A method of extracting latent semantic components from noisy categorical time-series data applied to human sleep stage data

Ikuhiro Yamaguchi, Akifumi Kishi, Fumiharu Togo and Yoshiharu Yamamoto

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

April 25, 2019
Method to extract latent semantic components from noisy categorical time-series data applied to human sleep stage data

Ikuhiro Yamaguchi  
Graduate School of Education  
The University of Tokyo  
Tokyo, Japan  
ikuhiro@p.u-tokyo.ac.jp

Akifumi Kishi  
Graduate School of Education  
The University of Tokyo  
Tokyo, Japan

Fumiharu Togo  
Graduate School of Education  
The University of Tokyo  
Tokyo, Japan

Yoshiharu Yamamoto  
Graduate School of Education  
The University of Tokyo  
Tokyo, Japan

Abstract—A method to extract latent semantic components from noisy categorical time-series data based on the Takens time-delay-embedding method and singular value decomposition is presented. A demonstration of this method to analyze a sleep stage time-series is demonstrated. The first component extracted by this method, i.e., the component with the largest singular value can be considered the circadian rhythm. The second component, which exhibits damping oscillation, can be interpreted as the ultradian rhythm, and matched with the moving-averaged \( c_2 \); which is an estimation of the cortico-thalamo-cortical loop strength calculated from the corresponding electroencephalogram data using the method previously reported. The sleep stage is treated as a nominal variable instead of an ordinal variable in this method; however, the quantitative variation of sleep state is extracted from the sleep stage time-series. We believe that this result suggests the validity and usefulness of both the methods, i.e., the method reported in the present study and the method reported in a previous study.

Index Terms—categorical time-series, latent semantic component, sleep, electroencephalography, cortico-thalamo-cortical loop

I. INTRODUCTION

Sleep is directly related to overall wellness and has been actively studied by medical scientists or medical engineers, using multifarious methods [1], [2]. However, in such methods, the categorization of sleep state into stages has been basic [1], and it is recognized that the dynamics of sleep stage transition is important [3]. The American Academy of Sleep Medicine (AASM) divides sleep into five stages: Wake, N1 (= Non rapid eye movement sleep 1), N2, N3, and REM (= rapid eye movement) [4]. From the viewpoint of signal processing, the sleep state stages can be interpreted as the reduction of high-dimensional data into low-dimensional data. The high-dimensional data is physiological, which includes the electroencephalogram (EEG), electro-oculography (EOG), and electromyogram (EMG). The low-dimensional data is qualitatively categorized data. Data reduction such as this is currently being actively researched in the field of machine learning or artificial intelligence [5].

However, further analysis after obtaining the sleep stage time-series is not easy because sleep stage data is categorical and not quantitative. That is, obtaining a sleep stage time-series has the advantage of dimensional reduction and the disadvantage that it is not quantitative. Therefore, a few quantification methods have been proposed. For example, in sleep restoration gain (SRG) [6], REM, N1, N2, and N3 are quantified as 0, 1/1.5, 1, and 1.5, respectively. Wake is quantified as -1.5 if the previous stage is not wake and -1 if the previous stage is wake. Using this method, we can quantize sleep quality. However, the method is arbitrary in the translation of categorical data into qualitative data. On the other hand, some probabilistic methods such as Markov model [3], [7]–[11] or spectral entropy [12]–[17] treat the sleep stage time-series just as categorical and not quantitative data.

The Markov model that treats the sleep stage time-series as pure Markov chain was proposed at first [7]. However, the model could not reproduce the observations well, and an improved model that considers the time duration of each stage as a probabilistic process was proposed [8]. This model agrees reasonably with the observations, and it has been applied to studies on chronic fatigue syndrome [3], sleep apnea [9], chronic fatigue syndrome with or without fibromyalgia [10], ultradian REM sleep rhythm [11], etc. Recently, multiorder Markov models were investigated to analyze sleep stage dynamics more accurately [11], [15].

For spectral entropy, an evaluation method using Walsh function or Haar function was proposed [12], and its relation with sleep fragmentation, daytime sleepiness [13], neonatal neurologic function [14], [16], sleep disorder [15], etc. have been reported.

