t)$:

$$x(t) = \sum_{i=1}^{L-1} d_i \cdot x(t - i)$$

This expression reveals the property of the forecast sequence $\{x(N + i)\}$.

Theorem 2. *The model of the tSSA's forecast is autoregressive $AR(L - 1)$. The coefficients d_i found in (16), completely define behavior of the forecast sequence.*

8 Computational experiment

Analyze forecast and decomposition quality of tSSA on sample multivariate time series. Also compare tSSA with the other methods: mSSA, VAR, and RNN. The RNN and VAR methods are able to make the forecast, but do not contain an explicit way of building time series decomposition.

To measure quality of the decomposition we introduce special metrics. They are motivated by the hankel residual matrices (8) introduced for the optimal decomposition problem, see Sec 5.

Definition 2. The *absolute hankel error* of the matrix M is

$$AHE = \|M - \text{Hankel}(M)\|_F.$$

Definition 3. The *relative hankel error* of the matrix M is

$$RHE = \frac{AHE}{\|M\|_F}.$$

The AHE is proportional to the total variance of matrix's anti-diagonals. The RHE is AHE's more interpretable modification. Recall the decomposition algorithm (5). It introduced the C_i matrices corresponded to the $c_i(t)$ components of the time series $x(t)$. The (12) is the optimal decomposition problem. It is basically AHE minimizing for the C_i matrices. Hence, the AHE and the RHE are computed for the C_i in the further experiments. Denote by \overline{RHE}_{ts} the mean RHE across all C_i of the time series named "ts". Denote by \overline{RHE} the mean \overline{RHE}_{ts} across time series of this computational experiment.

To measure forecast quality for the time series named "ts", use: the *mean squared error* MSE_{ts} and the *mean absolute percentage error* $MAPE_{ts}$. Denote by \overline{MSE}

and $\overline{\text{MAPE}}$ the mean MSE_{ts} and MAPE_{ts} across time series of this computational experiment.

Three sample multivariate time series are involved in the experiment. First, the electricity consumption and price, Fig. 5. Second, the inertial unit measurements [26]: three time series are from an accelerometer and three from a gyroscope, Fig. 6. The measurements represent walking movements. Third, measurements of the weather parameters, Fig. 7. It includes temperature, precipitation and air pressure. We assume the multivariate time series have shared phase space, see Def. 1. For the electricity time series it is related to the law of demand and supply. For the inertial unit time series it is related to coupled dynamical system describing the motion. Similar reasoning applies to the weather dynamics. The sample time series have the length $3 \cdot 10^3$ time ticks. About 20% of them are deferred test samples for the forecast quality evaluation.

