# Model Based Sparse Feature Extraction for Biomedical Signal Classification

Shengkun Xie[1,*] and Sridhar Krishnan[2]

[1]Ted Rogers School of Management, Ryerson University, Toronto, ON M5B 2K3, Canada

[2]Department of Electrical and Computer Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada

**Abstract:** This article focuses on model based sparse feature extraction of biomedical signals for classification problems, which stems from sparse representation in modern signal processing. In the presented work, a novel approach based on sparse principal component analysis (SPCA) is proposed to extract signal features. This method involves partitioning signals and utilizing SPCA to select only a limited number of signal segments in order to construct signal principal components during the training stage. For signal classification purposes, a set of regression models based on sparse principal components of the selected training signal segments is constructed. Within this approach, model residuals are estimated and used as signal features for classification. The applications of the proposed approach are demonstrated by using both the synthetic data and real EEG signals. The high classification accuracy results suggest that the proposed methods may be useful for automatic event detection using long-term observational signals. *keywords*: Sparse Principal Component Analysis, Sparse Feature Extraction, Signal Classification, Long-term Signals

**Keywords:** Sparse Principal Component Analysis, Sparse Representation, Signal Classification, Long-term Signals.

## 1. INTRODUCTION

In many medical diagnoses of human diseases including sleep disorder and epilepsy, long-term observational biomedical signals are often used for detecting related health events. In processing long-term biomedical signals, signal segmentation techniques are often applied to obtain a set of signal segments which are more stationary than the case that is without doing it. This enables application using some modeling techniques such as autoregressive moving average models. When applying segmentation to biomedical signals, the existence of non-stationarity and multi-scale structure may lead to low classification accuracy because signal segments from the same class may have different characteristics. This implies that model estimates using those signal segments will be highly volatile and it may be difficult to train a model based on the results of parameter estimates.

In signal classification, one may use as many signals as possible for the training process. However, this may not be a good idea for complex signals. Signal segments from long-term signals are often cross correlated, and cannot be treated as independent samples. This is particularly true for multichannel signals as they are sampled for the same investigation purpose. On one hand, the possibly high cross-correlation among signal segments restricts the use of many sophisticated statistical techniques for analyzing this type of signals; On the other hand, due to the huge amount of signal segments and their natural complexity, it is of great importance to efficiently manage and use these signals in computer aided diagnostic systems. This may call for novel techniques that are able to help reduce the dimensionality of data and to achieve reasonably good classification results.

Feature extraction of signals is a typical approach for data dimension reduction. The objective of feature extraction is to obtain a set of signal features so that a suitable classification method can be applied. The main reason for feature extraction of signals is that much of the sampled information does not necessarily contribute to high accuracy of classification results. Removal of redundant information will lead to a higher efficiency in signal analysis. In order to obtain high discriminative signal features, sparse representation of signals by compressed sensing, principal component analysis, independent component analysis, wavelet decomposition, empirical modes decomposition or matching pursuit has been proposed for various types of biomedical signals and images [1-3]. The aim of sparse representation is to use only a limited number of underlying signal components to represent a given signal [4-6]. In signal processing, these underlying signals are called basis functions and usually constructed by a time-frequency decomposition approach such as wavelet decompositions [7]. However, the traditional approach for sparse approximation of signals is based on decomposing a single signal at each time. Thus, it does not take inter-

*Address correspondence to this author at the Ted Rogers School of Management, Ryerson University, Toronto, ON M5B 2K3, Canada;
E-mail: shengkun.xie@ryerson.ca

E-ISSN: 1929-6029/17 

signal similarity into account, resulting in a large number of basis functions needed for complex signals. On the other hand, in signal classification, not all training signal segments are necessary. Often selected signal segments are contributable to classification, therefore it is important to identify those signal segments. Because of these reasons, developing simultaneous sparse representation of multiple signals becomes critical in real-world applications.

Due to the data dimension reduction property and its clustering effect, principal component analysis (PCA) [8] becomes a natural choice for clustering signal segments. Often first few principal components are able to explain major data variation, and their variances behave differently. Focusing on only the principal components can often lead to better clustering effect. PCA is done by searching signals that have larger correlations when each principal component is calculated. However, the limitation of using PCA is that it uses all signal segments when computing principal components. Due to signal redundancy, often only a subset of signal segments from the whole training data set is needed. Because of this, we propose sparse PCA (SPCA) to simultaneously select signal segments and construct signal principal components. We then use the extracted signal principal components to construct regression models in order for us to extract a set of features for classification. To achieve a better classification performance, we also propose a classification scheme based on the model residuals.

This paper extends our prior research on feature extraction via SPCA (e.g., [9, 10]. The essential difference is that the presented work proposes a systematic classification scheme based on SPCA and regression model residuals, and discuss on a general framework of sparse representation for signal classification. This work also examines the proposed method in a more detailed way by comparing the results to our prior work of using wavelet functional linear model [11]. This work is also related to our prior work of using dynamic PCA with non-overlapping moving window technique [12]. The main idea is quite similar and they all involve in signal segmentation and focus on extracting sparse signal features, with objective of achieving sparsity and reasonably good classification results. However the presented work uses SPCA to select a limited number of signal for constructing principal component, the prior work used PCA to extract signal principal components and combine PCs with signal energy measure to be feature vector for classification.

The remainder of the paper is organized as follows. In Section 2, the methods that are often used for sparse approximation of signals are discussed. Section 3 presents our classification scheme based on sparse variable approximation. In Section 4 we discuss the experimental results of using synthetic data and real EEG data. In Section 5 we report our concluding remarks.

## 2. SPARSE REPRESENTATION FOR SIGNAL CLASSIFICATION

Suppose one deals with a $G$-class classification problem, where the class is labeled by $g$, for $g = 1,\ldots,G$. In the training data set, it is assumed that there are $l_g$ signals in the $g$th class. Each signal is then partitioned into a set of signal segments, denoted by $\{X_{ig}\}_{i=1,\ldots,p_g, g=1,\ldots,G}$, where $X_{ig} \in \mathbb{R}^n$. Here $n$ is the length of the signal segments, $p_g$ is the number of signal segments in the class $g$. The total number of the signal segments for the training process is $p = \sum_{g=1}^{G} l_g p_g$. The objective of the classification is to determine the class membership of a test signal $Y$.

Sparse representation for signal classification (SRSC) [13] is an important technique to transform the signal $\{X_{ig}\}$ into a set of discriminative signal features. The objective is trying to identify a set of low dimensional discriminative signal features that facilitate the use of a simple classification method to achieve a reasonably good classification result. From this perspective, SRSC is quite different from sparse methods used for signal analysis such as signal de-noising or signal compression (e.g., [14]), where the focus is the signal re-construction with sparse components. In this section, we review some existing sparse methods that are often used in signal classification. These methods are subject to their own limitations when applied to signal classification.

