Statistics and Its Interface Volume 15 (2022) 225–236

Principal wave analysis for high-dimensional structured data with applications to epigenomics and neuroimaging studies

Yuping Zhang*

High-dimensional structured data are emerging and accumulating in biomedical research fields. Examples include epigenomics and neuroimaging studies. In these studies, it is often required to extract biologically meaningful patterns and identify relevant biological features from high-dimensional structured data. Motivated by this problem, we propose a new statistical learning method named Principal Wave Analysis (PWA). The practical merits of PWA are shown through simulation studies incorporating diverse types of signal patterns as well as its applications to epigenomic and neuroimaging data.

Keywords and phrases: Dimension reduction, Feature selection, Multiresolution analysis, Wavelets.

1. INTRODUCTION

By virtual of high-throughput modern technologies in biomedical research, high-dimensional data are often generated routinely for a group of subjects from a cohort. These data usually contain certain structures of interest that can facilitate meaningful data interpretation. For instance, in epigenomics, one may characterize histone modification, DNA methylation, or gene expression profiles for a group of subjects to identify biologically meaningful epigenomic signatures. As another example, in neuroimaging studies, one may measure non-stationary signals of brain activities for a cohort to identify certain patterns with interesting structures.

Characterizing the meaningful patterns embedded in high-dimensional data with rich structures often plays an important role in biomedical research to address some significant scientific questions. For instance, to understand epigenomic mechanism for certain biological condition, it is important to dissect the underlying molecular activities based on the measured biological signals across many genomic loci for a group of subjects. The patterns of transcription factor binding and chromatin states are not uniformly distributed along the genome, and can exhibit context-dependent signatures. In this paper, we focus on the problem of extracting population-level patterns and identifying important features from high-dimensional measures for a cohort of subjects over time or across genomic loci. That is, we consider a $N \times P \times T$ three-order tensor measured on a "uniform lattice", where N is the number of subjects, P is the number of features, and T is the number of time points or genome loci.

Various dimensional reduction methods can be applied to a high-dimensional structured data matrix. We briefly review some of the representative methods, including principal component analysis (PCA), sparse PCA, functional PCA (FPCA), and wavelet PCA. PCA is a classical dimension reduction method. Let Y indicate an $N \times P$ data matrix containing the measurements of P features for N observations. PCA can be implemented by a singular value decomposition of the data matrix X, or a spectral decomposition of the covariance (or correlation) matrix. The singular value decomposition is of the form: $Y = UDV^{\mathsf{T}}$, where $UU^{\mathsf{T}} = I_N, VV^{\mathsf{T}} = I_P$, D is a diagonal matrix with diagonal elements $d_1 \geq d_2 \geq \cdots \geq d_K > 0$, I_N and I_P are identical matrices with dimensions $N \times N$ and $P \times P$, respectively. Each principal component in PCA is a linear combination of all the original variables. To facilitate feature selection, several related methods have been developed to produce modified principal components with sparse loadings [11, 30, 22]. PCA also has been extended to functional data analysis scenario. Functional data have two key features [17]. One is that measurements are taken on the same subject repeatedly over different time or different space. The other is the smoothness assumption, where the underlying curve has a certain degree of smoothness. Given N random trajectories $X_i(t)$, FPCA [2, 20, 24] is defined as the following expansion: $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} A_{ik}\phi_k(t)$, where $i \in \{1, \ldots, N\}$, $t \in \mathcal{T}$, $A_{ik} = \int_{\mathcal{T}}(X_i(t) - \mu(t))\phi_k(t)dt$ are the functional principal components (scores) of X_i, A_{ik} are independent across i with $E(A_{ik}) = 0$ and $Var(A_{ik}) = \lambda_k$, λ_k are the eigenvalues in descending order, and ϕ_k are the corresponding orthogonal eigenfunctions. Specifically, the eigenvalues and eigenfunctions can be obtained via the spectral decomposition of the estimated auto-covariance surface $Cov(X(s), X(t))$ on a grid of time points. In some time series data, extreme observations may exist, such as spikes. Wavelet basis functions are known to be localized in both time and frequency domains simultaneously, and thus can extract localized features from

*ORCID: 0000-0001-8986-0354.

a time-varying signal. Wavelet PCA [8] is defined as performing PCA on wavelet coefficients. Specifically, let Y be an $N \times T$ data matrix, of which row vectors contain dyadic observations for individuals $i = 1, \ldots, N$, and $T = 2^J$ for some integer $J \geq 0$. Let W be an $T \times T$ wavelet transformation. The wavelet decomposition of Y is of the form $Y = BW$, where B is a scalar coefficient matrix consisting of all wavelet coefficients from the different locations and scales. Then, PCA is applied on wavelet coefficient matrix B, which is equivalent to performing a rotation of B to new axis, and the first coordinate is the projection with the greatest variance. When Y is noisy, wavelet PCA employed a two-step approach. First, it performs wavelet shrinkage on B, which results in smoothing in the wavelet domain with an estimate $\hat{B}$. Then, PCA is applied on $\hat{B}$. Note that, if there is no shrinkage on wavelet coefficients, wavelet PCA is equivalent to conventional PCA.