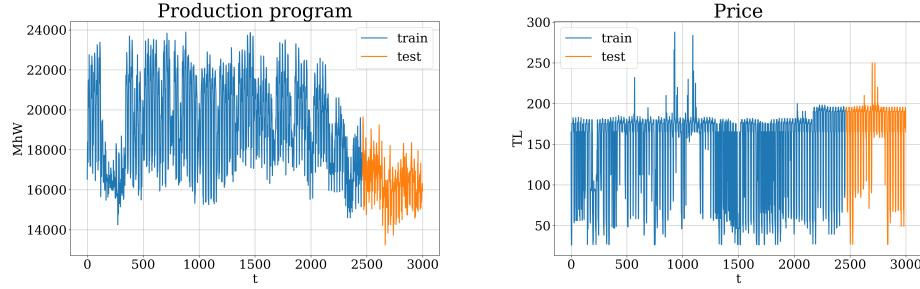


Figure 5 Time series for the electricity consumption and price

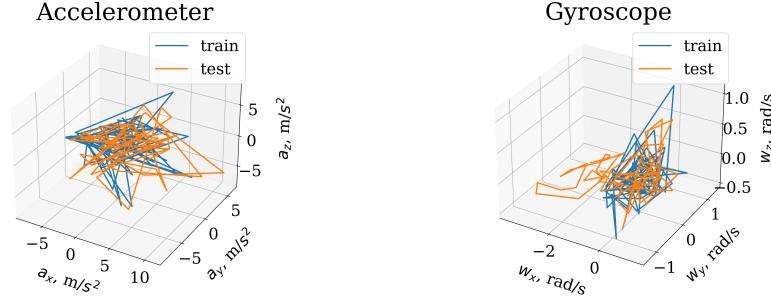


Figure 6 Time series for the inertial unit measurements

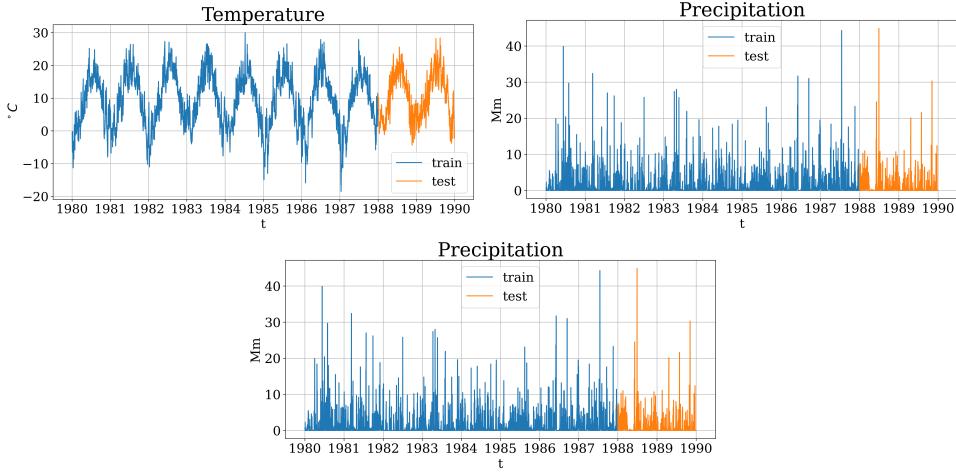


Figure 7 Time series for the weather parameters measurements.

The tSSA's parameter L is set to 500 for the electricity time series and to 1000 for the inertial unit time series. This is also a parameter for the other models. For RNN it is the minimum length of time series. For VAR it is a maximal lag number. For mSSA it is similar to tSSA.

9 Data availability

The electricity consumption data is available at <https://sourceforge.net/p/mvr/code/HEAD/tree/data/TurkElectricityConsumption.csv>. The inertial unit data is available at <https://data.mendeley.com/datasets/45f952y38r/5>. The weather data is available at <https://www.kaggle.com/datasets/guillemservera/global-daily-climate-data/data>.

10 Results and discussion

First, analyze the forecast quality. The trajectory tensor rank is equal to the forecast model's order, see Th. 2. Since efficient computation of the tensor rank is impossible, *the rank becomes a tSSA's parameter*. It was found that the forecast quality decreases with the greater ranks for the MAPE and MSE. See Fig. A1 and A2 in the Appendix

section. The optimal rank for each dataset was chosen based on the MSE. The CPD computation error decreases monotonously with greater ranks, see Fig. A4 in the Appendix section. This error is an estimation of how close the computed CPD is to the true one (1). As a result, the small-order forecast model delivers the best quality.

The Fig. 8, 9, 10 visualize forecast for the datasets. We can see that with the low rank for the weather time series, the forecast is a mean value. With the greater rank for the electricity and inertial unit time series, the forecast is a more complex curve. However, further increasing the rank always led to exploding forecasts for all datasets.

Tab. 1, 2, 3 contain the MAPE and MSE metrics for all models. Our method showed the best results in most cases. Close values were obtained with mSSA. The VAR model appeared unstable for the test samples. The RNN learned only the constant function from the electricity samples. But with the greater L for the inertial unit and weather samples, the RNN performed better.

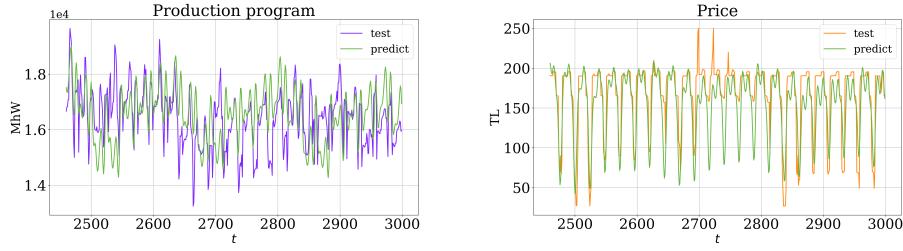


Figure 8 tSSA forecast for the electricity data. CPD rank = 30

Table 1 Forecast quality of the models on the electricity data

Method Metric	<i>tSSA</i>	<i>mSSA</i>	<i>VAR</i>	<i>RNN</i>
MSE _{Production, 10⁶}	1.24	1.51	7.81	2.70
MSE _{Price, 10³}	0.88	1.03	4.85	30.0
MSE, 10 ⁶	0.62	0.75	3.91	135.00
MAPE _{Production}	0.054	0.060	0.137	0.999
MAPE _{Price}	0.164	0.170	0.360	1.004
MAPE	0.109	0.115	0.249	1.002