### 2.1. Simple Sparse Approximation

Let $\{\phi_w \in \mathbb{C}^n : w \in \mathbb{N}\}$ be a collection of basis functions in the $n$-dimensional complex inner-product space. Sparse approximation is focused on identifying a sequence of basis function indices, $\lambda_1,\ldots,\lambda_M \in \mathbb{N}$ (often $M \ll d$) and estimating their coefficients, denoted by $c_m^{(ig)}$, for each $n \times 1$ signal $X_{ig}$ so that $X_{ig}$ can be represented by the following

$$X_{ig} = \sum_{m=1}^{M} c_m^{(ig)} \phi_{\lambda_m^{(ig)}} + e^{(ig)}. \quad (1)$$

where $\phi_{\lambda_m^{(ig)}}$ are selected basis functions from a given dictionary and $e^{(ig)}$ represents the noise component that is unexplained by sparse approximation. Since this signal approximation is applied to each signal segment independently, it is often called simple sparse approximation. In finding the sparse solution of (1), a greedy algorithm such as matching pursuit [7] or its variant (e.g., [15]) is often applied. As an alternative approach, variable selection via $L_1$ regularization can also be applied [16]. The most popular approach to this problem is to transform the original sparse approximation problem to a $L_1$ regularized optimization procedure [17]. This leads to the following optimization problem

$$\arg\min_{C_{ig}\in\mathbb{R}} \| X_{ig} - \Phi C_{ig} \|^2 + \lambda \| C_{ig} \|_1, \quad (2)$$

where $\Phi$ is the pre-defined $n \times n$ dictionary matrix. $\| X_{ig} - \Phi C_{ig} \|^2$ is the 2-norm that calculates the signal approximation error. $\lambda$ is the penalty parameter that controls the sparsity. The sparsity is often specified by the total number of non-zero elements in $C_{ig}$ that is represented by the 1-norm $\| C_{ig} \|_1$ in the equation (2) . The solution is obtained either through a greedy algorithm or via convex linear programming [18, 19].

In signal classification, the feature extraction problem can be formulated as sparse approximation of signals. When this is the case, a test signal with the same length as $X_{ig}$ is mapped onto $\phi_{\lambda_m^{(ig)}}$, to obtain signal features for classification. Note that achieving signal sparsity via a simple sparse representation of signal is done by controlling the signal reconstruction error. The exacted feature vector $C_{ig}$ is not necessarily discriminative. This is because the simple sparse approximation decomposes a signal at each time and it does not take inter-signal similarity into consideration.

### 2.2. Sparse Discriminative Feature Extraction

In order to take inter-signal similarity into account, the simultaneous sparse approximation (SSA) [20] was proposed. It looks for the sparse solution by considering all signal segments in all classes simultaneously. The signal segments are organized into a $n \times p$ data matrix, denoted by $X \equiv [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$, where $x_k = [X_{11}(k), \ldots, X_{l_1 1}(k), X_{(l_1+1)1}(k), \ldots, X_{l_1 \times p_1 1}(k), \ldots, X_{l_G \times p_G G}(k)]^{\mathrm{T}}$ is a $p \times 1$ column vector, for $k = 1, \ldots, n$, and $p = \sum_{g=1}^{G} l_g p_g$ is the total number of signal segments within the training set. The SSA problem can be formulated as follows

$$\arg\min_{C_{ij}\in^{n}} \| X - \Phi C \|^2 + \lambda \sum_{g=1}^{G} \sum_{i=1}^{l_g} \sum_{j=1}^{p_g} \| C_{ij} \|_1, \quad (3)$$

where $\lambda$ is a penalty parameter and $C = [C_{11}, \ldots, C_{l_1 1}, C_{(l_1+1)1}, \ldots, C_{l_1 \times p_1 1}, \ldots, C_{l_G \times p_G G}]$ is the vector of extracted features, where $C_{ij}$ is the feature vector of $X_{ij}$. Again the $\| C_{ij} \|_1$ is the 1-norm of $C_{ij}$ that calculates the total number of non-zero elements in the vector. In signal classification, those basis functions that are associated with non-zero components of $C_{ij}$ are identified and a given test signal segment is mapped onto those basis functions to obtain a set of features for determining class membership. The advantage of this approach is to analyze signals simultaneously, and it potentially leads to automatic selection of training signals so that only limited amount of signal segments will be used for obtaining signal features. The limitation of this type of sparse approximation is that the classification performance may be influenced by the choice of a dictionary as it is pre-defined.

In order to improve the effectiveness and the discrimination power of extracted features obtained from (3), one may further reduce the dimension of feature vector $C_{ij}$ through a $L_2$ regularization. This method selects the most discriminative signal features. It replaces the reconstruction error in the objective function of the optimization problem (3) by Fisher's discrimination power, which is defined as

$$D(C) = \frac{\| \sum_{g=1}^{G} p_g (m_g - m)(m_g - m)^{\mathrm{T}} \|^2}{\sum_{g=1}^{G} s_g^2}, \quad (4)$$

where

$$m_g = \frac{1}{l_g p_g} \sum_{i=1}^{l_g} \sum_{j=1}^{p_g} C_{ij}, s_g^2 = \frac{1}{l_g p_g} \sum_{i=1}^{l_g} \sum_{j=1}^{p_g} \| C_{ij} - m \|^2, \quad (5)$$

and

$$m = \frac{1}{p} \sum_{g}^{G} \sum_{i=1}^{l_g} \sum_{j=1}^{p_g} C_{ij}, \quad (6)$$

are respectively, the sample mean vector for class $g$, the sample variance vector for class $g$ and the grand mean vector for all classes. The $L_2$ regularization problem by using the Fisher's discrimination becomes

$$\arg\max_{C_{ij}\in R^n} D(C) + \lambda \sum_{g=1}^{G} \sum_{i=1}^{l_g} \sum_{j=1}^{p_g} \| C_{ij} \|_1, \quad (7)$$

where $\lambda$ is again a penalty parameter taking negative value. This approach does not use the pre-defined dictionary and it produces a set of empirical basis functions directly from the data. For feature extraction of a given test signal segment, it will map onto those extracted basis functions from (4).

### 2.3. Sparse Representation Based on Transform System

The signal sparse representation discussed above involves in finding a small set of basis functions to approximate the signal matrix $X$. In fact, the sparse approximation of signals can also be done by using a pre-defined sparse dictionary. This approach replaces a continuous basis function by a sparse vector. For instance, in sparse approximation of given signal $y(t)$ by discrete wavelet transform (DWT), which can be defined as $y(t)=\sum_{jk} y_{jk}\psi_{j,k}(t)$, where $j$ and $k$ are integers [21], the basis function $\psi_{j,k}(t)$ is generated by shrinking by a factor $2^{-j}$ and translating by $2^j k$ from the mother wavelet $\psi$, that is, $\psi_{j,k}(t)=2^{-j/2}\psi\left(2^{-j}t-k\right)$, where the $j$ subscript represents the dilation number and the $k$ subscript represents the translation number. In sparse representation of signals, the sparseness constraints are applied to wavelet coefficients and only those components that correspond to significantly large value are selected through wavelet shrinkage approaches, e.g. hard thresholding or soft thresholding [22], [23]. This approach is possible as signal energy is mainly preserved at a small number of wavelet coefficients. However, similar to the simple sparse approximation, it does not utilize inter-signal similarity when extracting signal features.