For a high-dimensional structured three-way array (tensor) with multiple subjects from a cohort, we have previously developed principal trend analysis (PTA) [27] with temporal smooth structure modeling for extracting the underlying patterns and identify important features simultaneously. Splines are known to be continuous with easily defined derivatives, and thus are well suited to modeling smooth functions [12, 19, 16, 21]. In the PTA framework, the smoothing splines are used, where we use a complete basis, but then shrink the coefficients toward smoothness. Then, smoothed principal trends can be characterized. PTA was successfully applied to unsupervised learning of smoothed transcriptional responses in longitudinal measurements of gene expression in patients [27]. The PTA framework has been extended to other sophisticated research scenarios [28, 26]. These methods have been successfully applied to analyzing longitudinal gene expression data measured for a cohort, where smooth trends are assumed. Further, other latent factor models were developed for three-way arrays via tensor decompositions [13, 23].

As aforementioned, in certain real data applications, the underlying meaningful signals can have different properties, such as jump discontinuities, spikes, varying frequency behaviour, and smooth signals containing jumps/spikes. PTA was proposed by integrating sparse latent factor models for dimension reduction and feature selection with spline-based methods [12, 19, 21] for temporal structure modeling. Thus, it is necessary to develop a new statistical learning method beyond the smoothness assumption for entire underlying curves to analyze more variously structured data with high-dimensional features measured for a cohort. Here, we relax the assumption that the entire underlying signals are smooth and develop a new statistical learning method to analyze data with multiple types of embedded structures and a large number of features measured for a cohort. On one hand, we want to extract the underlying signal patterns embedded in the data, where discontinuities are allowed. On the other hand, we want to identify the features that have contributions to those meaningful patterns. We propose a new method, named Principal Wave Analysis (PWA), to address the aforementioned objectives simultaneously. The proposed PWA framework blends the dimension reduction and feature selection, and characterizes data structures. The paper is organized as follows. In Section 2, we present the statistical framework of PWA and its computation. In Section 3, we demonstrate the practical merits of PWA through simulations. In Section 4, we show the applications of PWA to real data. One is from an epigenomics study of human embryonic stem cells, and the other is from a neuroimaging study. We conclude our paper in Section 5.

2. METHOD

2.1 The model of principal wave analysis

Let x_{npt} be an element of tensor $\boldsymbol{X}$ for feature p, subject n, time point (or genome locus) t, where $p \in \{1, ..., P\}$, $n \in \{1, ..., N\}$, $t \in \{1, ..., T\}$. We use $\boldsymbol{X}_n$ to denote the $P \times T$ matrix containing the observations for subject n in $\boldsymbol{X}$. We consider a factor model with basis expansions. Since the motivating examples from cohort studies in epigenomics and brain activity signals, their data exhibit spike and non-stationary phenomena, we employ wavelets in our modeling. Wavelets are multiresolution basis functions that are localized in both time and frequency domains simultaneously, and thus are suitable for irregular functions with discontinuities. We decompose the matrix-valued measurements $\boldsymbol{X}_n$ for each subject from the same population using a common factor model $\boldsymbol{X}_n = \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Theta}^\mathsf{T}\boldsymbol{W} + \boldsymbol{E}_n$, where $n = 1, \ldots, N$, $\boldsymbol{A}$ denotes a $P \times K$ matrix of factor scores, $\boldsymbol{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K]$, and K is the number of factors; $\boldsymbol{W}$ denotes a $T \times T$ wavelet basis matrix, and $\boldsymbol{W}\boldsymbol{W}^\mathsf{T} = \boldsymbol{I}_T$, $\boldsymbol{I}_T$ is a $T \times T$ identity matrix; $\boldsymbol{\Theta}$ denotes a $T \times K$ matrix of coefficients, $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K]$; $\boldsymbol{D}$ denotes a $K \times K$ diagonal matrix with scalars $d_1, \ldots, d_K$ on its diagonal; $\boldsymbol{E}_n$ is the error term for the decomposition for subject n. The underlying patterns are characterized by $\boldsymbol{S} = \boldsymbol{D}\boldsymbol{\Theta}^\mathsf{T}\boldsymbol{W}$, which are generated by the combinations of wavelets. We name $\boldsymbol{S}$ as Principal Waves (PWs).

Notably, let $\bar{\boldsymbol{X}}$ denote the sample mean, i.e. $\bar{\boldsymbol{X}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_n$, and let $\bar{\boldsymbol{E}} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{E}_n$, then we have $\bar{\boldsymbol{X}}\boldsymbol{W}^\mathsf{T} - \bar{\boldsymbol{E}}\boldsymbol{W}^\mathsf{T} = \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Theta}^\mathsf{T}$. Denoting $\bar{\boldsymbol{X}}\boldsymbol{W}^\mathsf{T} - \bar{\boldsymbol{E}}\boldsymbol{W}^\mathsf{T}$ by $\tilde{\boldsymbol{X}}$, in the light of singular value decomposition (SVD) of $\tilde{\boldsymbol{X}}$, with the constraints $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_P$, $\boldsymbol{\Theta}\boldsymbol{\Theta}^\mathsf{T} = \boldsymbol{I}_T$, $d_1 \geq d_2 \geq \ldots \geq d_K > 0$, the factor model is identifiable for the latent matrix $\tilde{\boldsymbol{X}}$ up to a sign change and orthogonal rotations.

