Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Incremental expectation maximization principal component analysis for missing value imputation for coevolving EEG data*

Sun Hee KIM, Hyung Jeong YANG†‡, Kam Swee NG

(*Department of Computer Science, Chonnam National University, Gwangju 500-757, Korea*)

†E-mail: hjyang@jnu.ac.kr

Received Oct. 14, 2010; Revision accepted Apr. 12, 2011; Crosschecked July 4, 2011

Abstract: Missing values occur in bio-signal processing for various reasons, including technical problems or biological characteristics. These missing values are then either simply excluded or substituted with estimated values for further processing. When the missing signal values are estimated for electroencephalography (EEG) signals, an example where electrical signals arrive quickly and successively, rapid processing of high-speed data is required for immediate decision making. In this study, we propose an incremental expectation maximization principal component analysis (iEMPCA) method that automatically estimates missing values from multivariable EEG time series data without requiring a whole and complete data set. The proposed method solves the problem of a biased model, which inevitably results from simply removing incomplete data rather than estimating them, and thus reduces the loss of information by incorporating missing values in real time. By using an incremental approach, the proposed method also minimizes memory usage and processing time of continuously arriving data. Experimental results show that the proposed method assigns more accurate missing values than previous methods.

Key words: Electroencephalography (EEG), Missing value imputation, Hidden pattern discovery, Expectation maximization, Principal component analysis

doi:10.1631/jzus.C10b0359 **Document code:** A **CLC number:** TP391

1 Introduction

Missing values in a set denote elements that are not observed in given data matrices. These arise due to technical problems or biological irregularities. One may remove incomplete records from the observed data set to handle missing values while they are imputed from other signals. Estimation methods of missing values are widely used and have been extensively developed since simply removing missing values may generate a biased model, causing loss of information. Imputation approaches, those that assign estimated values based on analysis and prediction, provide less loss of information from the original data than cases that exclude incomplete records with missing values (Horton and Lipsitz, 2001; Graham *et al*., 2007; Graham, 2009). Simple imputation methods generally use mean values of observed data. However, they underestimate standard deviation as well, since they do not consider the uncertainty in missing values.

Rubin (1987) proposed the multiple imputation (MI) method as an extension of a single imputation method, to consider the uncertainty. It replaces each missing value with a set of plausible values, instead of filling a single value. Yuan (2001) provided an imputation method that uses a combination of the regression method, the propensity score method, and Markov chain Monte Carlo (MCMC). However, these imputation methods mostly assume a normal data distribution of data and consider the posterior distribution of data in order to apply standard analysis,

‡ Corresponding author

* Project supported by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by National IT Industry Promotion Agency (No. NIPA-2011-C1090-1111-0008), the Special Research Program of Chonnam National University, 2009, and the LG Yonam Culture Foundation

such as general regression and analysis of variance (ANOVA).

In some cases, maximum-likelihood approaches generate good estimates from incomplete data via weighted estimation procedures, such as an expectation and maximization (EM) algorithm. These methods are more efficient because they do not require a model fitting test for posterior distribution of log-likelihood (Schafer, 1997; Little and Rubin, 2002). Adams *et al*. (2002) applied EMPCA by integrating principal component analysis (PCA) and the EM algorithm in pharmaceutical data for missing value imputation. Zhao *et al*. (2006) compared imputation performances between Robust EMPCA and a modified version of EMPCA using the data acquired from a waste water treatment process. Although Robust EMPCA is known to be superior to existing methods for estimating missing values, Zhao insisted Robust EMPCA performs well only in the case of non-negative data.

EEG signals contain the information for diagnosis with the form of a continuous graph (Smith, 2005). Many methods, including PCA, wavelet transform, adaptive auto regressive model, adaptive Gaussian representation, and independent component analysis, have been proposed to extract hidden information from EEG signals (Subha *et al*., 2010). However, these methods consider only complete EEG data. Data sets often include missing values during recoding of EEG signals for the following reasons: measuring equipment, movement, and communication disconnection. Previous methods also do not consider the characteristic of EEG signals that arrive consecutively from multiple electrodes, so real-time processing for EEG signals is essential.

In this paper, we propose an incremental EMPCA (iEMPCA) model. The proposed model incrementally updates the weights of PCA and the EM algorithm is used to estimate missing values using the maximum likelihood parameter in the model. The proposed method can perform real-time processing using an incremental model and summarizes large data by detecting hidden variables. It solves the problem of a biased model, which is inherited from missing data, and reduces the loss of information by imputing missing values. It decreases the computation complexity of the learning process by discovering major patterns. We measure similarities between original data and imputed data in EEG signals to confirm efficiencies of the proposed method, and compare it to existing methods.

2 Related works

Studies of missing value imputation that address clinical data often appear in fields of health science, biology and medical science (Musil *et al*., 2002). Among the methods for handling the missing values we commonly find simple imputation, multiple imputation, and maximum-likelihood estimation (MLE). Simple imputation methods generally use *K*-nearest neighbors and mean values of observed data. Dixon (1979) compared several simple imputation methods such as mean imputation, least squares estimation, and classical *K*-nearest neighbors. Abdala and Saeed (2004) used a weighted *K*-nearest neighbor algorithm to estimate the values of missing clinical laboratory data such as blood gasses, blood chemistry, and blood counts. Recently, Norazian *et al*. (2008) used mean and interpolation imputation to estimate missing values from air pollution data. However, because they did not use full information of observed values, these methods have drawbacks in that they may generate biased distribution of variance.