However, ‘trends’ ( [18], [19] ) of sleep stage dynamics, such as oscillation with a period of approximately 90 min or gradual change from deep to light sleep, cannot be analyzed well by these methods. Hence, we propose a method to extract latent semantic components from noisy categorical time-series data based on Takens time-delay-embedding method [20] and singular value decomposition (SVD) [21]–[30]. While Fourier transform, moving averages (MA) [18], [19] or empirical mode decomposition (EMD) [31], [32] are suitable for quantitative data time-series decomposition, SVD appears to be suitable...
for qualitative data time-series decomposition, which is now actively studied in the field of natural language analysis [22]. In this study, we show the calculation process of our method in section II, and show an application example of the method in section III.

II. METHOD

A. Data Expression

First, following AASM, we divide sleep states into five stages: Wake, N1, N2, N3, and REM. We express these stages as \((1\ 0\ 0\ 0\ 0), (0\ 1\ 0\ 0\ 0), (0\ 0\ 1\ 0\ 0)\) and \((0\ 0\ 0\ 0\ 1)\), respectively. Because we treat the stage data as nominal data, the order of this representation is arbitrary. The arbitrariness is ensured because the permutation of the data matrix columns is represented by an orthogonal matrix as described later. This expression is the same as that in [9]. Let \(X\) be a data matrix with row numbers corresponding to time, and row vectors corresponding to the expressed sleep stages. The following is an example.

\[
X = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 1 & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots 
\end{bmatrix}.
\]

(1)

We express each column vector of \(X\) in lowercase bold;

\[
X = [x_1 \ x_2 \ x_3 \ x_4 \ x_5].
\]

(2)

We express the \(t\)-th row component of the column vector \(x_j\) as \(x_j^{(t)}\). In this study, we assume the sleep stage time-series data as a 24 h cyclic data with the time interval of 30 s, and set \(T = 24 \times 60 \text{ min} \times 60 \text{ s} = 2880\).

B. Normalization

Next, we calculate the normalized data matrix \(Y\) as

\[
\bar{x}_j = \frac{1}{T} \sum_{t=1}^{T} x_j^{(t)}, \quad j = 1, 2, \ldots, 5.
\]

(3)

\[
\sigma = \left[ \frac{1}{T} \sum_{t=1}^{T} \left( x_j^{(t)} - \bar{x}_j \right)^2 \right]^{\frac{1}{2}},
\]

(4)

\[
y_j^{(t)} = \frac{1}{\sigma} \left( x_j^{(t)} - \bar{x}_j \right),
\]

(5)

\[
Y = [y_1 \ y_2 \ y_3 \ y_4 \ y_5].
\]

(6)

C. Embedding

Next, we embed \(Y\) into a high dimensional phase space and obtain an extended data matrix \(Z\) using Takens time-lag method [20];

\[
Z_{(m)}^{(t)} = Y^{(t-m)}, \quad m = 1, 2, \ldots, M - 1.
\]

(7)

\[
Z = [Z_{(0)} \ Z_{(1)} \ \cdots \ Z_{(M-1)}].
\]

(8)

In this study, we set \(M = 64\), corresponding to the 32-min time-window.

D. Singular Value Decomposition

We use the SVD [21]–[30];

\[
Z = USV^T,
\]

(9)

where \(U\) is \(T \times 5M\) orthogonal matrix, \(V\) is \(5M \times 5M\) orthogonal matrix, and \(S\) is \(5M \times 5M\) diagonal matrix with singular values \(\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{5M} \geq 0\);

\[
S = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \lambda_{5M}
\end{bmatrix}.
\]

(10)

E. Extracting Latent Semantic Components

We term the \(j\)-th column vector of matrix \(U\), \(u_j\), the \(j\)-th latent semantic component [22];

\[
U = [u_1 \ u_2 \ \cdots \ u_{5M}].
\]

(11)

They consist a basis of the phase space that satisfies the orthonormal condition:

\[
u_i^T u_j = \delta_{ij}.
\]

(12)

where \(\delta_{ij}\) is the Kronecker delta. From the following equation, it can be seen that the matrix \(V\) acts the data matrix \(Z\) as an orthonormal coordinate transformation to obtain the latent semantic components.

\[
ZV = [\lambda_1 u_1 \ \lambda_2 u_2 \ \cdots].
\]

(13)

F. Denoising

We can denoise data matrix \(Z\) to \(\tilde{Z}\) by removing small singular-valued components as noise. In this study, we remove \(u_j\), \(j \geq 3\);

\[
\tilde{Z} = U \tilde{S} V^T.
\]

(14)

\[
\tilde{S} = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \lambda_{5M}
\end{bmatrix}.
\]

(15)

We can define the denoised data vector \(\tilde{y}_j\) as follows:

\[
\tilde{Z} = [\tilde{Z}_{(0)} \ \tilde{Z}_{(1)} \ \cdots \ \tilde{Z}_{(M-1)}],
\]

