http://dx.doi.org/10.5392/IJoC.2013.9.4.001

Exploiting Patterns for Handling Incomplete Coevolving EEG Time Series

Ngoc Anh Nguyen Thi
School of Electronics and Computer Engineering
Chonnam National University, Gwangju 500-757, South Korea

Hyung-Jeong Yang*
School of Electronics and Computer Engineering
Chonnam National University, Gwangju 500-757, South Korea

Sun-Hee Kim
School of Computer Science, Carnegie Mellon University
Pittsburgh, Pennsylvania, 15213, USA.

ABSTRACT

The electroencephalogram (EEG) time series is a measure of electrical activity received from multiple electrodes placed on the scalp of a human brain. It provides a direct measurement for characterizing the dynamic aspects of brain activities. These EEG signals are formed from a series of spatial and temporal data with multiple dimensions. Missing data could occur due to fault electrodes. These missing data can cause distortion, repudiation, and further, reduce the effectiveness of analyzing algorithms. Current methodologies for EEG analysis require a complete set of EEG data matrix as input. Therefore, an accurate and reliable imputation approach for missing values is necessary to avoid incomplete data sets for analyses and further improve the usage of performance techniques. This research proposes a new method to automatically recover random consecutive missing data from real world EEG data based on Linear Dynamical System. The proposed method aims to capture the optimal patterns based on two main characteristics in the coevolving EEG time series: namely, (i) dynamics via discovering temporal evolving behaviors, and (ii) correlations by identifying the relationships between multiple brain signals. From these exploits, the proposed method successfully identifies a few hidden variables and discovers their dynamics to impute missing values. The proposed method offers a robust and scalable approach with linear computation time over the size of sequences. A comparative study has been performed to assess the effectiveness of the proposed method against interpolation and missing values via Singular Value Decomposition (MSVD). The experimental simulations demonstrate that the proposed method provides better reconstruction performance up to 49% and 67% improvements over MSVD and interpolation approaches, respectively.

Key words: *Multivariate Time Series, Electroencephalogram (EEG), Handling Missing Values, Interpolation, Linear Dynamical System, Kalman Filter, MSVD, Expectation maximization.*

1. INTRODUCTION

Human brain is one of the most vital organs of humans, controlling the coordination of muscles and nerves. To communicate between human brains and computers, electroencephalogram (EEG) based on brain computer interface is provided.

EEG signals reflect electrical activities of a brain. EEG signals have various clinical and advanced scientific applications such as medicine, pharmacy, psychology, linguistics, and biology. Due to the usefulness of EEG signals, the study of brain electrical activities, through electroencephalogram (EEG) records, is one of the most important tools for the diagnosis of neurological diseases [1].

To have an accurate insight and improved understanding of the mechanisms causing widespread brain disorder, careful analyses of EEG records are necessary. Large amounts of EEG signal processing have been investigated recently for distinguishing epileptic seizures, emotions, and brain functions. For example, Guler explored the ability of desired and trained Elman recurrent neural networks, combined with the Lyapunov exponents on the EEG signals [1]. Polat proposed a hybrid system to detect epileptic seizure in EEG signals via two steps: feature extraction using Fourier

* *Corresponding author, Email: hjyang@jnu.ac.kr*
Manuscript received Sep. 02, 2013; revised Nov. 20, 2013; accepted Nov. 29, 2013

transform, and decision making using the decision tree classifier [2]. Adeli et al. investigated discrete Daubechies and harmonic wavelets for the analysis of epileptic EEG records [3]. Besides, many investigators have developed different methods to better understand the dynamics of human brains through EEG analysis [4]-[7]. Since EEG signals are recorded from multiple electrodes placed on different locations along the scalp surface, the problem of missing values can be encountered frequently due to the disconnection of particular electrodes, lost signals, or the failure to report some of the measurements in time [8]. Unfortunately, most of the above approaches assume that the EEG signal is a complete data set as input. With the occurrence of missing values, they can cause inaccurate usage, distort results, or even degrade the performance of techniques.

Missing values are often simply discarded or ignored because they are deemed unimportant for analysis. However, if the presence of missing values is consecutive for a long period of time, the performance of methodologies on these dataset will not provide the accurate results. Hence, missing value estimation in EEG time series is necessary for the pre-processing in order to boost accuracy and reliable usage requirement .

On the other hand, the collected EEG dataset exhibits a large spatial and temporal relativity in their values. The size of the data is often a massively large matrix containing multiple dimensions. Since there are correlations among the variables of a large matrix expression of EEG signal over time, the necessity of variable selection is emphasized. It is not advisable to try to use the whole dataset for a process that would require only a small piece to be completed adequately. Therefore, it is better to mine meaningful variables from the large databases in order to save memory, processing power, and time as well.

To handle the above challenges successfully, we propose a new approach which satisfactorily fits on the available data when missing values are present. Our prime objective is to mine correlations and evolving behaviors of multiple electrodes by automatically identifying a few hidden variables, and then to exploit their dynamics for solving the problem of missing observation. Correlation implies that the observed dimensions of multiple electrodes are not dependent. Therefore, missing values can be deduced from others through hidden variables. Evolving behavior denotes that missing values can be estimated effectively based on neighbors' observations of next time ticks and following their moving trends. To evaluate the effectiveness of the proposed method by considering accuracy, reliability, and complexity aspects, this paper demonstrates the performance of imputation for consecutive missing observations on two real different datasets of electroencephalogram (EEG) signals. The proposed method can be effective to capture a few hidden variables automatically as well as to compactly illustrate how to learn their dynamics for solving consecutive missing values. Moreover, its computational time scales up linearly with the duration of the sequences. We compare the performance of the proposed method to the interpolation and missing singular value decomposition methods known as MSVD.