The wavelet bases used in the factor model can be obtained by translations and dilations of a single scaling function $\phi(t)$ with various types of forms [9, 6, 29]. With the high adaptability to different levels of smoothness, wavelet transforms are capable of capturing various types of discontinuities. We take the simple Haar wavelet transform as an example to present basic ideas. The Haar basis produces a piecewise-constant representation, where the scaling function (also known as the father) $\phi(t)$ is just the in-

dicator function of the interval $[0,1]$, the Haar base function (also known as Haar mother wavelet) is defined as $\psi(t) = I_{[0,\frac{1}{2})}(t) - I_{[\frac{1}{2},1]}(t)$. With binary dilation (scaling j) and dyadic translation (shifting l), we have $\psi_{j,l}(t) = 2^{j/2}\psi(2^j t - l)$, and $\phi_{j,l}(t) = 2^{j/2}\phi(2^j t - l)$, where $j = 0, 1, \ldots$ and $l = 0, \ldots, 2^j - 1$. Thus, $\psi_{j,l}$ and $\phi_{j,l}$ etc. form the basis vectors. Coefficients corresponding to $\psi_{j,l}$ are known as averages or sum coefficients, while coefficients corresponding to $\phi_{j,l}$ are differences or detail coefficients. In practice, to form the wavelet basis matrix, we only need $\psi_{0,0}$, and other columns are expressed in terms of $\phi_{\cdot,\cdot}$. To form a $T \times T$ wavelet basis matrix ($T = 2^J$), we need $2^J - 1$ basis vectors corresponding to detail coefficients and 1 basis vector corresponding to sum coefficient at 0^{th} level. The matrix $\boldsymbol{W}$ has these basis vectors as columns. For the numerical results in this paper, we used a more complicated transform based on Daubechies' extremal phase wavelets.

In this paper, we impose additional constraints on the elements of $\boldsymbol{A}$ and $\boldsymbol{\Theta}$. With a complete orthonormal wavelet basis, we may shrink and select the coefficients to represent different types of functions. Thus, we want to employ an L_1-norm regularization of coefficients toward a sparse representation. In addition, the number of features of data may be large, and only some of them behave interesting patterns over time or along the genome. Consequently, we want to select those important features simultaneously. We start with single-factor PWA, and consider the following optimization problem.

$$\min_{\boldsymbol{a},d,\boldsymbol{\theta}} \sum_{n=1}^{N} \|\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\|_F^2, \tag{1}$$

subject to $\|\boldsymbol{\theta}\|_1 \le c_1$, $\|\boldsymbol{a}\|_1 \le c_2$, $\|\boldsymbol{\theta}\|_2^2 = 1$, and $\|\boldsymbol{a}\|_2^2 = 1$, where $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_1$ is the L_1 norm, $\|\cdot\|_2$ is the L_2 norm, c_1 and c_2 are constant tuning parameters. We incorporate the L_2-norm equality constraints for $\boldsymbol{\theta}$ and $\boldsymbol{a}$ in the proposed PWA optimization framework. While the L_2-norm equality constraints make it a non-convex optimization problem, we can relax the constrains to $\|\boldsymbol{\theta}\|_2^2 \le 1$, and $\|\boldsymbol{a}\|_2^2 \le 1$. This is because the solutions $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{a}}$ minimizing the relaxed optimization problem also minimize the original one with the L_2-equality constraints on $\boldsymbol{\theta}$ and $\boldsymbol{a}$. Then, we have the following tri-convex optimization problem, which results in the same solution as problem 1.

Theorem 1. *The following optimization problem is tri-convex.*

$$\min_{\boldsymbol{a},d,\boldsymbol{\theta}} \sum_{n=1}^{N} \|\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\|_F^2,$$

subject to $\|\boldsymbol{\theta}\|_1 \le c_1$, $\|\boldsymbol{a}\|_1 \le c_2$, $\|\boldsymbol{\theta}\|_2^2 \le 1$, *and* $\|\boldsymbol{a}\|_2^2 \le 1$.

The tuning parameters c_1 and c_2 control the amount of sparsity in the estimated vectors $\boldsymbol{\theta}$ and $\boldsymbol{a}$, respectively. To shrink and select the coefficients of the complete orthonormal basis with $T \times T$ dimensions, the tuning parameter c_1 needs to be between 1 and $\sqrt{T}$. To select features from a dataset with the total number of features as P, the range of the tuning parameter c_2 needs to be between 1 and $\sqrt{P}$. By the Lagrangian duality theory, we can obtain $\boldsymbol{a}$, $\boldsymbol{\theta}$ and d by solving the following optimization problem (2) instead.

$$\min_{\boldsymbol{a},d,\boldsymbol{\theta}} \sum_{n=1}^{N} \|\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\|_F^2 + 2\lambda_1\|\boldsymbol{\theta}\|_1 + 2\lambda_2\|\boldsymbol{a}\|_1 + \lambda_3\|\boldsymbol{\theta}\|_2^2 + \lambda_4\|\boldsymbol{a}\|_2^2, \tag{2}$$

where λ_3 and λ_4 need to be chosen to make $\|\boldsymbol{a}\|_2 = 1$ and $\|\boldsymbol{\theta}\|_2 = 1$, respectively. The sparsity tuning parameters are represented by λ_1 and λ_2.