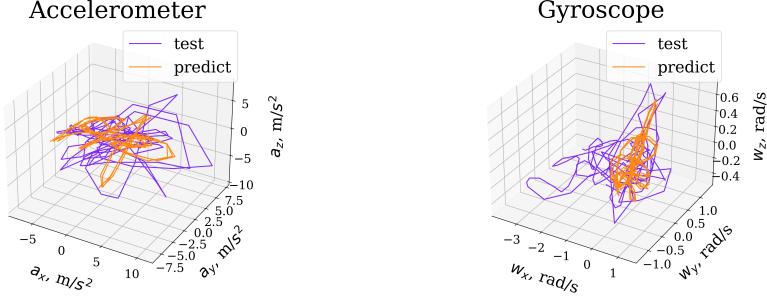


Figure 9 tSSA forecast for the accelerometer and gyroscope data. CPD rank = 20

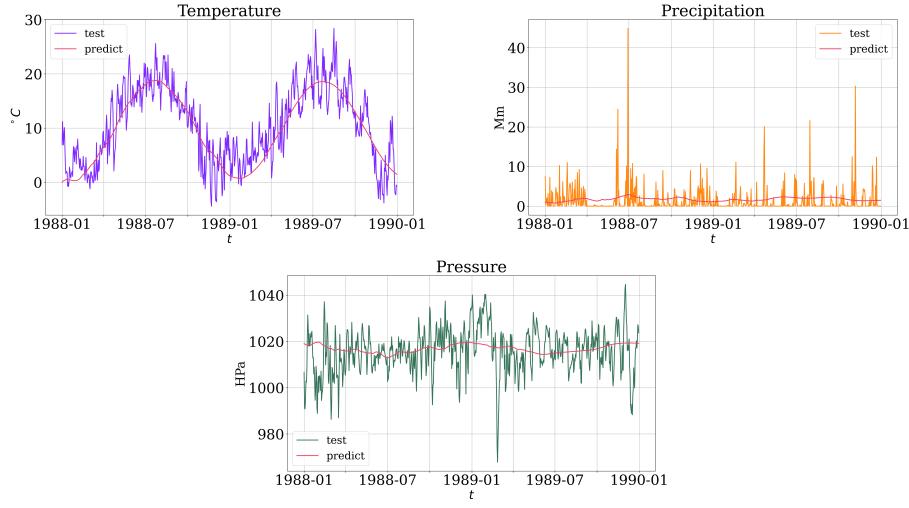


Figure 10 tSSA forecast for the weather data. CPD rank = 5

Now, analyze the decomposition quality. It was found that the RHE reaches a minimum for the tensor rank = 20 on all datasets. As the rank goes greater, the RHE does not change or even increase. See Fig. A5 in the Appendix section. The forecast quality had the same pattern. Therefore, small tensor ranks deliver the best quality for the decomposition and forecast problems. At the same time, the rank is a value of the shared phase space dimension, see Th. 1. The delay vectors define phase trajectory and have dimension equals L . Recall that L is set to 500 for the electricity samples and to 1000 for the inertial unit samples. As a result, tSSA has significantly reduced dimensionality of the sample time series.