## 3. SIGNAL CLASSIFICATION BASED ON SPARSE VARIABLE APPROXIMATION

In the discussion above we assumed that a set of basis functions are given so that a search algorithm can be applied to identify the most relevant basis functions to represent the signals. In this case, the successful application of sparse techniques heavily depends on the appropriate choice of dictionary. An alternative to this is to obtain a set of basis functions directly from **X**. To achieve this, PCA via singular value decomposition (SVD) of **X** can be used. However, in PCA, the loading matrix of **X** (i.e., a matrix consisting of eigenvectors of **X**) is typically non-sparse, so that the underlying principal components (i.e., extracted feature vectors) often have low discrimination power [24]. To improve the effectiveness of PCA method in signal classification, sparse PCA (SPCA) was proposed. This method was originally proposed in [25] for the purpose of multivariate data dimension reduction. In this paper, we discuss SPCA from signal classification perspective, which is different from data dimension reduction point of view. In our approach, we treat each signal segment as a set of realization of a random variable, therefore the data matrix $X$ is modeled as a set of realizations of a $p$-variate random vector. Our objective is to approximate this random vector by a sparse one. That is, using SPCA we aim to select partial variables from the $p$-variate random vector to construct a dictionary matrix. The selection of relevant variables reduce amount of signal segments from the training set to be used. This method is particularly important in long-term signal classification problem. Since not all information collected is contributable to signal classification, it is necessary to remove the redundant information from the training set if there is any.

In the PCA approach, sparsity is often achieved by retaining only a limited number (i.e., pre-defined) of principal components. Each principal component is iteratively obtained from maximizing the data variation explained for each principal component. Suppose that we consider the first $M$ major PCs and let $A=[\alpha_1,\alpha_2,\ldots,\alpha_M]$ and $B=[\beta_1,\beta_2,\ldots,\beta_M]$ be the score matrix and loading matrix, respectively, where $\alpha_i$ and $\beta_i$ are $p$-dimensional column vectors. Finding the solution of sparse loading matrix $B$ leads to the following optimization problem

$$\arg\min_{A,B}\sum_{i=1}^{n}\| x_i - AB^{\mathrm{T}}x_i\|^2+\lambda_2\sum_{m=1}^{M}\|\beta_m\|^2+\sum_{m=1}^{M}\lambda_{1,m}\|\beta_m\|_1, \quad (8)$$

subject to $A^{\mathrm{T}}A=I_{M\times M}$, where $\lambda_2$ and $\lambda_{1,m}$ are regularization parameters. Here, $I_{M\times M}$ is an identity matrix. The orthogonality of $A$ ensures an optimal design matrix for signal approximation, which is desired from signal representation perspective. The first term in (8) controls the signal reconstruction error, the second term aims to regularize PCA (also called ridge regression based PCA in the statistical literature) and the third term is responsible for constructing sparse principal component loadings. As usual $\|\beta_m\|_1$ is the 1-norm of $\beta_m$ that compute the total number of non-zero elements of vector $\beta_m$. If $p>>n$, $\lambda_2$ can be set to be infinity and the right hand side of Equation (8) can be simplified to the following

$$\arg\min\sum_{i=1}^{n}\| x_i - AB^{\mathrm{T}}x_i\|^2+\sum_{m=1}^{M}\lambda_{1,m}\|\beta_m\|_1. \quad (9)$$

In finding the optimal solution of $A$ and $B$, the value of $\lambda_{1,m}$ is automatically selected through LASSO regularization paths [26], therefore, there is no need to specify the value of $\lambda_{1,m}$. Due to the orthogonality of $A$,

the solution of $\operatorname{argmin}\sum_{i=1}^{n}\| x_i - AB^{\mathrm{T}}x_i \|^2$ is equivalent to the solution of $\operatorname{argmin}\sum_{m=1}^{M}\| X\alpha_m - X\beta_m \|^2$. Therefore, for a given $A$, each $\beta_m$ in $B$ that minimizes

$$\sum_{m=1}^{M}\left\{\| X\alpha_m - X\beta_m \|^2 + \lambda_2 \| \beta_m \|^2 + \lambda_{1,m} \| \beta_m \|_1\right\} \quad (10)$$

will also minimize (8). Note that, for a given $A$, the solution that minimizes (10) is the elastic net estimate of $B$ [26]. Thus, finding the solution of (8) can be achieved via an iterative process of finding the elastic net estimate of $B$. This iterative process updates the initial value of $A = UV^{\mathrm{T}}$ by computing the SVD of $X^{\mathrm{T}}XB = UDV^{\mathrm{T}}$.

Due to dimension reduction of principal components and regularization of PCA via $L_1$ penalty, SPCA (i.e., through the $L_2$ penalty) provides us a set of sparse latent components, i.e. sparse principal components, along with a set of sparse loading vectors. The sparsity of the loading matrix results in a limited number of signal segments that account for most of the inter-signal variation. Therefore, the sparse principal component loading vector $\beta_m$ may consist of many zero elements, which suggests that the signal segments with zero coefficients do not contribute to the construction of the $m$ th principal component. The small value of $M$ leads to sparse approximation of signal by a low dimensional feature vector in the classification step.

In order to further explain our proposed method, the $m$ th principal component loading vector $\beta_m$ is re-written as $\beta_m = [\beta_m^{(1)^{\mathrm{T}}}, \beta_m^{(2)^{\mathrm{T}}}, \ldots, \beta_m^{(G)^{\mathrm{T}}}]^{\mathrm{T}}$, where $\beta_m^{(g)} = [\beta_{1,m}^{(g)},\ldots,\beta_{p_g,m}^{(g)}]^{\mathrm{T}}$, for $1 \le g \le G$ and $1 \le m \le M$. Here $\beta_m^{(g)}$ is a sparse column vector that consists of the $m$ th principal component loadings of all signal segments in class $g$, and $\beta_{i,m}^{(g)}$ is also a row vector that corresponds to $m$ th principal component loading vector for the $i$ th signal segment of class $g$. We then construct multiple linear regression models for a given signal segment by using these $M$ principal components as regressors. Within this approach, each given new signal segment $Y_i \in \mathbb{R}^n$ is first normalized (subtract its sample mean and then divided by its sample standard deviation) and then regressed, respectively, by a set of extracted PCs that are belonged to the same class $g$. The regression models are constructed as follows

$$Y_i = \sum_{m=1}^{M} c_m^{(i,g)} X^{(g)}\beta_m^{(g)} + e_i^{(g)}, \quad (11)$$

for $g = 1,2,\ldots,G$ and $i = 1,\ldots,l_g$, where $e_i^{(g)}$ is the $g$ th model residual vector for signal segment $Y_i$.

The PCs are obtained by calculating $X^{(g)}\beta_m^{(g)}$, where $X^{(g)}$ is a $n \times l_g p_g$ sub-matrix, consisting of only the $g$ th class signal segments. Note that we do not use the full sparse loading vector $\beta_m$, but we use the partial sparse loading vector $\beta_m^{(g)}$ for the regression. This is because that the $g$ th regression model represents regression of a test signal segment by the $g$ th class signal segments only. It is equivalent to retaining only the PC loadings that correspond to the $g$ th class signal segments and shrinking the loadings of the other classes to be zero, for the $g$ th regression model. Since $\beta_m^{(g)}$ is a sparse vector, only the selected signal segments within the $g$ th class are used to construct signal principal components, therefore this approach filters out many signal segments that do not contribute to classification.