Please refer to the appendix materials (A1) for the proof of the triconvex property of optimization problem (2). The detailed algorithm can be found in Algorithm (1). The derivation of the PWA algorithm can be found in the appendix materials (A2). The solutions $\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{d}$ to optimization problem (2) also solve the problem (1).

Algorithm 1 The rank-one PWA algorithm

1: **procedure** TO FIND ($\hat{\boldsymbol{a}}$, $\hat{\boldsymbol{\theta}}$, $\hat{d}$)
2: initialize $\boldsymbol{a}^{(1)}$, $\boldsymbol{\theta}^{(1)}$ and $d^{(1)}$.
3: *repeat until convergence*: For k in $\{1, \ldots, k'\}$
4: $\boldsymbol{\theta}^{(k+1)} = \frac{\mathrm{soft}\left(\frac{1}{N}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n\right)^{\mathsf{T}}\boldsymbol{a}^{(k)}d^{(k)},\lambda_1\right)}{\left\|\mathrm{soft}\left(\frac{1}{N}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n\right)^{\mathsf{T}}\boldsymbol{a}^{(k)}d^{(k)},\lambda_1\right)\right\|_2}$, where $\mathrm{soft}(t,s)$ is a soft-thresholding function, defined as $\mathrm{sign}(t)(|t| - s)_+$.
5: $\boldsymbol{a}^{(k+1)} = \frac{\mathrm{soft}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_n\boldsymbol{W}\boldsymbol{\theta}^{(k+1)}d^{(k)},\lambda_2\right)}{\left\|\mathrm{soft}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_n\boldsymbol{W}\boldsymbol{\theta}^{(k+1)}d^{(k)},\lambda_2\right)\right\|_2}$.
6: $d^{(k+1)} = \frac{1}{N}(\boldsymbol{\theta}^{(k+1)})^{\mathsf{T}}\boldsymbol{W}(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}})\boldsymbol{a}^{(k+1)}$.
7: return $\hat{\boldsymbol{a}} = \boldsymbol{a}^{(k'+1)}$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(k'+1)}$ and $\hat{d} = d^{(k'+1)}$.
8: **end procedure**

2.2 Tuning parameter selection

The tuning parameters λ_1 and λ_2 (or equivalently, c_1 and c_2) control the amount of sparsity in the estimated vectors $\boldsymbol{\theta}$ and $\boldsymbol{a}$. Generally speaking, tuning parameter selection is a hard problem for unsupervised learning. Here, we transform the problem to the model selection of supervised learning by treating some of observations as "missing" and mimicking missing data imputation. Then, we have two options for the model selection. One is based on the extra-sample error, such as cross-validations. The other is based on in-sample error, such as Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). We elaborate the two options to estimate the penalty parameters in the proposed PWA framework.

We first introduce the cross-validation type of methods, which is similar to the extra-sample model selection approach in PTA [27]. Specifically, we can use the following resampling procedure to select tuning parameters. We construct m subsets of data from the original dataset $\boldsymbol{X}$, where

m is the number of folds stratified for cross-validation. Let $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^m$ denote these subsets of $\boldsymbol{X}$, each of which contains a nonoverlapping $\frac{1}{m}$ of the elements of $\boldsymbol{X}$. Each subset of data are obtained through random sampling from entries in $\boldsymbol{X}$. During the cross-validation procedure, each subset of data will serve as a validation set, treated as "missing" when the PWA is performed on the remaining $m-1$ folds. For each pair of candidate values of c_1 $(1 \leq c_1 \leq \sqrt{T})$ and c_2 $(1 \leq c_2 \leq \sqrt{P})$, we iterate the following steps. For each fold j, $(j \in \{1, \ldots, m\})$, we fit the PWA to $\boldsymbol{X}^j$ with the tuning parameter c_1 and c_2, and calculate $\hat{\boldsymbol{X}}^j$, the resulting estimate of $\boldsymbol{X}^j$. Each sample i of $\boldsymbol{X}^j$ is estimated by $\hat{\boldsymbol{a}}\hat{d}\hat{\boldsymbol{\theta}}^\top \boldsymbol{W}$. We then record the mean squared error of $\hat{\boldsymbol{X}}^j$. This mean squared error is obtained by computing the mean of the squared differences between elements of $\boldsymbol{X}^j$ and the corresponding elements of $\hat{\boldsymbol{X}}^j$. We then calculate the average mean squared error across $\boldsymbol{X}^1, \ldots, \boldsymbol{X}^m$ for the tuning parameters c_1 and c_2. Finally, the optimal values of c_1 and c_2 are those that correspond to the lowest mean squared error.

Alternatively, we can perform the model selection based on in-sample error. It is known that BIC is asymptotically consistent as a selection criterion, while AIC is not [4]. Thus, we propose to use a BIC type of approach to estimate the appropriate amount of sparsity in the vectors $\boldsymbol{\theta}$ and $\boldsymbol{a}$. Specifically, we propose the following formula as the BIC type of model selection criterion:

$$BIC_{type}(\lambda_1, \lambda_2) = \log\left(\frac{\sum_{i=1}^{N} \|\boldsymbol{X}_n - \hat{\boldsymbol{a}}\hat{d}\hat{\boldsymbol{\theta}}^\top \boldsymbol{W}\|_F^2}{NPT}\right) + \frac{\log(NPT)}{NPT}(\hat{df}(\lambda_1) + \hat{df}(\lambda_2)), \tag{3}$$

where $\hat{df}(\lambda_1)$ and $\hat{df}(\lambda_2)$ are estimates of the degrees of freedom for certain L_1-norm penalties employed on vectors $\boldsymbol{\theta}$ and $\boldsymbol{a}$, respectively. Here, we have $\hat{df}(\lambda_1) = |\hat{\boldsymbol{\theta}}|$, and $\hat{df}(\lambda_2) = |\hat{\boldsymbol{a}}|$, which are the numbers of nonzero elements in $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{a}}$, respectively. In practice, the BIC type of approach is usually less computationally intensive than the cross-validations. Thus, we employed the proposed BIC type of method (3) for model selections in our simulation studies and real applications.