Table 2 Forecast quality of the models on the inertial unit data

Metric \ Method	<i>tSSA</i>	<i>mSSA</i>	<i>VAR</i>	<i>RNN</i>
$\overline{\text{MSE}}_{\text{Accel}}$	7.351	6.980	8.108	6.604
$\overline{\text{MSE}}_{\text{Gyro}}$	0.610	0.636	0.631	0.639
$\overline{\text{MSE}}$	3.981	3.808	4.370	3.622
$\overline{\text{MAPE}}_{\text{Accel}}$	3.558	3.516	3.370	1.747
$\overline{\text{MAPE}}_{\text{Gyro}}$	3.773	3.943	10.427	5.641
$\overline{\text{MAPE}}$	3.666	3.730	6.899	3.694

Table 3 Forecast quality of the models on the weather data

Metric \ Method	<i>tSSA</i>	<i>mSSA</i>	<i>VAR</i>	<i>RNN</i>
$\overline{\text{MSE}}_{\text{Temperature}}$	14.17	20.12	40.68	3.21
$\overline{\text{MSE}}_{\text{Precipitation}}$	11.02	11.23	22.84	1.77
$\overline{\text{MSE}}_{\text{Pressure}}$	98.02	101.23	156.75	2.28
$\overline{\text{MSE}}$	41.07	44.19	73.42	2.42
$\overline{\text{MAPE}}_{\text{Temperature}}$	0.850	0.626	0.468	29.658
$\overline{\text{MAPE}}_{\text{Precipitation}}$	3.151	2.834	4.638	2.965
$\overline{\text{MAPE}}_{\text{Pressure}}$	0.007	1.254	3.983	7.073
$\overline{\text{MAPE}}$	1.336	1.571	3.030	13.232

Fig. 11 illustrates the decomposition on two components for the electricity consumption and price. The components are almost identical to the bias and scale factors. This is attributed to the *shared* phase space between the time series. Fig. 12 shows two-component decomposition for the accelerometer and Fig 13 shows the same for the gyroscope. The second components are identical to the rotation and scale. Moreover, all components here coincide with the phase trajectories. The Fig. 14 shows two-component decomposition for the weather parameters. Temperature and precipitation components are similar to each other and represent high- and low-amplitude periodical time series. Pressure components are periodical with equal amplitudes. They are both similar to the second components for other parameters.

6.

Tab. 4, 5, 6 shows tSSA have lower decomposition quality than mSSA. It is attributed to the two factors. First, the mSSA decomposition is connected with the factors partitioning like tSSA. Several heuristics exist to obtain usually high-RHE