Rubin (1978) proposed a multiple imputation method that imputes a missing value with more than one estimated value to compensate for this drawback. Rosenbaum and Rubin (1983) proposed the propensity score methods using discriminant analysis and logistic regression. Posterior distributions of Bayesian inference were applied in MCMC to express unknown parameters (Schafer, 1997; Yuan, 2001). This method iterates two steps, imputation and posterior. The imputation step estimates the missing values using the mean vector and covariance matrix. The posterior step computes posterior probability distribution of the unknown parameters. Raghunathan *et al*. (2001) proposed multivariate imputation using a regression sequence for missing values. This approach obtains the imputations by drawing values from the truncated predictive distributions and fitting a sequential regression on incomplete data sets generated from hypothetical populations. Ni *et al*. (2005) estimated the performance of the various multiple imputation methods in traffic data monitored by video cameras. Janssen *et al*. (2009) evaluated multiple imputation methods using deep vein thrombosis with

missing values to predict future event occurrence or the presence of a disease. However, these imputation methods require input parameters, which make them less accessible for general cases.

MLE is a well-known statistical estimation method that finds the parameter to maximize the likelihood function. Dempster *et al*. (1977) used an approach to compute maximum likelihood in various levels such as truncated and censored data when observed data are the incomplete data. Schneider (2001) estimated incomplete climate data and simulated surface temperature data using regularized expectation-maximization (EM) to compute the MLE. They both used an iterative method to estimate missing values using covariance matrices and mean values. Smith *et al*. (2003) estimated the missing data using EM and data augmentation (DA) in the traffic data collected in real time. However, compared with other methods, MLE approaches require relatively large-scale samples to generate precise estimations again, negatively affecting the accessibility of this approach for general cases.

Another approach was taken by Sharma *et al*. (2004) who focused on factor models to update pseudo-missing values of six permanent traffic counts (PTCs) based on neural networks, regression models, and an autoregressive integrated moving average (ARIMA) model. Al-Deek *et al*. (2004) and Zhong *et al*. (2005) estimated missing values with basic linear analysis techniques. Wang *et al*. (2006) proposed an imputing approach based on the support vector regression (SVR) method. This technique utilizes an orthogonality coding input scheme in a gene expression profile. Yamaguchi *et al*. (2008) demonstrated their model on clinical data of bladder cancer patients with missing values by applying self-organizing maps (SOM). They compared the model's discriminative ability by the receiver operating characteristic (ROC) area. Recently, Ryan *et al*. (2010) evaluated four alternative imputation strategies—one global (PCA based) and three local (nearest neighbor based). Ching *et al*. (2010) proposed a weighted local least square imputation (WLLSI) approach for missing value estimation to a breast cancer dataset. However, these methods require an entire data set to estimate missing values, and they were not designed to precisely handle the time series in cases of missing values.

In this study, we propose an incremental EM PCA that can impute the missing values of time series data in real time. The proposed method shows the short processing time of missing value imputation. iEMPCA simultaneously addresses such perennially challenging issues as missing value imputation, limited memory, and processing time in time series data.

3 Materials and methods

In this section, we discuss how the idea of EMPCA is developed in an incremental way. The purpose of iEMPCA is to recover missing values incrementally. Missing values can be inferred from other observed values based on correlations between different sequences.

3.1 Expectation maximization principal component analysis

Traditional PCA is widely known for both dimensionality reduction and feature extraction. PCA extracts eigen components, indicating data characteristics. It reduces the dimensionality by taking the characteristic components, which are representative data among the components. In short, PCA seeks an axis of the principal component that expresses the data sufficiently well (Smith, 2002). The dimensionality is reduced by projection to data matrix $\boldsymbol{X}$ with the axis of the principal components. The axis of the principal components is computed with the eigenvector and eigenvalue. The principal components can be computed using singular value decomposition (SVD) as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{T}\boldsymbol{V}^{\mathrm{T}},$$

where $\boldsymbol{X}_{n\times m}$ is the data matrix with n features and m observations, $\boldsymbol{T}_{n\times f}$ is a matrix containing the loading vectors, $\boldsymbol{V}_{m\times f}$ is a matrix containing the score vectors, f specifies the number of principal components, and m indicates the time point in time series data (Stanimirova *et al*., 2007).

However, if data matrix $\boldsymbol{X}$ contains the missing values, a traditional PCA approach is no longer coherent. Al-Deek and Roweis proposed an improved PCA approach based on expectation-maximization (termed EMPCA) to resolve the missing value problem (Roweis, 1998; Al-Deek *et al*., 2004). EMPCA is

computationally efficient in time and space by extracting a few eigenvectors and eigenvalues from large-scale collections of high-dimensional data. Also, it does not require computing the sample covariance of the data.

The EM-based PCA algorithm seeks the values of missing data using two steps. E-step includes filling a missing value with a base at the expectation values of data. The initial value begins with the mean of the row and the column. M-step maximizes a value obtained at E-step, and the missing values are refilled with the predicted values by PCA in the iterative process of EM.

$$\text{E-step:}\quad \boldsymbol{P}=(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C})^{-1}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{X}, \tag{1}$$

$$\text{M-step:}\quad \boldsymbol{C}=\boldsymbol{X}\boldsymbol{P}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}})^{-1}, \tag{2}$$

where $\boldsymbol{P}$ is a $k\times m$ matrix and $\boldsymbol{C}$ is an $n\times k$ matrix. Therefore, $\boldsymbol{X}$ is modeled as the combination of $\boldsymbol{C}$ and $\boldsymbol{P}$ with noise, $\boldsymbol{X}=\boldsymbol{C}\boldsymbol{P}+\boldsymbol{N}$, where $\boldsymbol{N}$ is noise (Zhao *et al*., 2006).

