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 oscilation 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 $(1\,0\,0\,0\,0)$, $(0\,1\,0\,0\,0)$, $(0\,0\,1\,0\,0)$, $(0\,0\,0\,1\,0)$ 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.

$$X = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ & & \vdots & & \\ 0 & 0 & 1 & 0 & 0 \\ & & \vdots & & \end{bmatrix}. \tag{1}$$

We express each column vector of X in lowercase bold;

$$X = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \boldsymbol{x_3} & \boldsymbol{x}_4 & \boldsymbol{x}_5 \end{bmatrix}. \tag{2}$$

We express the t-th row component of the column vector $\boldsymbol{x}_j$ 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 $T = 24\,\mathrm{h} \times 60\,\mathrm{min} \times 60\,\mathrm{s} \div 30\,\mathrm{s} = 2880$.

B. Normalization

Next, we calculate the normalized data matrix Y as

$$\bar{x}_j = \frac{1}{T}\sum_{t=1}^{T} x_j^{(t)},\ j = 1, 2, \cdots, 5. \tag{3}$$

$$\sigma = \left[\frac{1}{T}\sum_{t=1}^{T}\left(x_j^{(t)} - \bar{x}_j\right)^2\right]^{\frac{1}{2}}, \tag{4}$$

$$y_j^{(t)} = \frac{1}{\sigma}\left(x_j^{(t)} - \bar{x}_j\right), \tag{5}$$

$$Y = \begin{bmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \boldsymbol{y_3} & \boldsymbol{y}_4 & \boldsymbol{y}_5 \end{bmatrix}. \tag{6}$$

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];

$$Z_{(m)}^{(t)} = Y^{(t-m)},\ m = 1, 2, \cdots, M-1. \tag{7}$$

$$Z = \begin{bmatrix} Z_{(0)} & Z_{(1)} & \cdots & Z_{(M-1)} \end{bmatrix}. \tag{8}$$

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

D. Singular Value Decompositioin

We use the SVD [21]–[30];

$$Z = USV^{\mathrm{T}}, \tag{9}$$

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 $\lambda_1 \geqq \lambda_2 \cdots \geqq \lambda_{5M} \geqq 0$;

$$S = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_{5M} \end{bmatrix}. \tag{10}$$

E. Extracting Latent Semantic Components

We term the j-th column vector of matrix U, $\boldsymbol{u}_j$, the j-th latent semantic component [22] ;

$$U = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \cdots & \boldsymbol{u}_{5M} \end{bmatrix}. \tag{11}$$

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

$$\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{u}_j = \delta_{ij}. \tag{12}$$

where δ_{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.

$$ZV = \begin{bmatrix} \lambda_1\boldsymbol{u}_1 & \lambda_2\boldsymbol{u}_2 & \cdots \end{bmatrix}. \tag{13}$$

F. Denoising

We can denoise data matrix Z to $\tilde{Z}$ by removing small singular-valued components as noise. In this study, we remove $\boldsymbol{u}_j,\ j \geqq 3$;

$$\tilde{Z} = U\tilde{S}V^{\mathrm{T}}. \tag{14}$$

$$\tilde{S} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & 0 & \\ & & & \ddots \\ & & & & 0 \end{bmatrix}. \tag{15}$$

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

$$\tilde{Z} = \begin{bmatrix} \tilde{Z}_{(0)} & \tilde{Z}_{(1)} & \cdots & \tilde{Z}_{(M-1)} \end{bmatrix}, \tag{16}$$

$$\begin{bmatrix} \tilde{y}_1 & \tilde{y}_2 & \cdots & \tilde{y}_5 \end{bmatrix} = \tilde{Z}_0. \tag{17}$$

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 the 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].

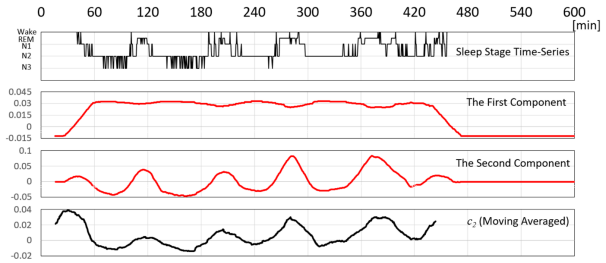

Fig. 1. Example of sleep stage time-series, first and second latent semantic components, and moving-averaged cortico-thalamo-cortical loop strength c_2 estimated based on an EEG.