The outline of the remainder of this paper is as follows; In section 2, we provide a summary of the available articles and research literatures related to time series with missing values. Section 3 describes the materials and proposed model setup. Section 4 illustrates detailed object experiments and experimental evaluations. Finally, in section 5, we summarize the whole paper and point out some possible future research directions.

2. RELATED WORK

In many real-world applications, time series have been given considerable attention within a variety of domains. They spread from network traffic data, currency exchange rates, sensor measurement, to biomedical and so on. In reality, the data set may contain missing observations since the process of data collection is not perfect due to poor record keeping, lost records, etc. Ignoring of missing data causes the loss of useful information of datasets. Therefore, numerous advanced filling mechanisms are proposed to overcome this problem in time series data.

One of the most straightforward procedures is to replace each missing variable with simple methods such as calculating appropriate mean values. Another alternating method for filling missing values is an interpolation method, which is related to the handling of missing elements using a curve fitting, known as linear interpolation and splines. These methods exploit the smoothness in a single time sequence. They can estimate the success for a short interval of missing values based on continuity of the sequence. Details of these approaches and its applicability can be found in [9]-[11]. However, these methods will either become invalid or face a big challenge when the observation gap of missing values is large. Furthermore, these approaches discard any relationship between the variables over time.

The well- known techniques related to dimension reduction and latent variables, namely Singular Value Decomposition (SVD) and Principal Component Analysis (PCA), could fill missing values through discovering correlations in multiple time sequences. They recover the possible missing values from estimation of other observations in one sequence of multivariate data. Many of the previous works has applied these methods for solving the presence of missing observations [12]-[14]. Kim *at el.* [15] proposes an incremental approach by updating the weights of PCA for every new data and the expectation-maximization to recover missing values. This model captured successfully correlations among different time sequences for imputing single missing value at various positions. However, this method ignores time-evolving patterns since its characteristics are not designed for tracking the ordering of the rows or the columns. Guha [16] identifies the correlations between multiple data streams by improving the discovery of correlations. They perform a dimensionality reduction with SVD periodically. However, this approach faces a challenge due to SVD re-computation and thus, it cannot easily handle the missing data. Papadimitriou [17] proposes SPIRIT model which can incrementally find correlations and hidden variables based on

PCA and uses them to recover missing values. Nati and Jaakkola [18] provide a simple EM algorithm that factors a dataset to low- rank matrices and approximate the missing value problem. Olga Troyanskaya [19] proposes a method by using an expectation-maximization-like procedure. It performs SVD over all complete columns, regresses incomplete columns against SVD to estimate missing values, then re-factors and re-imputes the completed data until a fix point is reached. This is extremely slow and only works if only very few values are missing.

The recent intuitive mechanism for handling missing values by using a statistical model is known as the linear dynamical system, which can be used to estimate values for missing time points. Phong and Singh [20] demonstrate the efficiency of their method by exploring the linear dynamical system model for gene expression with missing values. Aristidou [21] uses Kalman Filter, which predicts the occluded marker position on human motion dataset. Dorfmuller and Kalman [22]-[23] show how to use the previous maker position and a skeletal model to estimate the missing marker locations using the extended Kalman Filter. However, these models become ineffective when the marker's missing positions are held for a long period of time.

It can be observed from the literatures that there are a variety of techniques available for estimating missing values for time series data. However, only few literatures show the performance on different positions for the missing data, randomly in particular, for long-interval consecutive missing values. This research aims to propose a new approach to handle missing values automatically. Moreover, its computation time is significantly considered with the linear on the duration of the sequences.

3. PROPOSED METHOD

In this context, a brief introduction of the Linear Dynamic System (LDS), also known as Kalman Filter used to model multiple time series, is first presented in order to enhance how to find the hidden dynamics in multiple EEG time series signals. Then, a new proposed model is built to impute missing values in multiple coevolving EEG time series.

3.1 A Linear Dynamical System for EEG Time Series

A sequence of EEG is a multi-dimensional data since multiple electrodes are used to record the electrical activity along the scalp surface. A dynamical system can be modeled by a sequence of multi-dimensional EEG signals, denoted by $Y = \{y_1, y_2, y_3..., y_T\}$, where each vector y_t denotes the data at each time ticks *t = 1, 2, 3… T* of dimensionality of *M*. This means that data from EEG time series can be presented by a matrix $Y_{M\times T}$ of the variables M and observed time ticks T. We consider EEG time series data are obtained from EEG signals in such a dynamical system. It builds a statistical model to represent the state of the hidden variables which are evolving to a linear transformation leading to the observed numerical time sequences. The model can learn the dynamics of the time series data [24], [25]. It captures the correlations among multiple electrodes by choosing a proper number of hidden variables. In particular, LDS for multi-dimensional EEG time sequence is modeled by the following equations;

$$z_1 = \mu_0 + \omega_0 \quad (1)$$

$$z_{n+1} = A \cdot z_n + \omega_n \quad (2)$$

$$y_n = C \cdot z_n + \varepsilon_n, \quad (3)$$

where $\theta = \{\mu_0, Q_0, A, Q, C, R\}$ is the set of parameters. μ_0 is an initial state for hidden variables of the whole system. Vector y_n and z_n denote observed data sequences and hidden variables at time *t*, respectively. The transition dynamic matrix *A* relates to the transition of the state from the current time tick to the next time tick with noise $\{\omega_n\}$. Matrix *C* is the observation projection with the noise $\{\varepsilon_n\}$ at each time t, meaning that the series of hidden variables z_n are evolving over time ticks with linear transition matrix *A*. Moreover, the observed data sequences y_n are generated from these series of hidden variables with a linear projection matrix *C*. All noise, ω_0, ω_i and ε_i (i = 1…T) are zero-mean normally distributed random variables with covariance matrices Q_0, Q and R, respectively. In the model, only the observation of the system is presented. The state and all the noise variables are hidden. Overall, the definition and mathematical description of symbols used in the system is shown in Table 1.