PCA is applied on the completed data that are filled by the mean of the row and column. The data are then reconstructed to be used as alternates of missing values until convergence. The EMPCA algorithm can be explained as follows: First, missing values are initialized by means of row and column of the data matrix $\boldsymbol{X}$. PCA is then performed on the completed data set. E-step and M-step are repeatedly taken using Eqs. (1) and (2).

The data matrix $\boldsymbol{X}$ is reconstructed with the number of predefined principal components using Eqs. (3) and (4):

$$\boldsymbol{Y}=\boldsymbol{C}_{\text{new}}^{\mathrm{T}}\boldsymbol{X},\ \text{// Project } \boldsymbol{X} \text{ onto loading vector } \boldsymbol{C}_{\text{new}} \tag{3}$$

$$\hat{\boldsymbol{X}}=\boldsymbol{C}_{\text{new}}\boldsymbol{Y}.\ \text{// Reconstruction} \tag{4}$$

Next, the missing values are replaced with their reconstructed values as Eq. (5). The observed values of $\boldsymbol{X}$ remain unchanged.

$$\boldsymbol{X}^{r}=\hat{\boldsymbol{X}},\ \text{// Replace the missing values} \tag{5}$$

where r is the number of iterations.

$$\text{Error}=\text{MSE}(\boldsymbol{X},\boldsymbol{X}^{r}).\ \text{// Estimate the error rate} \tag{6}$$

The last three steps are repeated until convergence by Eq. (6). The data matrix $\boldsymbol{X}$ that contains the missing values is replaced with estimated values generated by the EMPCA algorithm.

3.2 iEMPCA for missing values

EEG time series are composed by the data matrix of the variable n and observed time tick t. For example, data set $\boldsymbol{X}$ is an $n\times t$ matrix. PCA reduces the data $\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ in n dimensions to $\boldsymbol{Y}=(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_k)$ in k dimensions, where $k<n$. The maximum information of data $\boldsymbol{X}$ is kept and principal components are not correlated to each other. The general form of the linear combination of the principal components is denoted as $\boldsymbol{Y}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}$.

The feature vector of the reduced dimensionality is expressed as $\boldsymbol{Y}=\boldsymbol{W}_k^{\mathrm{T}}\boldsymbol{X}$. $\boldsymbol{W}$ consists of $n\times k$ basis vectors. The basis vector is generally a standard basis vector. However, in the case of EEG time series data, a standard basis vector is far from the ideal way to express the data, since the data are measured by a time tick. Therefore, a new method that processes data according to time ticks is required for real-time analysis of time series data. Papadimitriou *et al*. (2005) proposed incremental PCA (iPCA) to resolve the problem of real-time analysis. It processes time series data by automatically updating the basis vector in each time tick. Table 1 shows the notations.

Table 1 Notations used in this paper

Symbol	Description
$\boldsymbol{W}$	Weight matrix
$\boldsymbol{w}_i$	ith participation weight vector
$\boldsymbol{x}_t$	n input values at time t
$\boldsymbol{X}$	$\boldsymbol{X}=(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$
$\boldsymbol{x}_{\text{replace}}$	replaced values at time t
$\boldsymbol{x}_{t,\text{miss}}$	n input values containing missing values at time t
$\boldsymbol{x}_{\text{mean}}$	Mean of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{t-1}$
$\hat{\boldsymbol{x}}_t$	Reconstruction of $\boldsymbol{x}_t$
$\boldsymbol{Y}$	Hidden variable matrix
$\boldsymbol{y}_t$	Vector of hidden variables for $\boldsymbol{x}_t$
λ	Exponential forgetting factor
$E_{i,t}$	Total energy of ith hidden variables up to time t
E_{hv}	Old total hidden variable energy
$E_{\boldsymbol{x}}$	Old total input data energy

In this study, we employ iPCA, which finds a new weight vector, $\boldsymbol{w}_i$, while incrementally updating the basis vector. The number of hidden variables is adjusted by updating the weight vector, which minimizes the average reconstruction error at each time tick (Papadimitriou *et al*., 2005). When we compute the ith hidden variable, $y_{i,\tau}$ ($1\le\tau\le t$) is expressed as

$$y_{i,\tau}=\sum_{n=1}^{N}(w^{\mathrm{T}})_{n,i}\cdot x_{n,\tau}. \tag{7}$$

We evaluate the magnitude of the ith hidden variable to incrementally determine the number of hidden variables by

$$E_{i,t}=\frac{1}{t}\sum_{\tau=1}^{t}y_{i,\tau}^2. \tag{8}$$

The threshold of magnitude follows the standard value from a minimum of 95% to a maximum of 98% as defined in Pan *et al*. (2004) and Sun *et al*. (2005). We re-express the magnitude threshold by energy. The energy threshold corresponds to a bound that contains the values of the upper and lower bounds of the energy. Its function is to determine how many hidden variables should be retained. The number of hidden variables will be increased if the maintained energy of hidden variables at the particular timestamp is less than the lower bound of the energy. Conversely, the number of hidden variables will be decreased if the estimated energy is too high. We may lose some of useful information if we set the lower bound energy too low, since fewer hidden variables will be obtained. Therefore, in our experiment, we keep track of the energy level of the hidden variables to obtain an efficient result, even though the first few hidden variables are sufficient to summarize the entire EEG signals.