### 3.1. Regression Model Residuals Based Classification

The conventional approach of using linear regression model for data classification often focuses on classifying the model coefficient vector $C^{(i,g)} = [c_1^{(i,g)}, c_2^{(i,g)}, \ldots, c_M^{(i,g)}]$. Each $c_m^{(i,g)}$ is calculated based on the regression model discussed above. This type of approach requires further feature selection on $C^{(i,g)}$ as well as an effective classifier in order to predict the class membership of signal $Y$. For the data we consider, this approach has been shown to be less effective for signal classification when a simple classifier is applied. This is often due to the fact that the signal segments are cross-correlated so that the extracted features are highly overlapped.

In residual analysis of regression, the characteristics of model residuals are important for model checking and model validation when the objective is modeling. For model based feature extraction problem, our goal is to discriminate signals. We still focus on the goodness of fit for the model and try to minimize the standard error for each model. After such minimization, we then consider the model residuals as useful features. In fact, the discriminative power of the model residuals becomes more important to us. Often it is the case that model residuals behave differently when different models are fitted to a test signal segment. Because of this, classification can be done using regression model residuals $e_i^{(g)}$, which are estimated by the mean absolute distance measure $\frac{1}{n}\| Y_i - \sum_{m=1}^{M} c_m^{(i,g)} X^{(g)}\beta_m^{(g)} \|_1$, a robust estimator for $e_i^{(g)}$. From the signal segmentation procedure, we know that

each signal $Y$ with the observational time $T$ is segmented into a total $\lfloor T/n \rfloor$ segments each of the length $n$. Using the model (11), we estimate the model residual $e_i^{(g)}$ for each $i$ th segment $Y_i$ of the signal $Y$ and compute the average of $e_i^{(g)}$, which is denoted by

$$R_Y^g = \frac{1}{\lfloor T/n \rfloor} \sum_{i=1}^{\lfloor T/n \rfloor} e_i^g. \quad (12)$$

The class membership of $Y$ is then determined by the one nearest neighbor classification method that uses the model residual $R_Y^g$ and the estimated model residuals obtained from the training data. Because the estimate of model residuals makes use of the labeling information of signals, it is then expected that the extracted features likely behave as clusters. This flowchart of our proposed classification scheme is represented in Figure **1**, and it is summarized as follows

**Classification scheme: input**: $X_{training} \equiv [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$ and a segmented signal, $Y_1, Y_2, \ldots, Y_{\lfloor T/n \rfloor}$, obtained from a test signal $Y_{test}$;

**initialize** $A$ and $B$; $A$ and $B$ are not converged,

**find** elastic estimate of $B$ by (10);

**compute** the SVD of $X^{\mathrm{T}}XB = UDV^{\mathrm{T}}$;

**update** $A$ by $UV^{\mathrm{T}}$;

**return** $B = [\beta_1, \beta_2, \ldots, \beta_M]$;

**rewrite** $\beta_m$ as $\beta_m = [\beta_m^{(1)^{\mathrm{T}}}, \beta_m^{(2)^{\mathrm{T}}}, \ldots, \beta_m^{(G)^{\mathrm{T}}}]^{\mathrm{T}}$, where $\beta_m^{(g)} = [\beta_{1,m}^{(g)}, \ldots, \beta_{p_g,m}^{(g)}]^{\mathrm{T}}$, for $1 \le g \le G$ and $1 \le m \le M$; and each element of the vector is non-zero;

**do** regression using the model (11);

**compute** $R_Y^g$ by Equation (12);

**return** $R_Y^g$;

**classify** $R_Y^g$ using a simple classifier.

**return** class membership of $Y_{test}$; With this classification scheme, we aim for a simple classification method such as the one nearest-neighbor (1-NN) or Fisher's linear discriminate (FLD), in order to determine the class membership of a test signal. The 1-NN classifier assigns a test signal to the class of its closest neighbor in the feature space by comparing the Euclidian distances of the feature vector for a test signal and the feature vectors obtained from the training set.

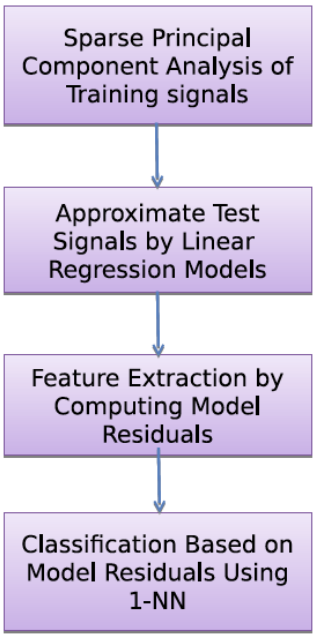



**Figure 1:** Flowchart of the proposed classification scheme.

The sparse feature extraction plus a simple classifier for signal classification is a typical approach for high dimensional data classification. In this work we emphasize on researching novel sparse feature extraction approach and developing its classification scheme. Within our proposed method, a signal is first partitioned into a set of signal segments and sparse PCA is used to learn a subset of most relevant signal segments. Since the principal component loading vectors can be treated as a dictionary, this approach can be considered as learning a window-wise sparse dictionary from the specific data. However, the difference between the traditional sparse representation and our approach is that we aim to obtain an average of signal dissimilarity measures (i.e., model residuals) using a localized regression based on the extracted sparse principal components, instead the traditional approach of sparse representation for classification focuses on extracting sparse signal features based on the signal similarity measure (i.e., model coefficients). Because the regression models are constructed for a signal based on different types of regressors that are characterized by different classes, the model residuals will behave similarly within the class and will be dissimilar among classes. This is why our approach leads to more appealing results when model residuals are used for classification.

## 4. EXPERIMENTAL RESULTS

In this section we demonstrate the utility of our developed method in signal classification. This experimental study is first based on the synthetic data,

and then on real EEG data. For the application to EEG data, the signals from the given database have been partitioned into signal segments of relatively small length, but there were from a long-term observational study.