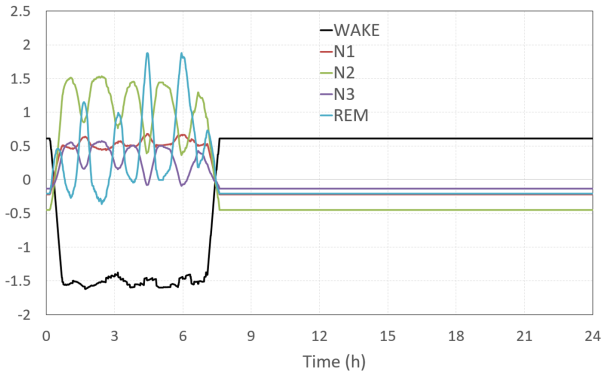

Fig. 2. Example of denoised vectors (Wake: $\tilde{y}_1$, N1: $\tilde{y}_2$, N2: $\tilde{y}_3$, N3: $\tilde{y}_4$, and REM: $\tilde{y}_5$) calculated from the data for Fig.1.

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 esptcially 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

[1] A. Rechtschaffen, ''A manual of standardized terminology, technique and scoring system for sleep stages of human subjects," Public Health Service, 1968.

[2] F. Mendonça, SS Mostafa, F. Morgado-Dias, AG Ravelo-García, T. Penzel, ''A Review of Approaches for Sleep Quality Analysis," IEEE Access, 7, pp. 24527-24546, 2019.

[3] A. Kishi, ZR Struzik, BH. Natelson, F. Togo, Y. Yamamoto, ''Dynamics of sleep stage transitions in healthy humans and patients with chronic fatigue syndrome," American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 294(6), pp. R1980-R1987, 2008.

[4] C. Iber, ''The AASM manual for the scoring of sleep and associated events: Rules," Terminology and Technical Specification, 2007.

[5] A. Supratak, H. Dong, C. Wu, Y. Guo, ''DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(11), 1998–2008, 2017.

[6] I. S. Badreldin, A. A. Morsy, ''Consistency of Sleep Restoration Gain (SRG) as a measure for assessing sleep quality," 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 4664–4667, 2012.

[7] WWK Zung, TH Naylor, D GIANTURC WP Wilson, ''A Markov chain model of sleep EEG patterns," ELECTROENCEPHALOGRAPHY AND CLINICAL NEUROPHYSIOLOGY, 19(1), 105, 1965.

[8] MC Yang, CJ Hursch, ''The use of a semi-Markov model for describing sleep patterns," Biometrics, 667–676, 1973.

[9] JW Kim, J-S Lee, PA Robinson, D-U Jeong, ''Markov analysis of sleep dynamics", Physical review letters, 102(17), 178104, 2009.

[10] A. Kishi, BH. Natelson, F. Togo, Z.R. Struzik, D.M. Rapoport, and Y. Yamamoto, Sleep-stage dynamics in patients with chronic fatigue syndrome with or without fibromyalgia, Sleep, 34(11), 1551 (2011).

[11] A. Kishi, I. Yamaguchi, F. Togo, and Y. Yamamoto, "Markov modeling of sleep stage transitions and ultradian REM sleep rhythm," Physiological Measurement, vol. 39(8), 084005, 2018.

[12] D.S. Stoffer, M.S. Scher, G.A. Richardson, N.L. Day, and P.A. Coble, A Walsh-Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling, Journal of the American Statistical Association, 83, 404, 954 (1988).

[13] M.R. Kirsch, K. Monahan, J. Weng, S. Redline, and K.A. Loparo, Entropy-based measures for quantifying sleep-stage transition dynamics: relationship to sleep fragmentation and daytime sleepiness, IEEE Transactions on Biomedical Engineering, 59,3, 787 (2012).

[14] R.A. Shellhaas, J.W. Burns, J.D.E. Barks, and R.D. Chervin, Quantitative sleep stage analyses as a window to neonatal neurologic function, Neurology, 82, 5, 390 (2014).

[15] A. Schlemmer, U. Parlitz, S. Luther, N. Wessel, and T. Penzel, Changes of sleep-stage transitions due to ageing and sleep disorder, Philosophical Transactions of the Royal Society, A: Mathematical, Physical and Engineering Sciences, 373, 2034, 20140093 (2015).

[16] RA Shellhaas, JW Burns, F Hassan, MD Carlson, JDE Barks, RD Chervin, ''Neonatal sleep–wake analyses predict 18-month neurodevelopmental outcomes," Sleep, 40(11), zsx144, 2017.

[17] I. Yamaguchi, A. Kishi, F. Togo, and Y. Yamamoto, "Spectral Analysis Method for Sleep-state Cycle Based on the Cortico-Thalamo-Cortical Loop Strength Estimation," International Conference on Noise and Fluctuations (ICNF), 2017, pp.1-4, 2017.