The reconstruction of $\boldsymbol{x}_t$ is as shown in Eq. (9):

$$\hat{\boldsymbol{x}}_t=\sum_{i=1}^{k}\boldsymbol{w}_i\cdot\boldsymbol{y}_t, \tag{9}$$

where $\hat{\boldsymbol{x}}_t$ denotes the reconstruction after projection on the k-dimensional space. The reconstruction error rate in Eq. (10) is used to update the basis vector:

$$e=\left\|\hat{\boldsymbol{x}}_t-\boldsymbol{x}_t\right\|. \tag{10}$$

If the new sample $\boldsymbol{x}_{t+1}$ contains missing values, the proposed method imputes the missing values using the incremental EMPCA approach. The model analyzes data as a unit of time tick in which the number of hidden variables is changed incrementally by updating the basis vector, as in iEMPCA (Algorithm 1). This approach uses the exponential forgetting factor λ to reflect more recent trends in EEG data. In this study, we assume λ=0.96; the typical choice of λ is between 0.96 and 0.98 (Pan *et al*., 2004; Sun *et al*., 2005). A large buffer space is not required with this incremental EMPCA approach, which helps reduce the amount of memory required.

Algorithm 1 iEMPCA algorithm

Input:
New input $\boldsymbol{x}_t\in\mathbb{R}^n$
Predefined lower bound energy f
Predefined upper bound energy F
Exponential forgetting factor λ
Output:
Estimated missing values $\boldsymbol{x}_{\mathrm{replace}}$
Algorithm
if input vector $\boldsymbol{x}$ includes missing values $\boldsymbol{x}_{t,\mathrm{miss}}$
$\boldsymbol{x}_{\mathrm{replace}}=\boldsymbol{x}_{t,\mathrm{miss}}=\boldsymbol{x}_{\mathrm{mean}}$ // Initialize the missing values
// by $\boldsymbol{x}_{\mathrm{mean}}$ while the observed values of $\boldsymbol{x}_t$ remain
// unchanged
Error=MSE($\boldsymbol{x}_t$, $\boldsymbol{x}_{\mathrm{replace}}$) // Estimate the error rate
while no convergence
E-step: $\boldsymbol{P}=(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{replace}}$
M-step: $\boldsymbol{W}_{\mathrm{new}}=\boldsymbol{x}_{\mathrm{replace}}\boldsymbol{P}^{\mathrm{T}}(\boldsymbol{P}\boldsymbol{P}^{\mathrm{T}})^{-1}$
Replace the missing values using Eqs. (3)–(6)
end while
$\boldsymbol{y}=\boldsymbol{W}_{\mathrm{new}}^{\mathrm{T}}\boldsymbol{x}_{\mathrm{replace}}$ // Compute hidden variables
$E_{\mathrm{hv}}=\lambda E_{\mathrm{hv}}+\boldsymbol{y}^2$
// Compute the energy of total hidden variables
$E_{\boldsymbol{x}}=\lambda E_{\boldsymbol{x}}+\boldsymbol{x}_{\mathrm{replace}}^2$
// Compute energy of total input data
if $E_{\mathrm{hv}}<fE_{\boldsymbol{x}}$
$k=k+1$
// Increase the number of hidden variables
else if $E_{\mathrm{hv}}>FE_{\boldsymbol{x}}$
$k=k-1$
// Decrease the number of hidden variables
end if
else // input vector $\boldsymbol{x}_t$ does not include missing values
// at time t
for i=1 to k // k is the number of hidden variables
Update the weight vector using Eqs. (7)–(10)
end for
// Adjust the number of hidden variables, k
if $E_{\mathrm{hv}}<fE_{\boldsymbol{x}}$
$k=k+1$
else if $E_{\mathrm{hv}}>FE_{\boldsymbol{x}}$
$k=k-1$
end if

Given multiple EEG time series data with missing values, the proposed iEMPCA method that estimates these missing values is applied. It extracts patterns which describe data characteristics that have changed over time. In addition, the proposed method detects the correlation amongst the data, by discovering hidden variables within the EEG data. iEMPCA does not have a large memory requirement; therefore, it is possible to determine the data characteristics and accurately estimate missing EEG time series values in real time.

4 Experimental results

In this section, we illustrate the EEG signals data acquisition process. Experimental results are presented in several aspects to show the effectiveness and efficiency of the proposed method.

4.1 Data sets

In our iEMPCA tests, we used three sets of multivariate EEG time series data (Table 2). The first EEG data set (publicly available at http://www.epileptologie-bonn.de/cms/) reflects a spasmodic epilepsy as provided by Krug. It was measured using 100 channels for 23.6 s. These data were recorded by a continuous multi-channel EEG after visual confirmation of the activity of the muscle or movement of eyes. Environments of the five subjects differ. Subjects 1 and 2 are eye open and closed conditions. Subject 3 is from the hippocampal formation of the opposite hemisphere of the brain, and subject 4 is from the epileptogenic zone. Both subjects 3 and 4 were recorded in seizure-free intervals. Subject 5 contains seizure activity. Subjects 1 and 2 were recorded extracranially and subjects 3, 4, and 5 were recorded intracranially. In this study, we illustrate experiment results from three subjects (1, 3, 5) to exclude similar results.

The second data set is Event Related Potential (ERP) provided on http://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html. It was recorded using 31 channels from participants sitting in a dimly lit room with a screen. The subjects performed two tasks—an animal categorization task and a recognition task. An 8-bit color vertical photograph was flashed for 20 ms using a programmable graphic board. This short presentation time avoids subjects using exploratory eye movement to respond. Participants gave their responses by pressing a ‘go/no go’ button. In the animal categorization task, participants had to respond whenever there was an animal in the picture. Both tasks were organized in a series of 100 images, in which 50 target images and 50 non-target images were mixed.