### 4.1. Synthetic Data

The synthetic data that we considered is available in *physionet* (http://www.physionet.org/physiobank/database/synthetic/tns/). There are the following types of signals: (1) correlated stationary signals, denoted by $S_1$; (2) signals with sinusoidal trends, denoted by $S_2$; (3) signals with different local standard deviations, denoted by $S_3$; (4) signals with spikes, denoted by $S_4$. $S_1$ signals were simulated using a first order autoregressive model with parameter value $\alpha_1 = 0.1$ for the group 1, and $\alpha_2 = 0.5$ for the group 2. $S_2$ signals were simulated again using first order autoregressive model with added sinusoidal trends that have amplitude $A_s$. The group 1 signal has $\alpha_1 = 0.9$ and $A_s = 2$, and the group 2 signal has $\alpha_2 = 0.1$ and $A_s = 2$. $S_3$ signals were simulated using first order autoregressive model with changing local standard deviations. For group 1 signal, the model parameter values of $\alpha_1 = 0.1$, standard deviation $\sigma_1$=1 with probability 0.95 and $\sigma_2 = 4$ with probability 0.05 were used, while the group 2 signal using $\alpha_1 = 0.9$, standard deviation $\sigma_1$=1 with probability 0.05 and $\sigma_2 = 4$ with probability 0.95. As about the $S_4$ signals, the group 1 signal contains spikes that have amplitude $A_{sp} = 1$ and were simulated using probability $p = 0.05$, while group 2 is spikes signal only, and spikes were simulated using probability $p = 0.05$ and amplitude $A_{sp} = 10$. For more details of simulation mechanism, we refer readers to [27], [28]. We choose these types of signals because they are representative signals that share common characteristics with biomedical signals. Figure **1** shows examples of time series plots of signal types $S_3$ and $S_4$. The objective of this illustration is to see the performance of sparse variable approximation coupled with model residuals in signal classification problems, in terms of both sparsity of signal segments used for principal component construction and classification accuracy. Table **1** reports the sparsity of the principal component loading vectors. It specifies which signal segments are included for calculation of principal components. The sparsity is defined as the ratio of the number of nonzero elements and the length of signal segments. A small value of sparsity indicates high level of the sparseness of the principal component loading vectors, therefore a small number of signal segments are selected. From Table **1**, one can see that the level of sparsity for non-stationary signals, i.e. $S_3$ and $S_4$, is much higher than the one for stationary signals, i.e. $S_1$ and $S_2$. That is, the sparsity measure as defined in this paper is much smaller. For example, less than 4% of signal segments are needed for constructing principal components of $S_4$. On the other hand, the sparsity measure is decreasing with the increase of the number of $ith$ principal component. This is due to the fact that data variance explained by the $ith$ principal components is decreasing with the increase of $i$. Based on the selected signal segments, the classification scheme discussed in Section 3 is applied to the classification problems for all types of signals. The obtained classification accuracy is reported in Table **2**. From the results in Table **2**, one can see that with an appropriate choice of signal segment length, the classification accuracies of all classification problems that we consider are nearly perfect. Within this experiment, we also observe that the classification accuracy is increased with the increase of signal segment length $n$. This is due to the fact that the longer the length of signal segments is, the better the chance that the difference between signals can be captured.

### 4.2. EEG Data

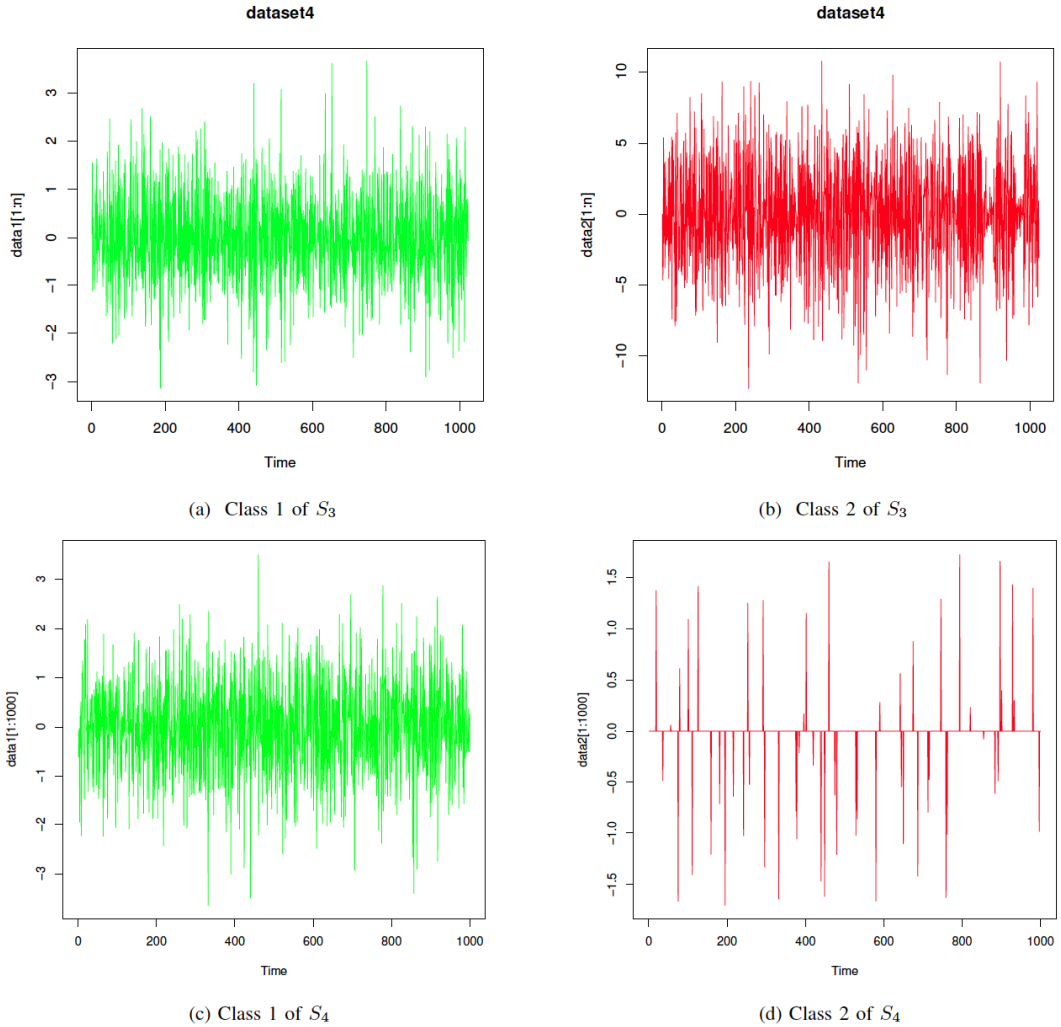
For application of the proposed method to biomedical signals, we use a set of EEG signals. EEG signals have been widely used for *long-term* monitoring of epilepsy as well as its diagnosis problem. They are typically multi-scale and non-stationary in nature. This database is from the University of Bonn, Germany.

**Table 1: The Sparsity of the Principal Component Loading Vectors. For $S_1$ and $S_2$ there are 4096 Signal Segments for Learning Principal Components. There are 1024 and 2048 Signal Segments for $S_3$ and $S_4$, Respectively**

| Dataset | 1st PC | 2nd PC | 3rd PC | 4th PC | 5th PC | 6th PC |
|---|---|---|---|---|---|---|
| $S_1$ | 0.2947 | 0.2776 | 0.2793 | 0.2763 | 0.2546 | 0.2341 |
| $S_2$ | 0.9919 | 0.2668 | 0.1045 | 0.0657 | 0.0598 | 0.0600 |
| $S_3$ | 0.1708 | 0.0480 | 0.0020 | 0.0000 | 0.0000 | 0.0000 |
| $S_4$ | 0.0142 | 0.0304 | 0.0132 | 0.0117 | 0.0107 | 0.0014 |