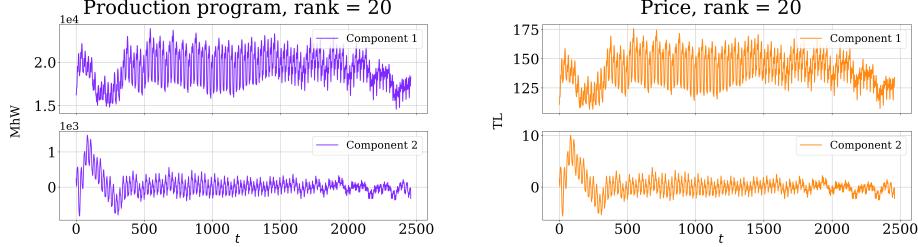


Figure 11 tSSA decomposition for the electricity data. CPD rank = 20

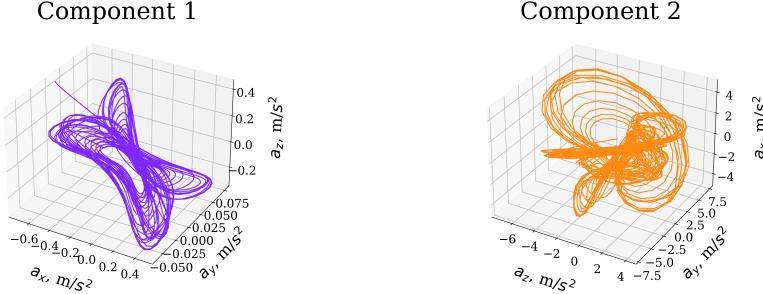


Figure 12 tSSA decomposition for the accelerometer data. CPD rank = 10

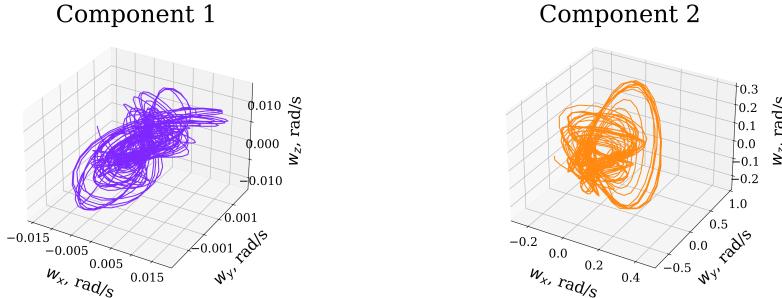


Figure 13 tSSA decomposition for the gyroscope data. CPD rank = 10

components. On the other hand, the tSSA optimal decomposition problem (12) is the Integer Least Squares (ILS) problem, see the end of Sec. 6. It can not be solved efficiently, only approximate methods are available. Moreover, the matrix Λ involved in the ILS has a large row dimension. To shorten computation time, only several hundred rows were left. It is still much greater than the size of the sought parameter vector in the ILS. But such approximation is not equal to the initial problem.

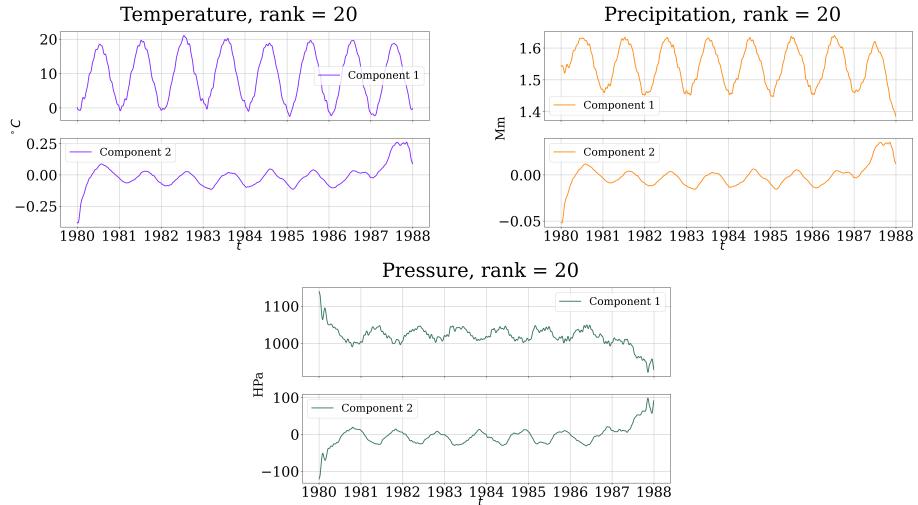


Figure 14 tSSA decomposition for the weather data. CPD rank = 20

Table 4 Decomposition quality of the models for the electricity data

Metric \ Method	tSSA	mSSA
$\overline{RHE}_{Production}$	0.507	0.308
\overline{RHE}_{Price}	0.511	0.310
\overline{RHE}	0.509	0.309

Table 5 Decomposition quality of the models for the inertial unit data

Metric \ Method	tSSA	mSSA
\overline{RHE}_{Accel}	0.438	0.394
RHE_{Gyro}	0.732	0.468
\overline{RHE}	0.585	0.431

Table 6 Decomposition quality of the models for the weather data

Metric \ Method	tSSA	mSSA
$\overline{RHE}_{Temperature}$	0.519	0.467
$\overline{RHE}_{Precipitation}$	0.508	0.538
$RHE_{Pressure}$	0.542	0.502
\overline{RHE}	0.523	0.502

11 Conclusion

The tSSA method is devoted to the forecast and the additive decomposition of the interdependent multivariate time series. It is based on dynamical systems theory and explicitly build the shared phase space of the time series. It also exploits multilinearity of the data via tensor data representation. The method has only two adjustable parameters and require the CPD computation of the trajectory tensor. The tSSA forecast is vector's scalar product. The tSSA decomposition is matrix factorization and factor's partitioning. Finally, the computational experiment showed tSSA's high forecast quality. For the electricity data, the tSSA has the best forecast. The MSE metric is 21% less than the second best forecast. The MAPE metric is 35% less than the other methods on average. The tSSA also reduced dimensionality of the electricity data in 16 times.

Future research will be devoted to a new way of building the shared phase space. Then, the current decomposition conception results in a computationally complex problem. Finding new concepts to build additive components is the other direction for the method improvement.