The self-regulation EEG data set is provided by Klaus in http://www.bbci.de/competition/ii/. The experiment was conducted to determine the position of a target in a computer screen. While recording EEG signals, a subject moved the cursor by thought from left to right or from top to bottom to hit the target on the screen. Self-regulation EEG data were collected using 64 channels from three subjects. The subject’s goal was to move the cursor to the height of the target. When the cursor reached the right edge, the screen went blank. In this work, we used 1–6 sessions of each subject. All EEG data were given as complete data sets without missing values. To measure the accuracy of the missing value imputation, we should know

Table 2 Data description of the three sets of multivariate EEG time series data

Data set	Subject	Session	Number of sensors	Number of time points	Trial/Task
Epilepsy	1, 3, 5		100	4097	1 task
ERP	1	1–4	31	234360	100*4 trials
	2	1–4	31	206440	100*4 trials
	3	1–4	31	227240	100*4 trials
	4	1–4	31	206640	100*4 trials
	5	1–4	31	215680	100*4 trials
	6	1–4	31	207320	100*4 trials
Self-regulation	1	1–6	64	172992	192*6 trials
	2	1–6	64	172992	192*6 trials
	3	1–6	64	174720	192*6 trials

the real values of missing values. However, it is very difficult to grasp the basic truth of missing values from real data. Therefore, we generated 5%, 10%, and 15% of missing values in the original observed data to verify the efficiency of missing value imputation as previous works conducted the missing value imputation (Schlogl and Supp, 2006; Acar *et al*., 2011).

4.2 Accuracy of imputation of missing values

We applied iEMPCA on EEG data sets containing missing values and then measured the accuracy of the resulting estimated values. The proposed method uses energy values from 0.95 to 0.98 to automatically extract the number of hidden variables that summarize the data set. We compared the accuracy to that of Robust EMPCA (Zhao *et al*., 2006) and missing value SVD (MSVD) (Troyanskaya *et al*., 2001) (Tables 3–5). The accuracy of the missing value imputation was estimated by the root mean square error (RMSE). In Epilepsy EEG data, the proposed method shows 0.0087 and 0.0034 lower imputation RMSE, on average, than Robust EMPCA and MSVD, respectively. This represents 20% and 9% improvement over Robust EMPCA and MSVD, respectively.

Table 3 RMSE of missing value imputation on epilepsy data

Subject	Rate of missing values (%)	Number of hidden variables			Error rate		
		Proposed	Robust EMPCA	MSVD	Proposed	Robust EMPCA	MSVD
1	5	22	22	80	0.0289	0.0482	0.0334
	10	21	22	80	0.0337	0.0529	0.0397
	15	21	22	80	0.0362	0.0545	0.0427
	Average	21	22	80	0.0329	0.0519	0.0386
3	5	14	22	52	0.0335	0.0388	0.0360
	10	14	22	52	0.0374	0.0437	0.0424
	15	13	22	52	0.0400	0.0453	0.0450
	Average	14	22	52	0.0370	0.0426	0.0411
5	5	25	22	55	0.0239	0.0256	0.0243
	10	25	22	55	0.0370	0.0387	0.0376
	15	22	22	55	0.0387	0.0404	0.0392
	Average	24	22	55	0.0332	0.0349	0.0337
Total average		20	22	62	0.0344	0.0431	0.0378

Table 4 RMSE of missing value imputation on ERP data

Subject	Number of hidden variables			Error rate		
	Proposed	Robust EMPCA	MSVD	Proposed	Robust EMPCA	MSVD
1	5	4	4	0.0287	0.0361	0.0298
2	3	3	3	0.0050	0.0134	0.0099
3	4	4	4	0.0032	0.0033	0.0038
4	5	6	6	0.0283	0.0360	0.0317
5	5	4	4	0.0131	0.0178	0.0164
6	4	3	3	0.0197	0.0252	0.0175
Total average	4	4	4	0.0163	0.0220	0.0182

Table 5 RMSE of missing value imputation on self-regulation data

Subject	Number of hidden variables			Error rate		
	Proposed	Robust EMPCA	MSVD	Proposed	Robust EMPCA	MSVD
1	15	7	6	0.0232	0.0222	0.0187
2	20	4	4	0.0308	0.0303	0.0279
3	11	5	5	0.0399	0.0377	0.0320
Total average	15	5	5	0.0313	0.0300	0.0262

The RMSE of MSVD is 12% lower than that of Robust EMPCA. MSVD uses 52 to 80 hidden variables for Epilepsy EEG data when it contains 5% of missing values, while the proposed method uses only 14 to 25 hidden variables (Table 3).

When applying iEMPCA in ERP EEG data, the RMSE of missing value imputation is 0.0055 and 0.0018 lower, on average, than Robust EMPCA and MSVD, respectively. It is 26% and 10% improvement over Robust EMPCA and MSVD, respectively. MSVD exhibits 0.0038 lower RMSE on average than Robust EMPCA (Table 4). The RMSE of MSVD is 17% lower than that of Robust EMPCA. However, in the case of self-regulation EEG data, the error rate of iEMPCA is higher than that of Robust EMPCA and MSVD. MSVD and Robust EMPCA show 16% and 4% higher accuracy, respectively, than the proposed method shown in Table 5. This is probably due to the fact that self-regulation EEG data are linear (Fig. 1), while the proposed method exhibits lower RMSE than MSVD and Robust EMPCA in the case of non-linear data such as Epilepsy EEG and ERP EEG data sets. MSVD shows the lowest RMSE among three methods for self-regulation EEG data.