**Table 2: Classification Errors for Classification Problems of Different Type of Synthetic Data Based on 10-Fold Cross-Validation Using Supervised Sparse Variable Approximation Classification Scheme. $n$ is the Optimal Length of Signal Segments**

| Classification problems | Overall error | False positive rate | False negative rate |
|---|---|---|---|
| $S_1$ ( $n=2^5$ ) | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| $S_2$ ( $n=2^5$ ) | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| $S_3$ ( $n=2^8$ ) | 0.016 ± 0.000 | 0.000 ± 0.000 | 0.031 ± 0.000 |
| $S_4$ ( $n=2^6$ ) | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| A,C ( $n=2^7$ ) | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| A,D ( $n=2^7$ ) | 0.000 ± 0.000 | 0.000 ± 0.000 | 0.000 ± 0.000 |
| B,C ( $n=2^7$ ) | 0.007 ± 0.013 | 0.004 ± 0.008 | 0.010 ± 0.025 |
| B,D ( $n=2^7$ ) | 0.004 ± 0.007 | 0.004 ± 0.008 | 0.004 ± 0.013 |
| A;B,C;D ( $n=2^7$ ) | 0.001 ± 0.002 | 0.000 ± 0.000 | 0.002 ± 0.004 |




(a) Class 1 of $S_3$ (b) Class 2 of $S_3$

(c) Class 1 of $S_4$ (d) Class 2 of $S_4$

**Figure 2:** (**a**) and (**b**) are the time series plot of the first $2^{10}$ time points of the signals with different local standard deviation. (**c**) and (**d**) are the plots for the signals with spikes.

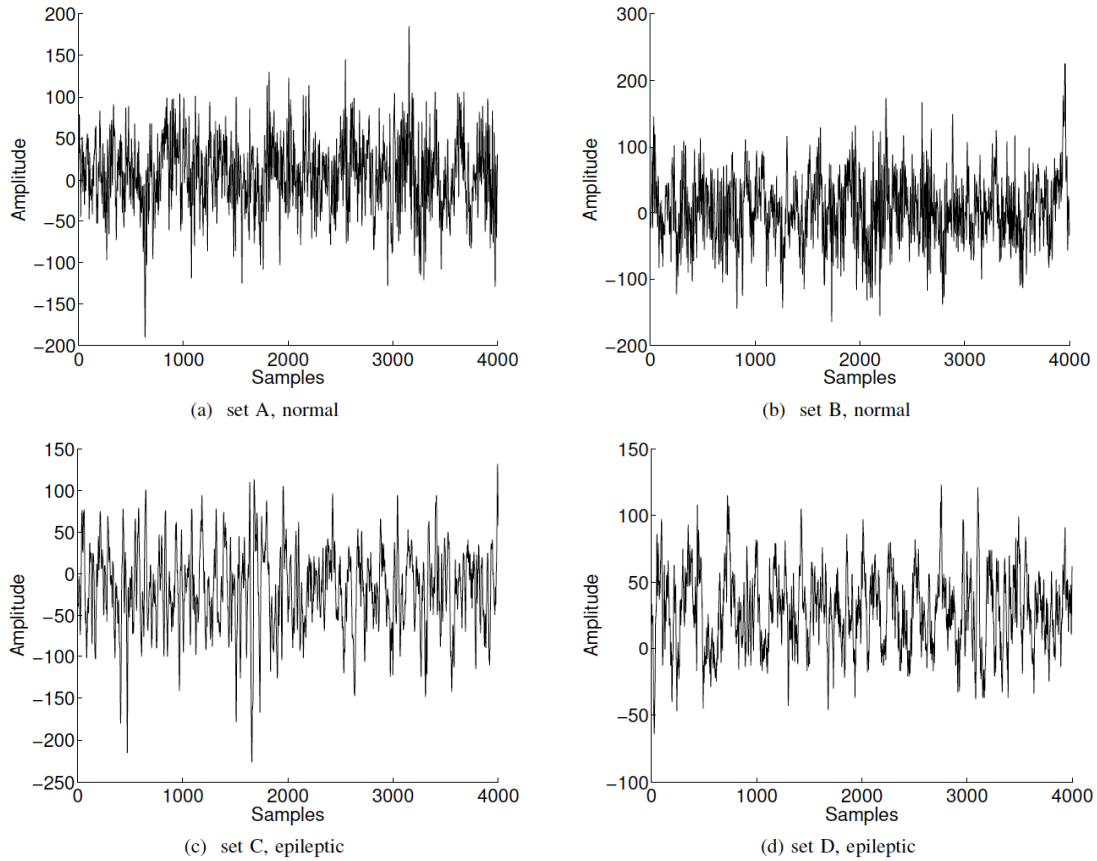



**Figure 3:** (**a**) and (**b**) are the time series plots of the 4096 time points of the signals sampled from a normal person. (**c**) and (**d**) are the plots for the same length of signals sampled from an epileptic patient.

There are four different sets of non-seizure signals, denoted as A, B, C, and D. Data in sets A and B are normal surface EEG signals with eyes closed and open, respectively. Data in sets C and D are transcranial EEG recordings coming from patients suffering from epilepsy. Each class contains 100 single channel scalp EEG segments of 23.6 second duration and sampled at 173.61 Hz (i.e., $T$ =4096).

We focus on classifying different types of signals, i.e. surface and transcranial EEG recordings. We consider five classification problems including A, C; A, D; B, C; B, D; and A; B, C; D. In the last classification problem, data set A and B is combined to be normal data and C and D are formed into abnormal data that indicates the presence of epilepsy. Figure **2** shows the EEG samples for both normal surface recordings and transcranial recordings. Using our proposed method, these signals are transformed into a two-dimensional feature space spanned by the two model residuals. The extracted features in terms of model residuals of two classes are shown in Figures **4** and **5** for the first four

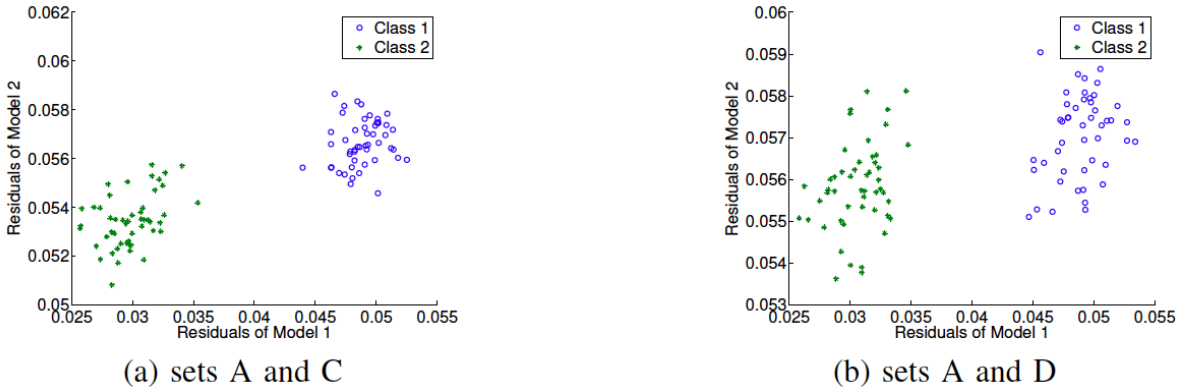
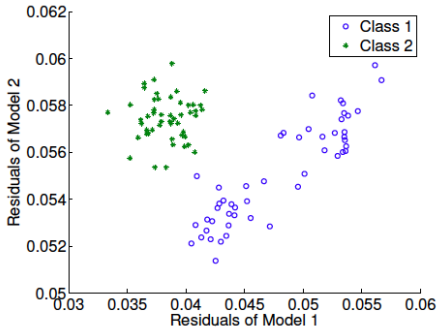
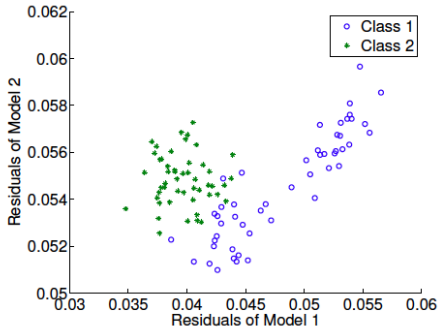



**Figure 4:** Scatter plots of signal residuals from the regression models applied to the test signal segments. The class 1 signals are normal EEG signals (i.e., set A) and the class 2 signals are epileptic EEG signals.