Table 1. Definition and mathematical description of Symbol table

Symbol	Definition and mathematical description
Y	A multi-dimensional observation sequences, $m \times T$
m	The dimension of the observation sequence
T	Time duration of sequences
W	Missing values indication matrix, $m \times T$
H	The dimension of hidden variables
μ_0	The Initial state for hidden variable, $H \times 1$
A	The Transition matrix, $H \times H$
C	Projection matrix from hidden state to observation, $m \times H$
Q	Transition covariance, $H \times H$
Q_0	Initial covariance, $H \times H$
R	Projection covariance, $m \times m$
Z	A sequence of hidden variables, $\{z_1, z_2, ..., z_T\}$
θ	A set including all necessary model parameters, $\theta = \{\mu_0, Q_0, A, Q, C, R\}$

3.2 Proposed Model Setup in the presence of missing values

The problem of missing time ticks will be first formulated in EEG data by the proposed model system. In the

time course experiment, consider a collection of M-dimensional EEG time series Y of a length T with lost measurements; the missing values of the observations are indicated by matrix W. The matrix W of missing observation is the same size as Y and is defined as below;

$$W(t,i)=\begin{cases}1 & if\ i-th\ dimensional\ observation\ of\ Y\ is\ observed\ at\ time\ t\\ 0 & if\ i-th\ dimensional\ observation\ of\ Y\ is\ missing\ at\ time\ t\end{cases} \quad (4)$$

The time sequences are modeled based on LDS, as seen in the above equations (1) – (3), with an extra missing indication matrix W [26]. We utilize an expectation-maximization algorithm to impute missing positions through estimating their expectation of missing values, $E\left[Y_{miss}|Y_{obs}\right]$, conditioned on the observed values, where Y_{miss} and Y_{obs} are the set of variables for the missing values and the set of the observed values in the sequence Y, respectively.

In order to handle the problem of missing values, the prime objective is to mine meaningful patterns via automatically identifying a few hidden variables so that their dynamics will be discovered to solve the problem of missing observations. This study focuses on exploiting the dynamic connectivity of the brain signals via two particular properties: namely, correlation and temporal continuity. To meet this problem, it needs to model the dynamics and hidden patterns of the observed time sequence by using sequences of hidden state variables Z. To model correlations, the model uses data sequences, which includes both observed and missing values, generated from a series of hidden variables via a $M \times T$ linear projection matrix C at each time point, shown in Fig. 1 where H is the number of hidden variables. Therefore, if some of the values are missing, they are inferred from the hidden variables since its mapping automatically discovers the correlations among the observation dimensions.

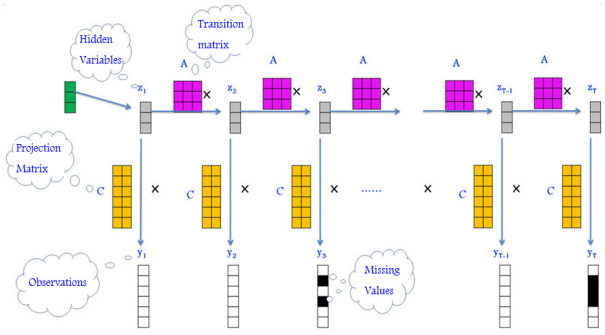


Fig. 1. Architecture of Proposed Model Setup

On the other hand, to model the temporal continuity property, *HxH* linear transition mapping A is adopted, since the hidden variables are time dependent with the values determined from the previous time ticks. This means that matrix A is related to the transition of the hidden variables' states over time, describing how the states move forward over time. Thus, the next time point only depends on the current time point. In this case, we first set an initial state for hidden variables at the begining time point with the set of parameters $\theta=\{\mu_0,Q_0,A,Q,C,R\}$. In the system, the joint distribution of Y_{obs}, Y_{miss}, and Z is given by the following equation:

$$P(Y_{miss},Y_{obs}\ and\ Z)=P(z_1)\cdot\prod_{i=2}^{T}P(z_i|z_{i-1})\cdot\prod_{i=1}^{T}P(y_i|z_i) \quad (5)$$

To achieve these goals above, the proposed model is given to find the optimal solution to maximize the expected log-likelihood of the observation sequence with respect to the model parameters $\theta=\{\mu_0,Q_0,A,Q,C,R\}$, the hidden variables $\hat{z}_n=E[z_n]$, $n=1...T$, and the missing observation $E\left[Y_{miss}|Y_{obs}\right]$.

In practicality, in order to achieve parameter estimation, it is necessary to find the maximum likelihood $L(\theta)=P(Y_{obs})$. Yet, it is known to be difficult to maximize the data likelihood in presence of missing values. Therefore, the expected log--likelihood of the observation sequence over the parameter θ is maximized by using Expectation-Maximization (EM) algorithm [27], which iteratively maximizes the expected complete log- -likelihood as in equation (6). To achieve the maximum likelihood estimation of the model parameters, the expectation–maximization (EM) method for learning LDS is utilized. The algorithm iterates between computing the conditional expectation of hidden variables through the forward-backward procedure in E-step and updating model parameters to maximize its likelihood in M-step for estimating missing values [25].

$$\begin{aligned}L(\theta;Y)=E_{Y,Z|\theta}\Big[&-(z_1-\mu_0)^T Q_0^{-1}(z_1-\mu_0)\\&-\sum_{t=2}^{T}(z_t-A\cdot z_{t-1})^T Q^{-1}(z_t-A\cdot z_{n-1})\\&-\sum_{t=1}^{T}(y_t-C\cdot z_t)^T R^{-1}(y_n-C.y_n)\Big]\end{aligned} \quad (6)$$