4.3 Hidden variable detection

iEMPCA detects hidden variables that summarize large EEG time series data sets. In this work, we compare iEMPCA with MSVD with 5%, 10%, and 15% of missing values in each EEG data set by hidden variable detection. iEMPCA shows similar patterns of hidden variables with recovered values compared with MSVD. The top graph in Fig. 2a is the first hidden variable detected by iPCA on the complete data set. The middle graph in Fig. 2a is the first hidden variable detected by iEMPCA after 5% of missing values were recovered. It shows similar patterns to the top of Fig. 2a. However, the pattern of the bottom graph of Fig. 2a, which shows the first hidden variable by MSVD after 5% of missing values were recovered by MSVD, differs from that of the top graph. MSVD is known to be the best existing method to estimate missing values for linear data (Troyanskaya *et al*., 2001). However, as seen from the above results, MSVD has fewer precise patterns close to the values of the original data compared to the iEMPCA method. Graphs in Fig. 2b show the first hidden variable of the first session data set of subject 1 on the ERP EEG data.

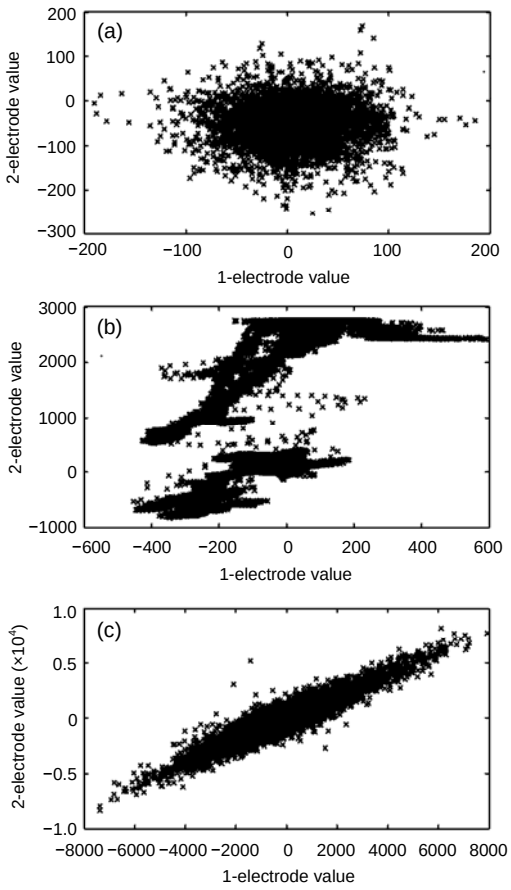

Fig. 1 Data distribution

Non-linear data: epilepsy EEG data (a) and ERP EEG data (b); linear data: self-regulation EEG data (c)

When comparing the pattern of the hidden variable detected by the three methods, iPCA and the proposed method can definitely divide each other in four different tasks. However, MSVD fails to detect the task division. Fig. 2c shows the first hidden variable of the first session data set of subject 1 on self-regulation EEG data. In the case of Fig. 2c, all three methods show similar patterns.

4.4 CPU time and computation complexity

The cost of iEMPCA is $O(r \cdot k)$, where r is the number of iterations and k is the number of hidden variables. In contrast, the complexity of MSVD is $O(r \cdot m \cdot n)$, where m is the number of samples and n is the dimensionality of $\boldsymbol{X}$. Therefore, the proposed method requires less memory than MSVD. In this experiment, 5% of each data set is treated as missing values. We compare the proposed method to MSVD and iPCA. iPCA uses the original data without

missing values, because it expects only complete data and does not take missing values into account. When applied to data with missing values, iPCA cannot obtain a correct experimental result.

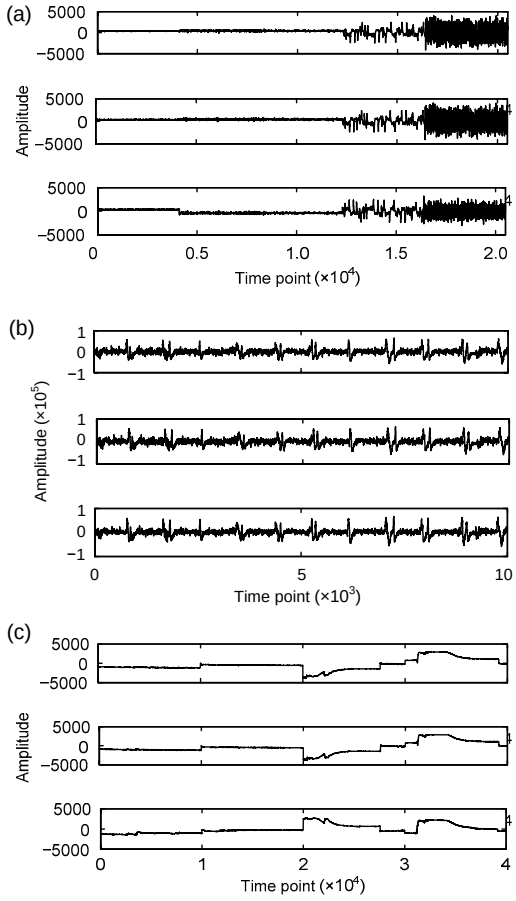

Fig. 2 Comparison of the detected first hidden variable
(a) Epilepsy EEG data; (b) First session data set on ERP EEG data; (c) First session on self-regulation EEG data. In each subfigure, from top to bottom: iPCA (original data), the proposed method (missing rate: 5%), MSVD (missing rate: 5%)

Fig. 3 shows the learning time of Epilepsy EEG, ERP EEG, and self-regulation EEG data. MSVD exhibits the longest execution time, since this approach must use all the data from data 1 to data t to learn if the new data t+1 enter the system. However, iPCA and our proposed method do not need as much learning time, since given new data t+1 they use weights and initial values at time point t. Therefore, iEMPCA can solve the problem of long processing time and limited memory by estimating missing values in real time.