(c) sets B and C




(d) sets B and D

**Figure 5:** Scatter plots of signal residuals from the regression models applied to the test signal segments. The class 1 signals are normal EEG signals(i.e., set B) and the class 2 signals are epileptic EEG signals.

classification problems. The fact that these extracted features become clusters in the feature space leads to a high classification accuracy.

Table **3** shows comparisons of classification errors obtained from different approaches that we take. For an appropriate choice of the length of signal segments, i.e. $n = 2^7$ (it is based on 10-fold cross-validation procedure), our proposed method (i.e., model residuals based) is able to achieve a nearly perfect classification accuracy using 1-NN as a classifier. However, when the method uses the model coefficient, it fails in classifying these EEG signals for both two classifiers. The FLD method performs poorly. This may imply the non-linearity of extracted features (i.e. model coefficients). Generally speaking, with an increase in $n$, the bias of the estimate of model residuals increases as the number of signal segments available decreases. Our method achieves a high accuracy with a small $n$ which implies a small bias of model residual estimate.

We also compare the classification result for the classification problem A; B, C; D to the ones that use

**Table 3: Classification errors for different classification problems of EEG data based on 10-fold cross-validation using, respectively, our classification scheme, classification using 1-NN and the coefficients estimated from supervised sparse variable approximation and the linear discriminate analysis of the coefficients. $n$ is the optimal length of signal segments**

| Classification Problems | Our Classification Scheme | Coefficients + 1-NN | Coefficients + FLD |
|---|---|---|---|
| A,C ( $n=2^7$ ) | 0.0000 | 0.3570 | 0.4923 |
| A,D ( $n=2^7$ ) | 0.0000 | 0.3308 | 0.4846 |
| B,C ( $n=2^7$ ) | 0.0070 | 0.3082 | 0.4924 |
| B,D ( $n=2^7$ ) | 0.0040 | 0.3121 | 0.4858 |
| A;B,C;D ( $n=2^7$ ) | 0.0000 | 0.3334 | 0.4850 |

**Table 4: The comparison of average classification error for the classification problem of using A and B, the normal surface signals, and C and D, the transcranial EEG recodrings between the method that uses wavelet variances as the input of various classifiers and the proposed method coupled with the optimal choice of length of signal segment $N=2^7$. Different lengths of signal segments are used for wavelet variance approach in order to demostrate the effect of signal segment length on classification accuracy.**

| | 1-NN | 3-NN | SVM | SPCA+1-NN |
|---|---|---|---|---|
| $N=2^7$ | | | | 0.10% ± 0.20% |
| $N=2^{10}$ | 0.38% ± 0.24% | 0.64% ± 0.31% | 10.66% ± 0.76% | |
| $N=2^{11}$ | 0.15% ± 0.24% | 0.29% ± 0.22% | 2.95% ± 0.63% | |
| $N=2^{12}$ | 0.00% ± 0.00% | 0.00% ± 0.00% | 0.00% ± 0.05% | |

wavelet based functional linear model [11] with various length choice of signal segments. In this comparison, the optimal classification result that uses $N = 2^7$ for this proposed method is comparable to the result that use wavelet based linear functional model with larger signal segment size (i.e. $N = 2^{12}$). The main difference between these two methods is that the present method requires a smaller segment size, but the method in [11] need a much bigger size of signal segment. This may be due to the process of selecting only discriminative signal segments.

## 5. DISCUSSIONS

The EEG data set that we use in this work has been also analyzed and presented in many research publications including [29-31]. However, most of the existing works including [29-31] utilize this data set for the seizure detection problem by considering signals with seizure onset, which we do not consider in this work. We found that the signal power of the seizure type was much higher than that of other types such as normal and seizure-free epileptic signals. The classification between seizure and non-seizure types of signals has been widely studied and many promising results have been obtained. Therefore, we do not consider the classification problem that involves seizure type signals. This is because our proposed scheme is more suitable for application to signals with similar sample standard deviations so that the selection of signal segments is mainly based on the signal similarity measured by the inter-signal correlation, rather than the energy of a signal. Because we did not classify the seizure signals, we can not directly compare the performance of proposed method to other existing studies that focus on the epileptic seizure detection problem. However, we conclude that this presented work is superior for the data we consider as the classification result is almost perfect. There are many recent publications including [32-34], to name a few, that present research outcome using various advanced learning algorithms, but none of them focus on sparse signal segment extraction for the training process.

## 6. CONCLUDING REMARKS

In this paper we presented a novel classification scheme based on sparse PCA for long-term observational signals. Our major contribution was to propose a simultaneous signal segments selection and principal component construction along with the construction of regression models for test signal segments. The construction of regression models aims at obtaining model residuals and use them for classification. This makes our work different from existing work where model coefficients were used. Our experiments showed that the model residuals based method perform much better than the model coefficients based methods. The promising results suggest that the proposed method may be applied to event detection problems using biomedical signals.

The major benefit of using this proposed method is the data dimension reduction for the training data set. It leads to more robust estimates of signal features for classification because only the relevant signal segments are selected for constructing principal components. The proposed technique are different from many existing sparse techniques (e.g., [35]) because it focuses on the extraction of selective signal segments from the training data set and aims for signal features with high discrimination power, rather than other techniques that emphasize on signal representation by sparse components and the classification problem that make use of the sparse components. The limitation of this method is that it may be not suitable for classification of long-term signals with significant difference of signal powers among classes. Fortunately, many other basic feature extraction methods such as Fourier and wavelet methods are good candidates when one has to deal with classification of signals with different signal powers. Also our approach of extracting signal features using linear regression may lead to another aspect of limitations as we consider only the linear dependency among signal segments. It is of our interest to further investigate the impact when dependence structure among signal segments is non-linear. This will be left as our future work.

## REFERENCES

[1] Bao LJ, Zhu YM, Liu WY, Croisille P, Pu ZB, Robini M, Magnin IE. Denoising human cardiac diffusion tensor magnetic resonance images using sparse representation combined with segmentation. Phys Med Biol 2009; 54: 1435-1456. https://doi.org/10.1088/0031-9155/54/6/004

[2] Provost J, Lesag F. The Application of Compressed Sensing for Photo-Acoustic Tomography. IEEE Transactions On Medical Imaging 2009; 28(4): 585-593. https://doi.org/10.1109/TMI.2008.2007825

[3] Huang HF, Hu GS, Zhu L. Sparse Representation-Based Heartbeat Classification Using Independent Component Analysis. Journal of Medical Systems 2010; 0148-5598: 1-13,.