References

- [1] Rinck, P.A.: Magnetic Resonance in Medicine: A Critical Introduction. round table foundation, ??? (2018)
- [2] Logothetis, N.K.: What we can do and what we cannot do with fmri. *Nature* **453**, 869–878 (2008) <https://doi.org/10.1038/nature06976>
- [3] Wang, M., Wang, J., Yu, M., Yang, F.: Spatiotemporal wind speed forecasting using conditional local convolution and multidimensional meteorology features. *Scientific Reports* **14** (2024) <https://doi.org/10.1038/s41598-024-78303-8>

- [4] Enders, W.: Applied Econometric Time Series. Wiley series in probability and statistics. Wiley, New York (2010)
- [5] Dagum, E., Bianconcini, S.: Seasonal Adjustment Methods and Real Time Trend-Cycle Estimation, p. 283. Springer, Switzerland (2016)
- [6] Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.: Stl: A seasonal-trend decomposition procedure based on loess (with discussion). *Journal of Official Statistics* **6**, 3–73 (1990)
- [7] Box, G.E.P., Jenkins, G.M.: Time Series Analysis: Forecasting and Control. Holden-Day series in time series analysis and digital processing. Holden-Day, Minnesota (1970)
- [8] Hyndman, R.J., Athanasopoulos, G.: Forecasting: Principles and Practice. OTexts, Australia (2021)
- [9] Volterra, V., Brelot, M.: Leçons sur la Théorie Mathématique de la Lutte pour la vie. Cahiers Scientifiques. Fasc Series. Gauthier-Villars et cie., Michigan (1931)
- [10] Smith, W.W., Smith, J.M.: Handbook of Real-Time Fast Fourier Transforms: Algorithms to Product Testing. Wiley, ??? (1995)
- [11] Raid, A., Khedr, W., El-dosuky, M., Ahmed, W.: Jpeg image compression using discrete cosine transform - a survey. "International Journal of Computer Science & Engineering Survey" **5** (2014) <https://doi.org/10.5121/ijcses.2014.5204>
- [12] Haberman, R.: Applied Partial Differential Equations: With Fourier Series and Boundary Value Problems. Featured Titles for Partial Differential Equations. Pearson, ??? (2013)

- [13] Naveed, K., Rehman, N.: Wavelet based multivariate signal denoising using mahalanobis distance and edf statistics (2020) <https://doi.org/10.48550/arXiv.2005.11616>
- [14] Hao, H., Wang, H.L., Rehman, N.U.: A joint framework for multivariate signal denoising using multivariate empirical mode decomposition. *Signal Processing* **135**, 263–273 (2017) <https://doi.org/10.1016/j.sigpro.2017.01.022>
- [15] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computation* **9**(8), 1735–1780 (1997) <https://doi.org/10.1162/neco.1997.9.8.1735>
- [16] Tealab, A.: Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal* **3**(2), 334–340 (2018) <https://doi.org/10.1016/j.fcij.2018.10.003>
- [17] Asteriou, D., Hall, S.: Vector Autoregressive (VAR) Models and Causality Tests, pp. 333–346 (2016). https://doi.org/10.1057/978-1-349-41547-9_15
- [18] Zellner, A.: An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**(298), 348–368 (1962) <https://doi.org/10.1080/01621459.1962.10480664>
- [19] Stepanov, D., Golyandina, N.: SSA-based approaches to analysis and forecast of multidimensional time series, pp. 293–298 (2005)
- [20] Takens, F.: Detecting Strange Attractors in Turbulence. In: *Dynamical Systems and Turbulence*, Warwick 1980. Lecture Notes in Mathematics, vol. 898, pp. 366–381. Springer, Berlin (1981). Chap. 21. <https://doi.org/10.1007/bfb0091924>
- [21] Krim, H., Viberg, M.: Two decades of array signal processing research: The parametric approach. *Signal Processing Magazine, IEEE* **13**, 67–94 (1996) <https://doi.org/10.1109/78.508422>

[//doi.org/10.1109/79.526899](https://doi.org/10.1109/79.526899)

- [22] Ignatov, A.D., Strijov, V.V.: Human activity recognition using quasiperiodic time series collected from a single tri-axial accelerometer. *Multimedia tools and applications* **75**, 7257–7270 (2016) <https://doi.org/10.1007/s11042-015-2643-0>
- [23] Hassani, H., Mahmoudvand, R.: Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics* **01** (2013) <https://doi.org/10.1142/S2335680413500051>
- [24] Emde-Boas, P.: Another NP-complete Partition Problem and the Complexity of Computing Short Vectors in a Lattice. Report. Department of Mathematics. University of Amsterdam. Department, Univ., Amsterdam (1981)
- [25] Grafarend, E., Zwanzig, S., Awange, J.: Integer Least Squares, pp. 579–603. Springer, Cham (2022)
- [26] Nahid, A.-A., Sikder, N., Rafi, I.: KU-HAR: An Open Dataset for Human Activity Recognition. Mendeley Data (2021). <https://doi.org/10.17632/45f952y38r.5>