In summary, the proposed method is performed to achieve the best parameters $\theta=\{\mu_0,Q_0,A,Q,C,R\}$ for the model. The method applied in this paper conducts three main steps: expectation, recovering missing values, and maximization. In more detail, the algorithm first guesses an initial set of model parameters in the expectation step. It uses Kalman filtering and Kalman smoothing to estimate the hidden variables based on observation and current parameters for each iteration. The general idea is to use a forward-backward algorithm to compute the posterior expectations of hidden variables, $E(z_n|Y;\theta)$, tick by tick based on the computation of the previous time tick. Given the data with missing values, the estimation finds the marginal distribution for hidden state variables after initializing missing values to be a random number with interpolation method. Both prior and conditional distributions in the model are Gaussian, thus, the

posterior up to the current time tick is $p(z_n|y_1,...,y_T)$, which is also Gaussian given by:

$$\hat{\alpha}(z_n) = N(\mu_n, V_n) \tag{7}$$

We obtain the following forward-backward propagation equations. The values here are μ_n, V_n and P_{n-1}, given by:

$$P_{n-1} = A \cdot V_{n-1} \cdot A^T + Q \tag{8}$$

$$K_n = P_{n-1} \cdot C^T \cdot \left(C \cdot P_{n-1} \cdot C^T + R\right)^{-1} \tag{9}$$

$$V_n = \left(I - K_n\right) \cdot P_{n-1} \tag{10}$$

$$\mu_n = A \cdot \mu_{n-1} + K_n \cdot \left(y_n - C \cdot A \cdot \mu_{n-1}\right) \tag{11}$$

The initial values are given by following equations:

$$K_1 = Q_0 C^T \left(G Q_0 C^T + R\right)^{-1} \tag{12}$$

$$\mu_1 = \mu_0 + K_1(y_1 - C \cdot A \cdot \mu_0) \tag{13}$$

$$V_1 = \left(I - K_1\right) \cdot Q_0 \tag{14}$$

Smoothing involves an initial forward recursion followed by a backward recursion. In the forward step, the values of the Kalman Filter equations are stored. In the backward step, these values are then used to initialize the Kalman smoother equations given by:

Smoothing involves an initial forward recursion followed by a backward recursion. In the forward step, the values of the Kalman Filter equations are stored. In the backward step, these values are then used to initialize the Kalman smoother equations given by:

$$\hat{\mu}_n = \mu_n + J_n \cdot \left(\hat{\mu}_{n+1} - A \cdot \mu_n\right) \tag{15}$$

$$\hat{V}_n = V_n + J_n \cdot \left(\hat{V}_{n+1} - P_n\right) \cdot J_n^T \tag{16}$$

$$J_n = V_n \cdot A^T \cdot P_n^{-1} \tag{17}$$

The expectations are taken from the posterior marginal distribution $p(z_n|y_1,...,y_T)$ from the propagation of belief. Therefore, the expectations are obtained by the following equations:

$$E\left[z_n\right] = \hat{\mu}_n \tag{18}$$

$$E\left[z_n z_{n-1}^T\right] = J_{n-1}\hat{V}_n + \hat{\mu}_n \hat{\mu}_{n-1}^T \tag{19}$$

$$E\left[z_n z_n^T\right] = \hat{V}_n + \hat{\mu}_n \hat{\mu}_n^T \tag{20}$$

In the recovering step, missing values are recovered by using Markov property in the graphical model from the estimation of hidden variables, **(Fig. 1)** with the following equations;

$$E\left[Y_{miss} | Y_{obs}, Z; \theta\right] = C \cdot E\left[Z\right]_{\{i,j\}}, \ (\{i,j\} \in W \tag{21}$$

In the maximization step, the algorithm updates the parameter θ^{new} by maximizing the expected log-likelihood using some sufficient statistics from the posterior distribution. To estimate the parameters, the expected log-like hood $L(\theta;Y)$ in equation **(6)**, with respect to the components of θ^{new}, is maximized. Taking the derivatives of equation **(6)** and making them be zero provide the following results:

$$\mu_0^{new} = E\left[z_1\right] \tag{22}$$

$$Q_0^{new} = E\left[z_1 z_1^T\right] - E\left[z_1\right]E\left[z_1^T\right] \tag{23}$$

$$A^{new} = \left(\sum_{n=2}^{T} E\left[z_n z_{n-1}^T\right]\right)\left(\sum_{n-1}^{T-1} E\left[z_n z_n^T\right]^{-1}\right) \tag{24}$$

$$Q^{new} = \frac{1}{T-1}\sum_{n=2}^{T}\left(E\left[z_n z_n^T\right] - A^{new}E\left[z_{n-1} z_n^T\right] - E\left[z_n z_{n-1}^T\right]\left(A^{new}\right)^T + A^{new}E\left[z_n z_{n-1}^T\right]\left(A^{new}\right)^T\right) \tag{25}$$

$$C^{new} = \left(\sum_{n=1}^{N} y_n E\left[z_n^T\right]\right)\left(\sum_{n=1}^{T} E\left[z_n z_n^T\right]\right)^{-1} \tag{26}$$

$$R^{new} = \frac{1}{T}\sum_{n=1}^{T}\left(y_n y_n^T - C^{new}E\left[z_n\right]y_n^T - y_n E\left[z_n^T\right]\left(C^{new}\right)^T + C^{new}E\left[z_n z_n^T\right]\left(C^{new}\right)^T\right) \tag{27}$$

On the whole, the proposed method for solving the problem of missing values in EEG time series can be summarized as follows;

- Estimate hidden variables Z (E-step): Given the fixed parameters, θ and Y containing missing values, the forward-backward procedure to estimate posterior $P(Z|Y;\theta)$ and its sufficient statistics $E\left(z_n|Y;\theta\right)$, $E\left(z_n z_n^{'}|Y;\theta\right)$, $E\left(z_n z_{n+1}^{'}|Y;\theta\right)$ are used.
- Recovering missing values: Given fixed Z, missing values Y_{miss} $E\left(Y_{miss}|Z;\theta\right)$ using $E\left(z_n|Y;\theta\right)$ are estimated.
- Update model parameters (M-step): Given fixed Y and Z, new model parameters, $\theta^{new} \leftarrow \text{argmax}\, E\left[\log(Y,Z,\theta\right]$, are estimated.