2.3 PWA with multiple factors

Using the PWA with single factor as the core building block, we can sequentially build the PWA model with multiple factors. For instance, with the rank-one PWA result, we can obtain the rank-two PWA through fitting a rank-one decomposition on the residues. The residues are obtained by removing the rank-one PWA estimates from the original dataset $\boldsymbol{X}$. Similarly, we can employ a cross-validation or BIC type of approach to select appropriate tuning parameters for the rank-two PWA. We explicitly present the rank-K PWA $(K \geq 2)$ in Algorithm (2).

Algorithm 2 The rank-K PWA algorithm

1: **procedure** TO FIND $(\hat{\boldsymbol{a}}_K, \hat{\boldsymbol{\theta}}_K, \hat{d}_K)$
2: let $\hat{\boldsymbol{X}}_n^{(1)} = \boldsymbol{X}_n$ and initialize $\hat{\boldsymbol{a}}_1$, $\hat{\boldsymbol{\theta}}_1$ and $\hat{d}_1$ through the rank-one PWA algorithm.
3: *loop*: For k in $\{1, \ldots, K-1\}$
4: set $\hat{\boldsymbol{X}}_n^{(k+1)} = \hat{\boldsymbol{X}}_n^{(k)} - \hat{\boldsymbol{a}}_k \hat{d}_k \hat{\boldsymbol{\theta}}_k^\top \boldsymbol{W}$
5: apply the rank-one PWA algorithm to $\hat{\boldsymbol{X}}_n^{(k+1)}$ to obtain $\hat{\boldsymbol{a}}_{k+1}$, $\hat{\boldsymbol{\theta}}_{k+1}$, $\hat{d}_{k+1}$
6: return $\hat{\boldsymbol{a}}_K$, $\hat{\boldsymbol{\theta}}_K$ and $\hat{d}_K$.
7: **end procedure**

For high-dimensional structured three-way arrays with multiple subjects from a cohort, PTA and PWA can extract the underlying patterns and identify important features simultaneously. In the PTA framework, the smoothing splines are used, where we use a complete basis, but then shrink the coefficients toward smoothness. Then, smoothing principal trends can be characterized. PTA was successfully applied to unsupervised learning of longitudinal gene expression data by assuming smoothing transcriptional dynamics [27]. In the PWA framework, wavelets are used to form a complete orthonormal basis matrix. We then shrink and select the coefficients toward a sparse representation. As a result, we obtain principal waves without a smoothness assumption on the entire underlying signal patterns embedded in the data. We use the following simulation studies and real applications to show the practical merits of PWA.

3. SIMULATIONS

We performed simulation studies to demonstrate the performance of PWA for both rank-one and rank-K $(K > 1)$ scenarios. For the rank-one scenario, the simulated dataset contains one signal component with various types of characteristics. We simulated longitudinal datasets denoted by an $N \times P \times T$ tensor $\boldsymbol{X}$ with $N = 10$ subjects, $T(= 2048)$ time points, and $P(= 100)$ features, respectively. We adopted various functions that can reproduce phenomena found in real world signals as discussed in the literature [7]. Specifically, we considered four types of functions. The first type of functions represent jump discontinuities, i.e. blocks. We assigned values to x_{npt}, $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T\}$ and $p \in \{1, \ldots, P\}$, following Equation 4:

$$x_{npt} = \frac{1}{2} b_{1,p} f(t) + b_{2,p} \epsilon_{npt}, \tag{4}$$

and

$$f(t) = \sum_j h_j \left[1 + sign\left(\frac{t}{T} - t_j\right)\right], \tag{5}$$

where $(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2)$, $(t_j) = (0.1, 0.13, 0.15, 0.23, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$, $b_{1,p}$ is $I(0 < p \leq 60)$, $b_{2,p}$ is $I(0 < p \leq P)$, $\epsilon_{npt} \sim N(0, 1)$. The signal component in this equation has block patterns.

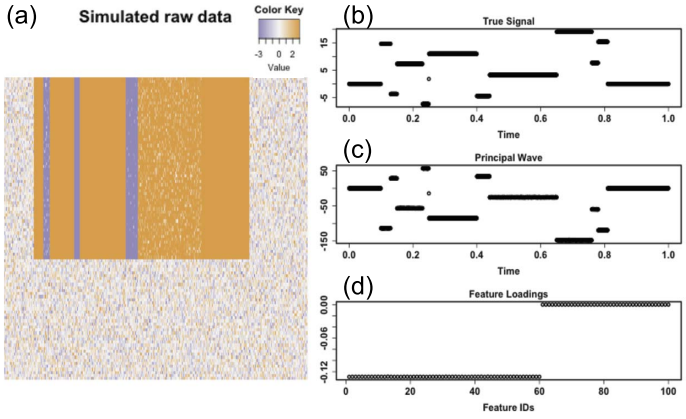

Figure 1. The simulation study for a signal scenario with blocks. (a) Simulated raw data represented by one individual sampled from 10 subjects. Each row indicate one feature. Columns are ordered by time. (b) True signals generated by Equation 5. (c) The principal wave (PW) identified by PWA. (d) Feature loadings identified by PWA.