(16)

\[
[y_1 \ y_2 \ \cdots \ y_5] = \tilde{Z}_0.
\]

(17)

III. APPLICATION EXAMPLE

A. Latent Semantic Components

Fig. 1 shows an example of a sleep stage time-series, the first and second latent semantic components, and the moving-averaged cortico-thalamo-cortical loop strength \(c_2\), which is estimated from EEG obtained by the method previously proposed by us [17], [33]–[35]. The first and second latent semantic components can be interpreted as the circadian rhythm and ultradian rhythm, respectively. Note that the proposed method can extract some damping oscillation-like variation of the sleep state even if the method treats sleep stage data as nominal.
data not ordinal data. Therefore, they were termed as latent semantic components (LSC). We believe that the similarity between the second LSC and $c_2$ suggests the validity and practicality of both methods, i.e., the method reported in the present study and the method reported previously by us.

B. Denoised vectors

Fig. 2 shows the denoised vectors (Wake: $\tilde{y}_1$, N1: $\tilde{y}_2$, N2: $\tilde{y}_3$, N3: $\tilde{y}_4$, and REM: $\tilde{y}_5$) calculated from the data for Fig.1. During sleep (duration: 0 to 8 h), N1, N2, N3, and REM appear to be cyclic signals with the same period, approximately 90 min, and phase shift, i.e., they can be interpreted as signals belonging to the same cluster ‘Sleep.’ In contrast, Wake does not belong to the Sleep cluster. This result agrees well with previous findings:

- WSE2 is closer to CSE than WSE5 [17]. Here, WSE2 is Walsh Spectral Entropy calculated from binarized (Sleep/Wake) sleep stage time series, while WSE5 is calculated from original five stage data. CSE is a proposed spectral entropy that is calculated from EEG based on a cortico-thalamo-cortical loop model [17].
- The distribution of $c_2$ value obtained from 26 subjects could approximate by the Gaussian Mixture Model with two peaks corresponding ‘Sleep’ and ‘Wake’ [34].

IV. DISCUSSION

A. Novelty of this study

The proposed method is based on a singular value decomposition (SVD) and time-lag embedding, that have been studied for many decades and applied in various fields such as finance, physiology, and genomics [21]–[30]. In sleep research, similar methods have been used to analyze the physiological data especially EEG data. However, to the best of our knowledge, this is the first paper that proposes the application of such a method to the sleep stage time-series. SVD is basically the same as Latent Semantic Analysis (LSA) or Principal Component Analysis (PCA), and is sometimes applied in the preprocessing stage in machine learning or artificial intelligence. But there is no SVD application to sleep stage time-series, maybe because sleep stage data has some aspects of duality, as outlined below.

B. Input and Output

The technology for discriminating the sleep stage from the biomedical signal has been actively studied using machine learning and artificial intelligence. The goal is to output the sleep stage, and few studies have actively analyzed the sleep stage. Analyses performed in the medical field are mainly simple calculations such as those for sleep efficiency. The sleep stage seems to be becoming standard data in medical practice. We expect that the proposed method advances mathematical analysis in which sleep stage data is regarded as input.

C. Quantitative and Qualitative

Although the source data for determining the sleep stage is quantitative data measured by electronic devices, the sleep stage comprises categorical qualitative data. The sleep stage data looks like ordinal data like N1<N2<N3. However, the difference between Non-REM and REM is not quantitative, but completely qualitative. In this paper, we presented a method to quantify the sleep stage using the dynamics of its time-series, by referring to Latent Semantic Analysis (LSA). LSA is used in various fields and may be suitable for quantifying qualitative data as it is the starting point of topic model in natural language analysis. Verification of this quantification method would be an next issue.

D. Probabilistic noise and Deterministic trend

Although we described this as a denoising method in this paper, it is possible to utilize this as a detrending method that is important for exact theoretical analysis [18], [19]. This theoretical extension is the next issue.

E. Empirical data analysis and Theoretical modeling

Comprehensive study of sleep requires mathematical models that represent functions of neural circuits such as suprachiasmatic nuclei and corticothalamic loops [37]–[39]. Herein, observational data is analyzed, but it has also been shown that the analysis results are consistent with the corticothalamic model. This research could be useful in creating more reliable physiological and mathematical sleep models.
ACKNOWLEDGMENT

This work is partly supported by JSPS KAKENHI Grant Nos. 18K17887 to IY.

REFERENCES