[18] K. Kiyono, "Theory and applications of detrending-operation-based fractal-scaling analysis," International Conference on Noise and Fluctuations (ICNF), 2017, pp. 1-4, 2017

[19] M. Hll, K. Kiyono, H. Kantz, ''Theoretical foundation of detrending methods for fluctuation analysis such as detrended fluctuation analysis and detrending moving average," Phys. Rev. E 99, 033305, 2019.

[20] F. Takens, ''Detecting strange attractors in turbulence," Dynamical systems and turbulence, Warwick 1980, pp. 366-381, 1981.

[21] DW Tufts, R. Kumaresan, I. Kirsteins, ''Data adaptive signal estimation by singular value decomposition of a data matrix," Proceedings of the IEEE, 70(6), pp. 684-685, 1982.

[22] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, R. Harshman, ''Using latent semantic analysis to improve access to textual information," Proceedings of the SIGCHI conference on Human factors in computing systems, 281–285, 1988.

[23] S. Nishisato, ''Gleaning in the field of dual scaling," Psychometrika, 61(4), 559-599, 1996.

[24] A. Ziehe, KR Müller, ''TDSEPan efficient algorithm for blind separation using time structure," International Conference on Artificial Neural Networks, pp. 675-680, 1998.

[25] SJ Roberts, W. Penny, I. Rezek, ''Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing," Medical & biological engineering & computing, 37(1), pp. 93-98, 1999.

[26] B. Podobnik, D. Wang, D. Horvatic, I. Grosse, HE Stanley, ''Time-lag cross-correlations in collective phenomena," EPL (Europhysics Letters), 90(6), 68001, 2010.,

[27] AM Sabatini, ''Analysis of postural sway using entropy measures of signal complexity," Medical and Biological Engineering and Computing, 38(6), pp. 617-624, 2000.

[28] BR Greene, S. Faul, WP Marnane, G. Lightbody, I. Korotchikova, GB Boylan, ''A comparison of quantitative EEG features for neonatal seizure detection," Clinical Neurophysiology, 119(6), pp. 1248-1261, 2008.

[29] P. Caraiani, ''The predictive power of singular value decomposition entropy for stock market dynamics," Physica A: Statistical Mechanics and its Applications, 393, pp. 571-578, 2014.

[30] J. Xue, J. and Li, D. Yu, M. Seltzer, Y. Gong, ''Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6359-6363, 2014.

[31] Z. Wu, NE Huang, ''Ensemble empirical mode decomposition: a noise-assisted data analysis method," Advances in adaptive data analysis, 1(01), pp. 1-41, 2009. Shimizu

[32] E. Shimizu, T. Nakamura, J. Kim, K. Yoshiuchi, Y. Yamamoto, ''Application of Empirical Mode Decomposition to Mother and Infant Physical Activity," Methods of information in medicine, 57(03), pp. 152-157, 2018.

[33] I. Yamaguchi, Y. Ogawa, H. Nakao, Y. Jimbo, K. Kotani, Linear analysis of the corticothalamic model with time delay, Electronics and Communications in Japan, 97(8), 32-44 (2014).

[34] I. Yamaguchi, A. Kishi, F. Togo, T. Nakamura, Y. Yamamoto, ''A Robust Method with High Time Resolution for Estimating the Cortico-Thalamo-Cortical Loop Strength and the Delay when Using a Scalp Electroencephalography Applied to the Wake-Sleep Transition," Methods of information in medicine, 57(03), 122–128, 2018.

[35] J.W. Kim and P.A. Robinson, Compact dynamical model of brain activity, Physical Review E, 75, 3, 031907 (2007).

[36] M Niknazar, GP Krishnan, M Bazhenov, SC Mednick, ''Coupling of thalamocortical sleep oscillations are important for memory consolidation in humans," PloS one, 10(12), e0144720, 2015.

[37] S. Postnova, ''Sleep Modelling across Physiological Levels," Clocks & Sleep, 1(1), 166–184, 2019.

[38] I. Yamaguchi, T. Isomura, H. Nakao, Y. Ogawa, Y. Jimbo, K. Kotani, ''Suppression of macroscopic oscillations in mixed populations of active and inactive oscillators coupled through lattice Laplacian," Journal of the Physical Society of Japan, 88(5), 054004, 2019.

[39] DP Yang, L McKenzie-Sell, A Karanjai, PA Robinson, ''Wake-sleep transition as a noisy bifurcation," Physical Review E, 94(2), 022412, 2016.