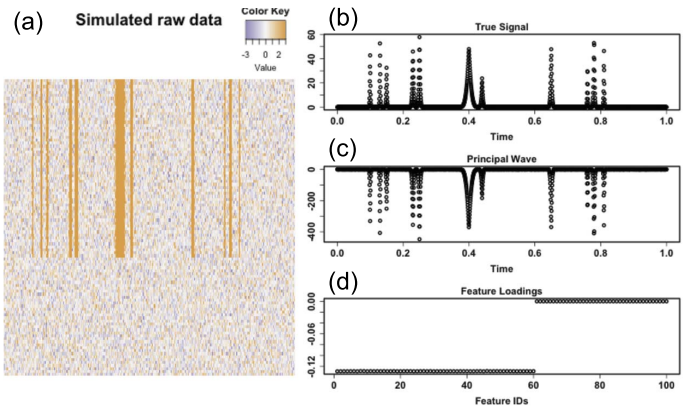

Figure 2. The simulation study for a signal scenario containing spikes. (a) Simulated raw data represented by one individual sampled from 10 subjects. Each row indicate one feature. Columns are ordered by time. (b) True signals generated by Equation 7. (c) The principal wave (PW) identified by PWA. (d) Feature loadings identified by PWA.

Figure 1(a) shows the heatmap of simulated raw data from one individual. The true latent pattern is shown in Figure 1(b). We applied PWA to this dataset and used the proposed BIC-type of criterion to select tuning parameters. The PW is shown in Figure 1(c), representing the estimated signal. The corresponding feature loadings are shown in Figure 1(d). One can see that the top 60 features have negative loadings, and the remaining 40 features have zero loadings. On one hand, PWA can identify informative features having contributions to the PW. On the other hand, the feature loadings multiplied by the PW reported by PWA reflect the signal patterns of the input data.

The second type of functions consist of spikes, which are similar to the bump signals observed in nuclear magnetic resonance spectroscopy experiments. In this simulation, we assigned values to x_{npt}, $n \in \{1,\dots,N\}$, $t \in \{1,\dots,T\}$ and $p \in \{1,\dots,P\}$, following Equation 6:

$$x_{npt} = b_{1,p} g(t) + b_{2,p}\epsilon_{npt}, \tag{6}$$

and

$$g(t) = \sum_j h_j \left(1 + \frac{1}{w_j}\left|\frac{t}{T} - t_j\right|\right)^{-4}, \tag{7}$$

where $(h_j) = (4, 5, 3, 4, 5, 4.2, 2.1, 4.3, 3.1, 5.1, 4.2)$, $(t_j) = (0.1, 0.13, 0.15, 0.23, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$, $(w_j) = (0.005,\ 0.005,\ 0.006,\ 0.01,\ 0.01,\ 0.03,\ 0.01,\ 0.01,\ 0.005,\ 0.008, 0.005)$, $b_{1,p}$ is $I(0 < p \le 60)$, $b_{2,p}$ is $I(0 < p \le P)$, $\epsilon_{npt} \sim N(0,1)$. The heatmap of simulated raw data from one individual is shown in Figure 2(a). The true latent pattern is shown in Figure 2(b). We applied PWA to this dataset and used the proposed BIC-type of criterion to select tuning parameters. The PW is shown in Figure 2(c). The corresponding feature loadings are shown in Figure 2(d). One can see that the top 60 features have negative loadings, and the remaining 40 features have zero loadings. PWA not only identifies informative features, but also captures the signal patterns of the input data characterized by the feature loadings multiplied by the PW.

In the third simulation study, we utilized functions that represent smooth signals with jumps/spikes in them. Specifically, we assigned values to x_{npt}, $n \in \{1,\dots,N\}$, $t \in \{1,\dots,T\}$ and $p \in \{1,\dots,P\}$, following Equation 8.

$$x_{npt} = b_{1,p} m(t) + b_{2,p}\epsilon_{npt}, \tag{8}$$

where $b_{1,p}$ is $I(0 < p \le 60)$, $b_{2,p}$ is $I(0 < p \le P)$, $\epsilon_{npt} \sim N(0,1)$, and

$$m(t) = 4\left[\sin\left(\frac{4\pi t}{T}\right) - sign\left(\frac{t}{T} - 0.3\right) - sign\left(0.72 - \frac{t}{T}\right)\right]. \tag{9}$$

The corresponding heatmap of raw simulated data is shown in Figure 3(a), illustrated by one subject. The true latent pattern embedded in the simulated data is shown in Figure 3(b), which is based on the corresponding signal component in Equation 8. We applied PWA to this dataset, and found that the identified PW reflects the embedded signals as shown in Figure 3(c). The features exhibiting the temporal signals can be identified by PWA with nonzero loadings.