4. EXPERIMENTAL EVALUATION

4.1 Dataset Acquisition

To demonstrate the effectiveness of the proposed method by considering accuracy, reliability, and the complexity aspects, we evaluate its performances in recovering

consecutive missing values on two real different datasets of EEG signals. We compare our proposed method to the interpolation and MSVD methods. The study is examined on random ranges of multiple coevolving EEG time series with consecutive missing values using a variety of parameter settings over different real datasets.

The first dataset is the publicly available Epilepsy EEG database at *http://epileptologiebon-n.de/cms/frontcontent.php?idcat=193&lang=3&changelang=3*. Further details are found in the work of [28]. This dataset contains no missing values consisting of five classes denoted by A, B, C, D, and E. Classes A and B are composed of segments taken from the surface EEG recordings that were carried out on five healthy volunteers. Classes C, D, and E originated from the EEG archive of presurgical diagnosis. Class D was recorded from within the epileptogenic zone. Class C was from the hippocampal formation of the opposite hemisphere of the brain. While classes C and D contained only the activity measured during seizure free intervals, class E contained seizure activity. Each of the class contained 100 single channels with 4,097 time points. In this study, we utilize the two sets named A and E from the complete dataset.

The second dataset is taken from *http://www.bbc-i.de/competition/ii/*. This dataset was recorded from a normal subject during a no-feedback session. The subject sat in a normal chair with relaxed arms resting on the table and fingers in the standard typing position at the computer keyboard. This BCI (self-paced) dataset contained two classes. Class label '0' is used for upcoming left hand movements and class labels '1' is used for upcoming right hand movements. Each class contained 28 EEG channels in the following order: F3, F1, Fz, F2, F4, FC5, FC3, FC1, FCz, FC2, FC4, FC6, C5, C3, C1, Cz, C2, C4, C6, CP5, CP3, CP1, CPz, CP2, CP4, CP6, O1, O2 [29].

4.2 Performance Evaluation on Reconstruction Error

To conduct our research experiment, two aspects are considered to assess the effective performance of the proposed method against MSVD and the interpolation approach. The comparisons are carried out based on the estimation performance on different amounts of entries with missing values.On the other way, we evaluate this problem based on a fixed percent of entries with missing values with different average missing lengths. For each experiment setup, we created different positions of consecutive missing observation on random channels of BCI and Epilepsy EEG datasets. The experiments are repeated 10 times in order to avoid the random effect.We reported the average of the mean square error (MSE) in order to evaluate the quality of the proposed method. The MSE is calculated following the equation: $\sum \|\tilde{y}_t - y_t\|^2 \Big/ \sum \|y_t\|^2$, where t denotes each time tick, $\tilde{y}_t$ is reconstructed data, and y mentions the input data.

For each dataset in each experiment, the Fukunaka's principle rule is used as a tool to achieve a proper number h for the hidden dimension of the model by taking the Singular Value Decomposition (SVD) of the original data $Y = U \cdot S \cdot V^T$, where both U and V are orthonormal matrices, S is a diagonal matrix with singular values on the diagonal. To get the number of h, small singular values are typically set to zero. Therefore, we order the singular values, and then choose h at the one with the 98th percentile of the total sum of squared singular values.

The comparison is based on the difference among the three methods' reconstruction errors at 5%, 10%, and 15% of missing values, as shown in Table 2. In all of the cases, both the proposed method and MSVD use the same number of hidden variables with 98% energy; the average length of the consecutive missing values is 35 time points. Table 2 demonstrates that in all different amounts of missing data over the range of 5%, 10%, and 15%, the reconstruction errors of the proposed method give the best results that have smaller errors than the interpolation and MSVD methods. Specifically, in the BCI (self-paced) dataset, the proposed method presents 0.00483 and 0.00733 average reconstruction errors, which are lower than those of the MSVD and interpolation, respectively. On this dataset, it shows approximately 77% and 84% improvement compared to that of MSVD and interpolation.

Similarly, in subject 2 of the BCI (self-paced) dataset, the performance of the proposed method also presents 75% and 85% improved reconstruction against that of MSVD and interpolation. In the epilepsy dataset, the perfromances obatin higher reconstruction errors compared to the BCI (self-paced) dataset since it is more complicated. However, the average MSE of missing imputaions also shows significant reconstruction with lower errors at 0.069 and 0.1495 over those of the MSVD and interpolation, and shows 49% and 67% improvement over that of the MSVD and interpolation on subject A. On epilepsy of subject E, Table 2 shows that the proposed method presents estimated values of 0.0702 and 0.1574, which performed 49% and 68% improvements in the average MSE compared to MSVD and interpolation.