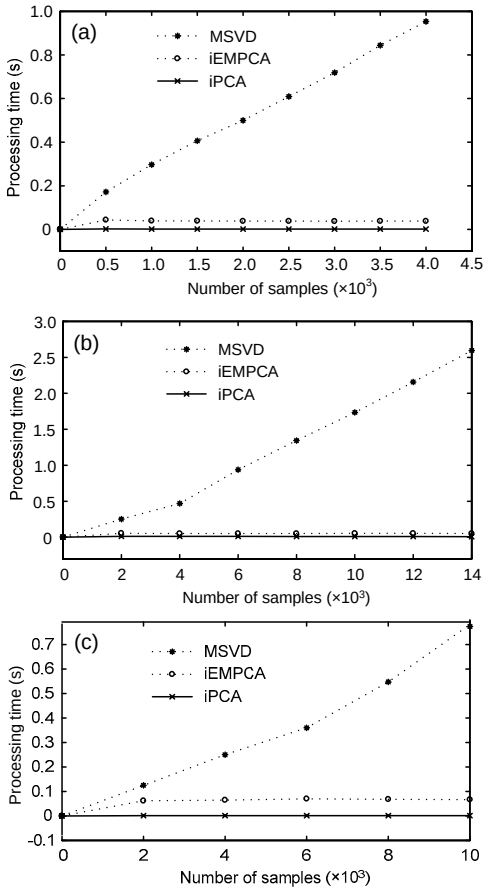

Fig. 3 Learning time measurement for epilepsy (a), ERP (b), and self-regulation (c) EEG data

5 Conclusions

We propose iEMPCA for multiple time sequences that contain missing values. iEMPCA automatically estimates missing values, and summarizes by finding of hidden variables. It estimates the approximation values of missing values and distinguishes the specific pattern using discovered hidden variables. The reduction of data based on the hidden variables can be used as learning data. The proposed method reduces the memory requirement, since it can discover the hidden variables after automatically estimating missing values in real time. iEMPCA is more accurate than other methods for the imputation of missing values in the case of non-linear data. Therefore, approximation for missing values in large quantities of EEG time series data can be estimated close to the original data. Multivariate temporal EEG

can be summarized using a few hidden variables through our method. Moreover, the proposed method can reduce the processing complexity and memory requirement.

References

Abdala, O.T., Saeed, M., 2004. Estimation of missing values in clinical laboratory measurements of ICU patients using a weighted *K*-nearest neighbors algorithm. *Comput. Cardiol.*, **31**:693-696. [doi:10.1109/CIC.2004.1443033]

Acar, E., Dunlavy, D.M., Kolda, T.G., Mørup, M., 2011. Scalable tensor factorizations for incomplete data. *Chemometr. Intell. Lab. Syst.*, **106**(1):41-56. [doi:10.1016/j.chemolab.2010.08.004]

Adams, E., Walczak, B., Vervaet, C., Risha, P.G., Massart, D.L., 2002. Principal component analysis of dissolution data with missing elements. *Int. J. Pharm.*, **234**(1-2):169-178. [doi:10.1016/S0378-5173(01)00966-8]

Al-Deek, H.M., Venkata, C., Chandra, S.R., 2004. New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. *Trans. Res. Rec. J. Transp. Res. Board*, **1867**:116-126. [doi:10.3141/1867-14]

Ching, W.K., Li, L., Tsing, N.K., Tai, C.W., Ng, T.W., Wong, A.S., Cheng, K.W., 2010. A weighted local least squares imputation method for missing value estimation in microarray gene expression data. *Int. J. Data. Min. Bioinform.*, **4**(3):331-347. [doi:10.1504/IJDMB.2010.033524]

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B*, **39**(1):1-38.

Dixon, J.K., 1979. Pattern recognition with partly missing data. *IEEE. Tran. Syst. Man. Cybern.*, **9**(10):617-621. [doi:10.1109/TSMC.1979.4310090]

Graham, J.W., 2009. Missing data analysis: making it work in the real world. *Ann. Rev. Psychol.*, **60**(1):549-576. [doi:10.1146/annurev.psych.58.110405.085530]

Graham, J.W., Olchowski, A.E., Gilreath, T.D., 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.*, **8**(3):206-213. [doi:10.1007/s11121-007-0070-9]

Horton, N.J., Lipsitz, S.R., 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *Am. Stat.*, **55**(3):244-254. [doi:10.1198/000313001317098266]

Janssen, K.J.M., Vergouwe, Y., Donders, A.R.T., Harrell, F.E.Jr., Chen, O., Grobbee, D.E., Moons, K.G.M., 2009. Dealing with missing predictor values when applying clinical prediction models. *Clin. Chem.*, **55**(5):994-1001. [doi:10.1373/clinchem.2008.115345]

Little, R.J.A., Rubin, D.B., 2002. Statistical Analysis with Missing Data (2nd Ed.). John Wiley and Sons, New York, p.200-222.

Musil, C.M., Warnerm, C.B., Yobas, P.K., Jones, S.L., 2002. A comparison of imputation techniques for handling missing data. *West. J. Nurs. Res.*, **24**(7):815-829. [doi:10.1177/019394502762477004]

Ni, D., Leonard, J.D., Guin, A., Feng, C., 2005. Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. *J. Transp. Eng.*, **131**(12):931-938. [doi:10.1061/(ASCE)0733-947X(2005)131:12(931)]

Norazian, M.N., Shukri, Y.A., Azam, R.N., Al Bakri, A.M.M., 2008. Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, **34**(3):341-345. [doi:10.2306/scienceasia1513-1874.2008.34.341]

Pan, J.Y., Kitagawa, H., Hamamoto, M., Faloutsos, C., 2004. AutoSplit: Fast and Scalable Discovery of Hidden Variables in Stream and Multimedia Databases. 8th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, p.519-528. [doi:10.1007/978-3-540-24775-3_62]

Papadimitriou, S., Sun, J., Faloutsos, C., 2005. Streaming Pattern Discovery in Multiple Time-Series. 31st Int. Conf. on Very Large Data Bases, p.697-708.