The fourth simulation study was motivated by the Doppler shift phenomena in real world waves. Take a passing siren as an example, one may notice the sudden change in pitch. This phenomenon is caused by the well-known Doppler effect in physics, which indicates an increase (or

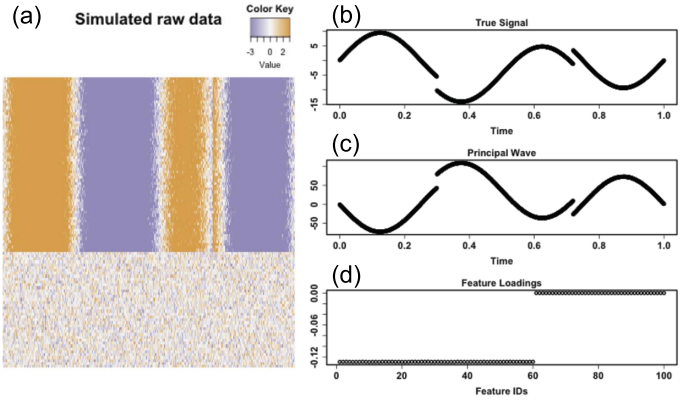

Figure 3. The simulation study considering smooth signals with jumps/spikes in them. (a) Simulated raw data represented by one individual sampled from 10 subjects. Each row indicate one feature. Columns are ordered by time. (b) True signals generated by Equation 9. (c) The principal wave (PW) identified by PWA. (d) Feature loadings identified by PWA.

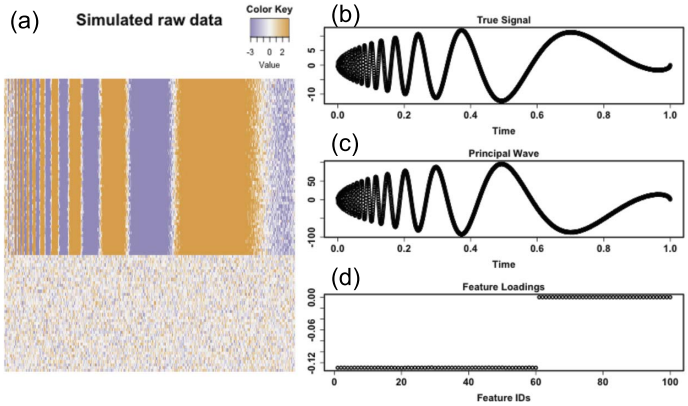

Figure 4. The simulation study motivated by the Doppler shift phenomena in real world waves. (a) Simulated raw data represented by one individual sampled from 10 subjects. Each row indicate one feature. Columns are ordered by time. (b) True signals as generated by Equation 11. (c) The principal wave (PW) identified by PWA. (d) Feature loadings identified by PWA.

decrease) in the frequency of waves as the source and observer move toward (or away from) each other. To mimic the Doppler effect in this simulation study, we incorporated a signal component using functions that have a varying frequency behavior. Precisely, we assigned values to x_{npt}, $n \in \{1, \ldots, N\}$, $t \in \{1, \ldots, T\}$ and $p \in \{1, \ldots, P\}$, following Equation 10.

$$x_{npt} = b_{1,p} l(t) + b_{2,p} \epsilon_{npt}, \tag{10}$$

and

$$l(t) = \frac{[t(T-t)]^{\frac{1}{2}}}{T} \sin\left[\frac{2.1T\pi}{t + 0.05 \cdot T}\right], \tag{11}$$

where $b_{1,p}$ is $I(0 < p \leq 60)$, $b_{2,p}$ is $I(0 < p \leq P)$, $\epsilon_{npt} \sim N(0, 1)$. We plotted the heatmap of raw data from one individual as shown in Figure 4(a). The true latent pattern function based on Equation 10 is represented in Figure 4(b). By applying the PWA to this dataset, one can see that the identified PW characterizes the embedded signal as demonstrated in Figure 4(c). Furthermore, as shown in Figure 4(d), the features with nonzero loadings in PWA are consistent with those features carrying signals according to the simulation.

To demonstrate the performance of PWA with multiple ranks, we simulated a dataset with aforementioned four types of signals simultaneously, i.e. jump discontinuities, spikes, varying frequency behavior, and jumps/spikes in smooth signals. Specifically, we generated x_{npt} based on Equation 12

$$x_{npt} = b_{1,p} f(t) + b_{2,p} g(t) + b_{3,p} m(t) + b_{4,p} l(t) + \epsilon_{npt}, \tag{12}$$

where $n \in \{1, \ldots, N\}$, $p \in \{1, \ldots, P\}$, $t \in \{1, \ldots, T\}$, $b_{1,p} = I(0 < p \leq 50)$, $b_{2,p} = I(50 < p \leq 90)$, $b_{3,p} = I(90 < p \leq 120)$, $b_{4,p} = I(120 < p \leq 140)$, and $f(t)$, $g(t)$, $m(t)$ and $l(t)$ are based on Equations 5, 7, 9 and 11, respectively.