Table 2. Reconstruction error over different rates of missing values 5%, 10% and 15%

Dataset	Class	Method	Reconstruction Error on different missing rates			Average MSE
			5%	10%	15%	
BCI (Self- paced	1	Proposed	**0.00062**	0.0014	0.0021	0.00137
		MSVD	0.0032	0.0067	0.0088	0.0062
		Interpolation	0.008	0.0088	0.0093	0.0087

	2	Proposed	**0.00059**	0.0013	0.002	0.0013
		MSVD	0.0027	0.0054	0.0075	0.0052
		Interpolation	0.0081	0.0087	0.009	0.0086
Epilepsy	A	Proposed	**0.0351**	0.0699	0.1107	0.0719
		MSVD	0.0726	0.144	0.2065	0.14103
		Interpolation	0.122	0.2307	0.3115	0.2214
	E	Proposed	**0.0314**	0.0701	0.1116	0.07103
		MSVD	0.0717	0.1434	0.2086	0.14123
		Interpolation	0.1443	0.2297	0.3113	0.22843

Secondly, the comparision is based on the difference average missing length settings with fixed 10% entries missing values among the three methods' reconstruction errors. Fig. 2 shows the efficiency of the proposed method based on a fixed 10% missing entries with different consecutive missing lengths, ranging from 10 to 100 consecutive missing observations for BCI (self-paced) dataset in Fig. 2(a), and from 10 to 80 consecutive missing values for Epilepsy dataset in Fig. 2(b), respectively. Again, the proposed method performs the best reconstruction among the three methods with MSE increasing slightly along with the increasing consecutive missing length.

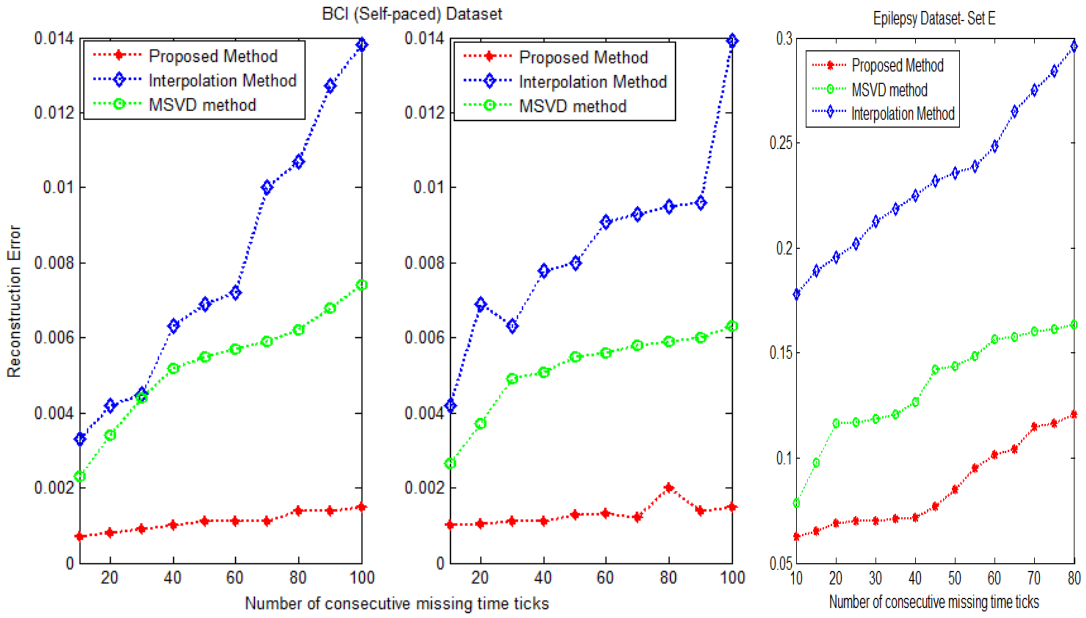


Fig. 2. Average reconstruction errors versus different average missing lengths on BCI and Epilepsy.

To test the efficiency of the proposed method, Fig. 3 shows an intuition of qualitative reconstruction on channel 4 from subject 1 of the BCI dataset of the three approaches. We demonstrate different hidden variables, which correspond to 96%, 97%, 98%, and 99% of energy in the original dataset. In all of the cases, the best recovery is performed on the 98% energy with 100 consecutive missing time points. The proposed method performs 74% and 85% higher accuracy for reconstruction than MSVD and interpolation on the BCI (self-paced) dataset, respectively. In the figure, the dashed portion indicates original signals, the green line is the reconstructed signals by the MSVD method, and the red line depicts the reconstructed signals of our proposed method. It shows that our proposed method (red line) achieves the best reconstruction since it reaches very close to the original signals against MSVD and interpolation methods.