Appendix A Additional figures

The Fig. A1, A2, A3 show the forecast quality depending on the tensor ranks for the MAPE and MSE metrics. Fig. A4 shows the CPD computation error depending on the tensor ranks. This error is an estimation of how close the computed CPD is to the true one (1).

The Fig. A5 shows the decomposition quality depending on the tensor ranks for the \overline{RHE} metric.

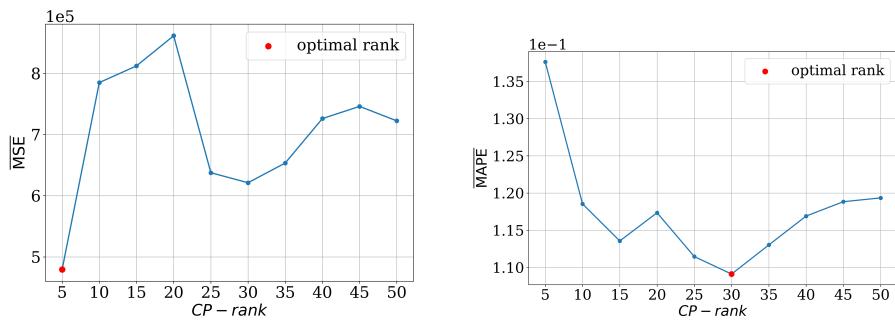


Figure A1 $\overline{\text{MSE}}$ and $\overline{\text{MAPE}}$ metrics for the tSSA forecast depending on the CPD rank. An optimal point is marked with red. Electricity data

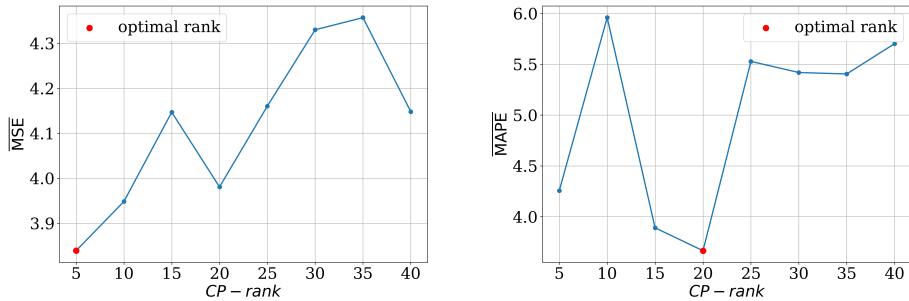


Figure A2 $\overline{\text{MSE}}$ and $\overline{\text{MAPE}}$ metrics for the tSSA forecast depending on the CPD rank. An optimal point is marked with red. Inertial unit data

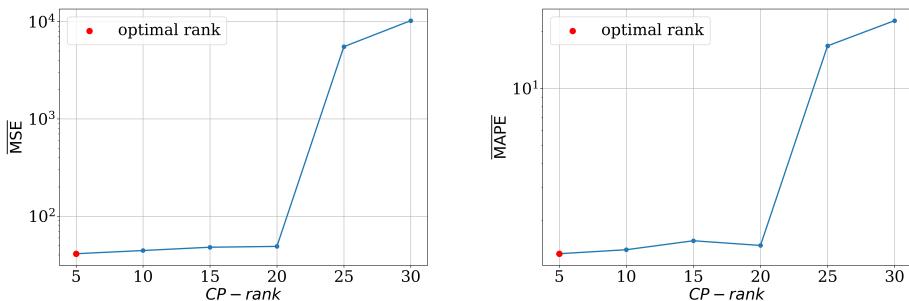


Figure A3 $\overline{\text{MSE}}$ and $\overline{\text{MAPE}}$ metrics for the tSSA forecast depending on the CPD rank. An optimal point is marked with red. Weather data