Raghunathan, T.E., Lepkowksi, J.M., van Hoewyk, J., Solenbeger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**(1):85-95.

Rosenbaum, P.R., Rubin, D.B., 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B*, **45**(2):212-218.

Roweis, S., 1998. EM algorithms for PCA and SPCA. *Adv. Neur. Inform. Process. Syst.*, **10**:626-632.

Rubin, D.B., 1978. Multiple Imputation in Sample Surveys—a Phenomenological Bayesian Approach to Nonresponse. Proc. Survey Research Methods Section, p.20-34.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York, p.249-250.

Ryan, C., Greene, D., Cagney, G., Cunningham, P., 2010. Missing value imputation for epistatic MAPs. *BMC Bioinform.*, **11**(1):197-234. [doi:10.1186/1471-2105-11-197]

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, London, p.478-479.

Schlogl, A., Supp, G., 2006. Analyzing event-related EEG data with multivariate autoregressive parameters. *Progr. Brain Res.*, **159**:135-147. [doi:10.1016/S0079-6123(06)59009-0]

Schneider, T., 2001. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**:853-871. [doi:10.1175/1520-0442(2001)014<0853:AOICDE>2.0.CO;2]

Sharma, S., Lingras, P., Zhong, M., 2004. Effect of missing values estimations of traffic parameters. *Transp. Plan. Technol.*, **27**(2):119-144. [doi:10.1080/0308106042000218203]

Smith, B.L., Scherer, W.T., Conklin, J.H., 2003. Exploring imputation techniques for missing data in transportation management systems. *Transp. Res. Rec. J. Transp. Res. Board*, **1836**:132-142. [doi:10.3141/1836-17]

Smith, L., 2002. A Tutorial on Principal Components Analysis. Cornell University, USA. Available from http://www.cs.otgo.ac.nz/cosc453/student_tutorials/principal_compone

nts.pdf [Accessed on Sept. 10, 2009].

Smith, S.J.M., 2005. EEG in the diagnosis, classification, and management of patients with epilepsy. *J. Neurol. Neurosurg. Psych.*, **76**:ii2-ii7. [doi:10.1136/jnnp.2005.069245]

Stanimirova, I., Daszykowski, M., Walczak, B., 2007. Dealing with missing values and outliers in principal component analysis. *Talanta*, **72**(1):172-178. [doi:10.1016/j.talanta.2006.10.011]

Subha, D.P., Joseph, P.K., Acharya, U.R., Lim, C.M., 2010. EEG signal analysis: a survey. *J. Med. Syst.*, **34**(2):195-212. [doi:10.1007/s10916-008-9231-z]

Sun, J., Papadimitriou, S., Faloutsos, C., 2005. Online Latent Variable Detection in Sensor Networks. 21st Int. Conf. on Data Engineering, p.1126-1127. [doi:10.1109/ICDE.2005.100]

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, B., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**(6):520-525. [doi:10.1093/bioinformatics/17.6.520]

Wang, X., Li, A., Jiang, Z., Feng, H., 2006. Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinform.*, **7**:32. [doi:10.1186/1471-2105-7-32]

Yamaguchi, T., Mackin, K.J., Matsumoto, K., Okusa, H., 2008. SOM for classifying data sets with missing values: application to clinical data of bladder cancer patients. *Artif. Life Robot.*, **13**(1):271-274. [doi:10.1007/s10015-008-0578-5]

Yuan, Y.C., 2001. Multiple Imputation for Missing Data: Concepts and New Development SAS/STAT 8.2. Available from http://www.sas.com/statistics [Accessed on May 18, 2010].

Zhao, L., Chai, T., Cong, Q., 2006. Operating Condition Recognition of Predenitrification Bioprocess Using Robust EMPCA and FCM. Sixth World Congress on Intelligent Control and Automation, p.9386-9390. [doi:10.1109/WCICA.2006.1713818]

Zhong, M., Sharma, S., Liu, Z., 2005. Assessing robustness of imputation models based on data from different Jurisdictions: examples of Alberta and Saskatchewan, Canada. *Transp. Res. Rec. J. Transp. Res. Board*, **1917**:116-126. [doi:10.3141/1917-14]

2010 JCR of Thomson Reuters for *JZUS-A* and *JZUS-B*

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME | HELP | RETURN TO LIST

2010 JCR Science Edition

Journal: Journal of Zhejiang University-SCIENCE A

Mark	Journal Title	ISSN	Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Citable Items	Cited Half-life	Citing Half-life
☐	J ZHEJIANG UNIV-SC A	1673-565X	442	0.322		0.050	120	3.7	7.1

Journal: Journal of Zhejiang University-SCIENCE B

Mark	Journal Title	ISSN	Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Citable Items	Cited Half-life	Citing Half-life
☐	J ZHEJIANG UNIV-SC B	1673-1581	770	1.027		0.137	124	3.5	7.5

JZUS-A is an international “Applied Physics & Engineering” reviewed-Journal, covering research in Applied Physics, Mechanical and Civil Engineering, Environmental Science and Energy, Materials Science, and Chemical Engineering. *JZUS-B* is an international “Biomedicine & Biotechnology” reviewed-Journal, covering research in Biomedicine, Biochemistry, and Biotechnology. *JZUS-A* and *JZUS-B* were covered by SCI-E in 2007 and 2008, respectively.