[4] Scholler S, Purwins H. Sparse Approximations for Drum Sound Classification. IEEE Journal Of Selected Topics In Signal Processing 2011; 5(5): 933-940. https://doi.org/10.1109/JSTSP.2011.2161264

[5] Rubinstein R, Bruckstein AM, Elad M. Dictionaries for Sparse Representation Modeling. Proceedings of the IEEE 2010; 98(6): 1045-1057. https://doi.org/10.1109/JPROC.2010.2040551

[6] Yaghoobi M, Blumensath T, Davies ME. Dictionary Learning for Sparse Approximations With the Majorization Method. IEEE Transactions On Signal Processing 2009; 57(6): 2178-2191. https://doi.org/10.1109/TSP.2009.2016257

[7] Mallat S, Zhang Z. Matching Pursuit with Time-Frequency Dictionaries. IEEE Transaction On Signal Processing 1993; 41(12): 3397-3415. https://doi.org/10.1109/78.258082

[8] Pearson K. On lines and planes of closest fit to systems of points in space. Phil Mag 1901; 2(6): 559-572. https://doi.org/10.1080/14786440109462720

[9] Xie S, Krishnan S, Lawniczak A. Sparse Principal Component Extraction and Classification of Long-term Biomedical Signals. In: Proceedings of the 25th IEEE International Symposium on Computer Based Medical System 2012; 1-6. https://doi.org/10.1109/cbms.2012.6266371

[10] Xie S, Krishnan S. Learning Sparse Dictionary for Long-term Signal Classification and Clustering, in: Proceedings of the 11th International Conference on Information Science, Signal Processing and their Applications 2012; 1151-156. https://doi.org/10.1109/isspa.2012.6310458

[11] Xie S, Krishnan S. Wavelet Based Sparse Functional Linear Model with Applications to EEGs Seizure Detection and Epilepsy Diagnosis. Medical & Biological Engineering & Computing 2013; 51(1): 49-60. https://doi.org/10.1007/s11517-012-0967-8

[12] Xie S, Krishnan S, Dynamic Principal Component Analysis with Non-overlapping Moving Window and Its Applications to Epileptic EEG Classification. The Scientific World Journal 2014; (2014): Article ID 419308, 10.

[13] Huang K, Aviyente S. Sparse representation for signal classification. In Adv NIPS 2006.

[14] Tošic I, Frossard P. Dictionary Learning for Stereo Image Representation. IEEE Transactions On Image Processing 2011; 20(4): 921-934. https://doi.org/10.1109/TIP.2010.2081679

[15] Pati Y, Rezaiifar R, Krishnaprasad P. Orthogonal Matching Pursuit : recursive function approximation with application to wavelet decomposition. in Asilomar Conf. on Signals, Systems and Comput 1993.

[16] Chen SS, Donoho DL, Saunders MA. Atomic Decomposition by Basis Pursuit. Siam Review 2001; 43(1): 129-159. https://doi.org/10.1137/S003614450037906X

[17] Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statist Soc B 1996; 58(1): 267-288.

[18] Tropp JA. Greed is good: Algorithmic results for sparse approximation. IEEE Trans Inform Theory 2004; 50(10): 2231-2242. https://doi.org/10.1109/TIT.2004.834793

[19] Tropp JA, Gilbert AC, Strauss MJ. Algorithms for simultaneous sparse approximation, Part I: greedy pursuit. Signal Process 2006; 86(3) 572-588. https://doi.org/10.1016/j.sigpro.2005.05.030

[20] Tropp JA, Gilbert AC, Strauss MJ. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. Signal Processing 2006; 86(3): 572-588. https://doi.org/10.1016/j.sigpro.2005.05.030

[21] Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell 1989; 11: 674-693. https://doi.org/10.1109/34.192463

[22] Donoho D, Johnstone I, Kerkyacharian G, Picard D. Wavelet shrinkage: Asymptopia? J R Statist Soc B 1995; 57: 301-369.

[23] Donoho D, Johnstone I. Minimax estimation via wavelet shrinkage. Ann Statist 1998; 26: 879-921. https://doi.org/10.1214/aos/1024691081

[24] Chipman HA, Gu H. Interpretable Dimension Reduction. Journal of Applied Statistics 2005; 32(9): 969-987. https://doi.org/10.1080/02664760500168648

[25] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. Journal of Computational and Graphical Statistics 2006; 15(2): 262-286. https://doi.org/10.1198/106186006X113430

[26] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B 2005; 67(2): 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[27] Hu K, Ivanov PCh, Chen Z, Carpena P, Stanley HE. Effects of trends on detrended fluctuation analysis. Phys Rev E 2001; 64: 011114. https://doi.org/10.1103/PhysRevE.64.011114

[28] Chen Z, Ivanov PCh, Hu K, Stanley HE. Effects of nonstationarities on detrended fluctuation analysis. Phys Rev E 2002; 65: 041107. https://doi.org/10.1103/PhysRevE.65.041107

[29] Gautama T, Mandic DP, Van Hulle M. ndications of nonlinear structures in brain electrical activity". Phys Rev E 2003; 67: 046204. https://doi.org/10.1103/PhysRevE.67.046204

[30] Nigam VP, Graupe D. A neural-network-based detection of epilepsy. Neurol Res 2004; 26: 55-60. https://doi.org/10.1179/016164104773026534

[31] Zhu G, Li Y, Wen P. Epileptic seizure detection in EEGs signals using a fast weighted horizontal visibility algorithm. Computer Methods and Programs in Biomedicine 2014; 115(2): 64-75. https://doi.org/10.1016/j.cmpb.2014.04.001

[32] Ahangi A, Karamnejad M, Mohammadi N, Ebrahimpour R, Bagheri N. Multiple classiïOer system for EEG signal classiïOcation with application to brain-computer interfaces. Neural Comput & Applic 2013; 23: 1319-1327. https://doi.org/10.1007/s00521-012-1074-3

[33] Yuan Q, Zhou W, Li S, Cai D. Epileptic EEG classification based on extreme learning machine and nonlinear features. Epilepsy Res 2011; 96(1-2): 29-38. https://doi.org/10.1016/j.eplepsyres.2011.04.013

[34] Ghaffari A, Ebrahimi Orimi H. EEG signals classification of epileptic patients via feature selection and voting criteria in intelligent method. J Med Eng Technol 2014; 38(3):146-55. https://doi.org/10.3109/03091902.2014.890677

[35] Yang JY, Peng YG, Xu WL, Dai QH. Ways to sparse representation: An overview. Science in China Series F: Information Sciences 2009; 52(4): 695-703. https://doi.org/10.1007/s11432-009-0045-5

Received on 23-12-2016 Accepted on 01-02-2017 Published on 28-02-2017

https://doi.org/10.6000/1929-6029.2017.06.01.2