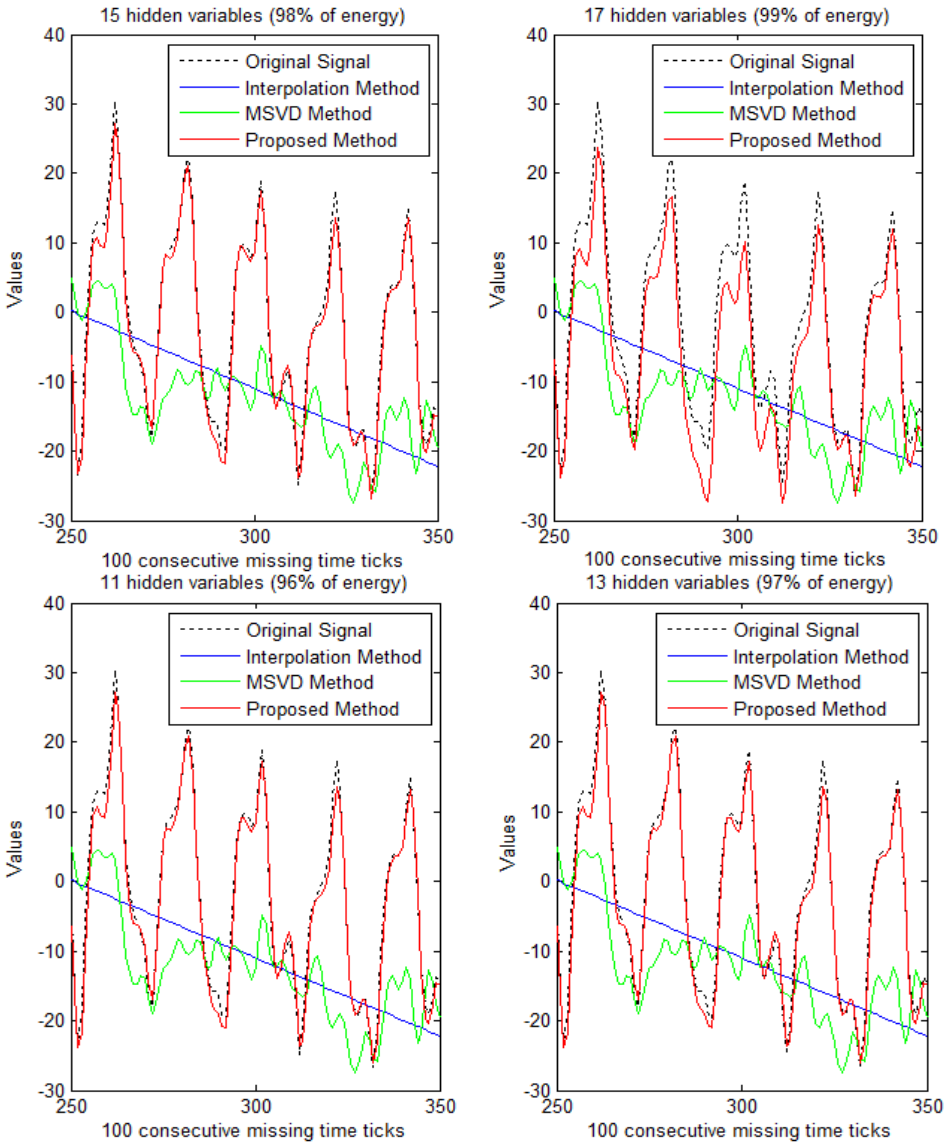
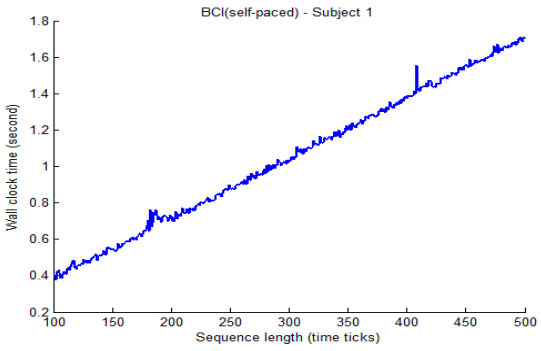


Fig. 3. Reconstructions vs. original signals

Since the proposed methodology captures correlations via discovering a few powerful patterns as well as learning dynamics from multiple coevolving EEG time sequences, our proposed method provides better results than the compared approaches.



4.3 Computation time performance

The computational complexity of the proposed method is shown in Fig. 4. It can be observed that wall clock times lie on almost a straight line over the duration of sequences. The computation time of the proposed method increases slowly with the input and the time duration T of the time sequences. In this experiment, we use 98% energy on all sets, and the learning step runs at the same number of iterations equal to 20; we also set 10% missing values of the original dataset for each run together with 30 average occlusion lengths.

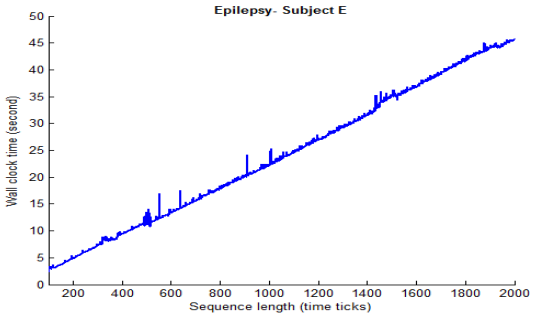


Fig. 4. Computation time versus the duration

Although the proposed method presents good results in all cases of the experiments and linear computation time, it still gets a hard time because the parameters of the proposed method are learned by the EM procedure. It takes more time for iterations to converge since the computation time of forward-backward procedures need cubic time in the dimension of observations. Therefore, the time complexity of the proposed method will receive poor scaling with large dimensional data. We will improve the speedup of the proposed model in future study.

5. CONCLUSION

In this paper, we proposed a method to tackle the problem of consecutive missing values for multiple EEG time series, specifically on their reconstruction. The method solves consecutive missing values of real multi-dimensional EEG time series. It can also automatically discover a few useful patterns and learn their dynamics successfully in order to solve the problem with consecutive missing observations. In most cases, this method provides the best result compared to the alternative techniques such as interpolation and MSVD. On the other hand, performance results show almost a linear speedup as we increase the input of the data set.

In this paper, we have only examined the EEG time series data for missing value reconstruction. For further work, the proposed method will be examined with other EEG time series mining tasks such as compressing and forecasting. Furthermore, we will broaden our research to involve different types of data such as Electrocardiography (ECG) and similar multiple coevolving data in the time series model. Finally, we will continue our exploration on the theme of mining large coevolving sequences with the goal of developing fast algorithms.

ACKNOWLEDGEMENTS

This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2013-H0301-13-3005) supervised by the NIPA (National IT Industry Promotion Agency). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013-052849).

REFERENCES

[1] N. F. Guler, E. D. Ubeyli, and I. Guler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," In Expert System Appl, vol. 29, 2005, pp. 506-514.

[2] K. Polat and S. Gunes, "Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform," In Applied Mathematics and Computation-AMC, vol. 187, no. 2, 2006, pp. 1017-1026.