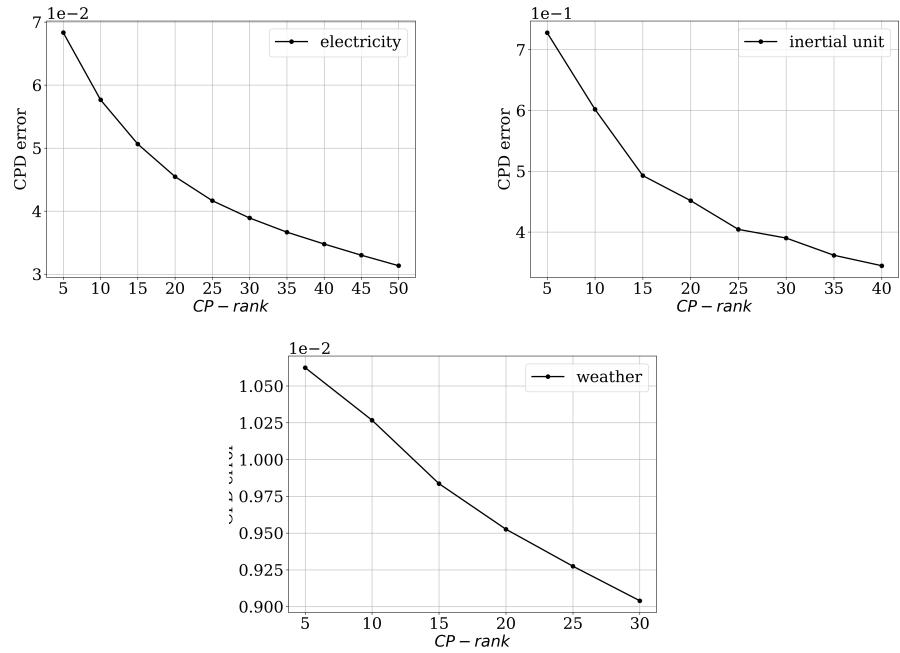


Figure A4 Relative CPD computation error depending on the CPD rank. The left is for the electricity data, the right is for the inertial unit data, the bottom is for the weather data.

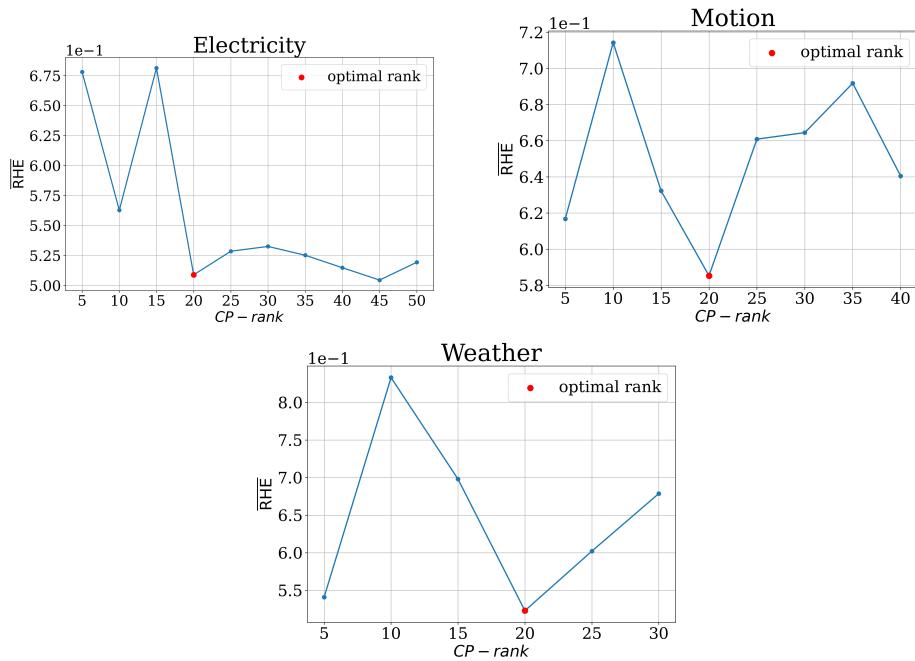


Figure A5 The RHE metrics depending on the CPD rank. The left is for the electricity data, the right is for the inertial unit data, the bottom is for the weather data. An optimal point is marked with red