[3] H. Adeli, Z. Zhou, and N. Dadmehr, "Analysis of EEG records in an epileptic patient using wavelet transform," In Journal of Neuroscience methods, vol. 123, 2003, pp. 69-87.

[4] N. Sivasankari and Dr. K. Thanushkodi, "Automated Epileptic Seizure Detection in EEG signals Using FastICA and Neural Network," In Int. J. Advances, Soft computer, Appl, vol. 1, no. 2, 2009, pp. 91-99.

[5] N. F Guler, E. D Ubeyli, and I. Guler, "Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficient," In Journal of Neuroscience methods, vol. 148, no. 2, 2005, pp. 113-121.

[6] N. Kannathal, M. L. Choo, U. R. Acharya, and P. K. Sadasivana, "Entropies for detection of epilepsy in EEG," In Computer Methods Programs Biomed, vol. 80, no. 3, 2005, pp. 187-194.

[7] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," In Expert Systems with Applications, vol.32, no. 4, 2007, pp. 1084-1093.

[8] E. Acar, D. M. Dunlavy, T. G Kolda, and M. Mørup, "Scalable Tensor Factorizations with Missing Data," In Proceedings of the 2010 SIAM International Conference on Data Mining, 2010, pp. 701-702.

[9] A. Sorjamaa, J. Hao, N. Reyhani, Y. Ji, and A. Lendasse, "Methodology for long-term prediction of time series," In Neurocomputing, vol. 70, 2007, pp. 16-18.

[10] C. Chatfield, The Analysis of Time Series: An Introduction, 2008.

[11] C. F. Gerald and P. O. Wheatley, Applied Numerical Analysis in 7th Edition, 2007.

[12] G. Liu and L. McMillan, "Estimation of missing markers in human motion capture," In International Journal of Computer Graphics, vol. 22, no. 9, 2009, pp. 721-728.

[13] S. I. Park and J. K. Hodgins, "Capturing and animating skin deformation in human motion," In ACM Trans. Graph, vol. 25, no. 3, 2006, pp. 881- 889.

[14] M. Kurucz, A. Benczúr, and K. Csalogany, "Method for large scales SVD with missing values," In KDDCup. 07, San Jose, California, USA, 2007.

[15] S. H. Kim, H. J. Yang, and Ng. KamSwee, "Incremental expectation maximization principal component analysis for missing value imputation for coevolving EEG data,"

Journal of Zhejiang University-SCIENCE C (Computer & Electronics), vol. 12, no. 8, 2011, pp. 687-697.
[16] S. Guha, D. Gunopulos, and N. Koudas, "Correlating synchronous and asynchronous data stream," In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD, 2003.
[17] S. Papadimitriou, J. Sun, and C. Faloutsos, "Streaming pattern discovery in multiple time series," In Proceeding of the 31st VLDB, 2005.
[18] N. S Nati and T. Jaakkola, "Weighted low-rank approximation," In 20th International Conference on Machine Learning, AAAI Press, 2003, pp. 720- 727.
[19] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," In Bioinformatics, vol. 7, 2001, pp. 520-525.
[20] C. Phong and R. Singh, "Missing value estimation for time series microarray data using linear dynamical systems modeling," In Advanced Information Networking and Applications-AINA, 2008, pp. 814-819.
[21] A. Aristidou, J. Cameron, and J. Lasenby, "Predicting missing marker to drive real- time center of rotation estimation," In AMDO '08: Proceeding of the 5th international conference on articulated motion and deformable objects, Springer-Verlarg, 2008, pp. 238-249.
[22] K. Dorfmuller-Ulhaas, *Robust Optical User Motion Tracking Using a Kalman Filter*, In Technique report, Technical Report of Augsburg University, 2003.
[23] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," In Journal of Basic Engineering, vol. 81, 1960, pp. 35-45.
[24] L. Ralaivola and F. d'Alche-Buc, "Time series Filtering, Smoothing and Learning using the Kernel Kalman Filter," In Proceedings of International Joint Conference on Neural Network, Montreal, Canada, 2005.
[25] L. Li, W. Fu, F. Guo, T. C Mowry, and C. Faloutsos, "Cut- And- Stitch: Efficient Parallel Learning of Linear Dynamical Systems on SMPs," In KDD'08, Las Vegas, Nevada, USA, 2008.
[26] L. Li, J. McaCann, N. Pollard, and C. Faloutsos, "Dynammo: Mining and summarization of coevolving sequences with missing values," In KDD, New York, NY, USA, 2009.
[27] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," In Journal of Time Series Analysis, vol. 3, 1982, pp. 253-264.
[28] R. G. Andrzejak, "Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity," in Physical Review E, 2001.
[29] B. Blankertz, G. Curio, and K. R. Müller, "Classifying Single Trial EEG: Towards Brain Computer Interfacing," In Advances in Neural Inf, 2002.

Ngoc Anh Nguyen Thi
She received the B.S, in Faculty Mathematics-Informatics from Da Nang Education University, Viet Nam in 2006, and M.S. at Dept. Electronics and Computer Engineering, Chonnam National University, Korea. She is currently a Ph.D. student at Dept. of Electronics and Computer Engineering, Chonnam National University, Korea. Her research interests focus on the intelligent computing in many applications such as pattern recognitions, bioinformatics, data analysis of data mining and machine learning.

Hyung-Jeong Yang
She received her B.S., M.S. and Ph. D from Chonbuk National University, Korea. She was a Post-doc researcher at Carnegie Mellon University, USA. She is currently an associate professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design.

Sun-Hee Kim
She received M.S in Dongguk University, Korea. She received Ph. D degree at Dept. Electronics and Computer Engineering, Chonnam National University, Korea. She is currently a Post-doc researcher at School of Computer Science, Carnegie Mellon University, USA. Her research interests focus on data mining, sensor mining and stream mining.