We applied the PWA with multiple ranks to this dataset, and employed the proposed BIC type of method to select the tuning parameters based on the first local minimum for each rank sequentially. We illustrate the BIC type of curves in the left panel of Figure 5 with $\|\boldsymbol{\theta}\|_1$ fixed as 0.2, 0.1, 0.2, and 0.2 for the first, second, third, fourth rank of PWA, respectively. The number of features with nonzero loadings for the first, second, third and fourth rank of PWA are 50, 40, 30, and 20, respectively. These numbers are consistent with the sizes of components with signals embedded in the simulation procedure indicated by Equation 12. In the middle panel of Figure 5, one can see those feature loadings identified by the first, second, third, and fourth rank of PWA, from top to bottom respectively. The corresponding PW for each rank of PWA is shown in the right panel of Figure 5, from top to bottom accordingly. Figure 5 shows that, in each rank of PWA, the features with nonzero loadings along with their corresponding PW can reveal the underlying signal patterns in the simulated dataset. As shown in Figure 6, we further plotted the explained variances for the first, second, third and fourth rank of PWA, as well as the residues of the data after applying the first four ranks of PWA. This indicates that the top four ranks of PWA is appropriate to characterize the data structure and variance.

For comparison, we sought to use FPCA on this dataset. We used the R-package "fdapace" to implement the FPCA. FPCA cannot be directly applied to a tensor dataset. To overcome the limitation, we arranged the simulated data as

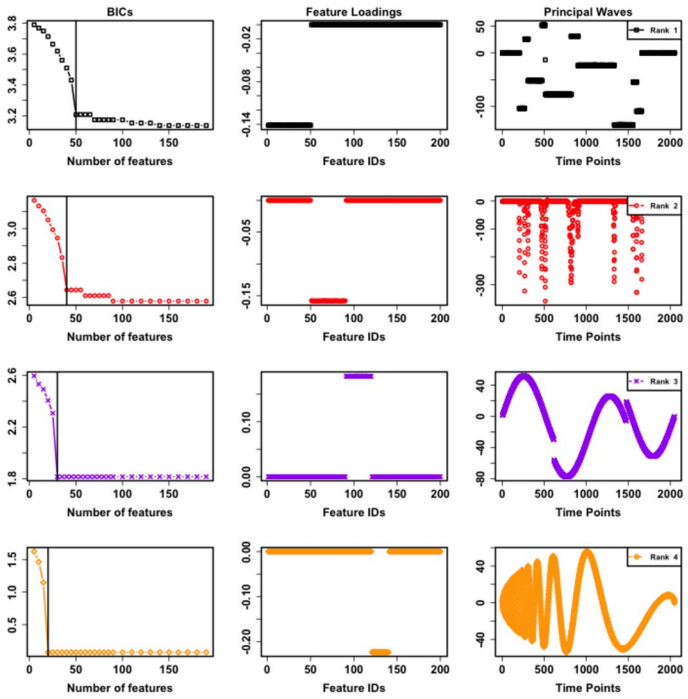

Figure 5. Simulation studies for PWA with multiple ranks. The subfigures in the left panel illustrate BIC curves for the first, second, third and fourth rank of PWA, from top to bottom respectively. The vertical line in each BIC-type model selection subfigure indicates the selected number of nonzero features in the corresponding rank of PWA. The subfigures in the middle panel show the feature loadings identified by the first, second, third and fourth rank of PWA, from top to bottom respectively. The subfigures in the right panel show the identified principal waves accordingly.

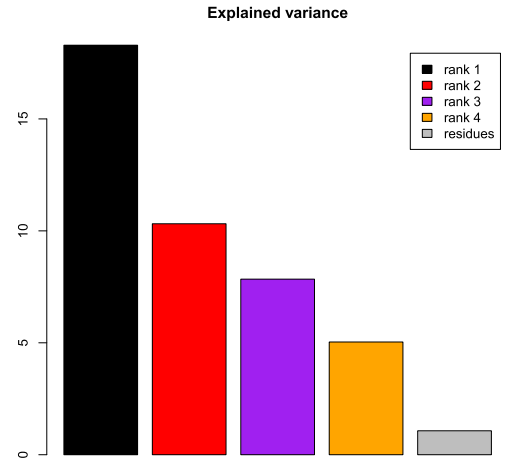

Figure 6. Simulation studies for PWA with multiple ranks. Explained variances.

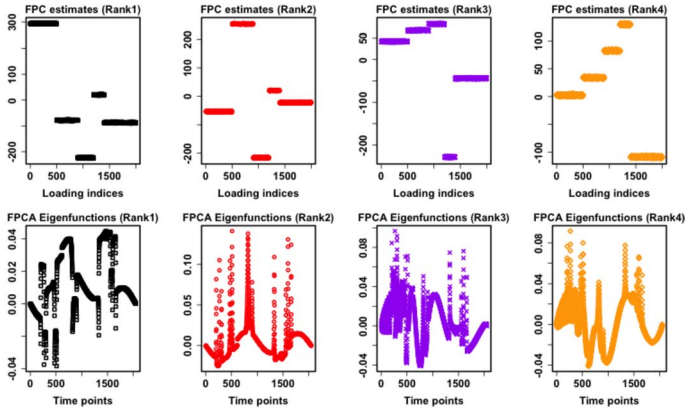

Figure 7. Simulation studies for FPCA with multiple ranks. The top four ranks of loadings (upper panel) and eigenfunctions (lower panel) are shown from left to right.

a $(N \times P) \times T$ big matrix with T columns. We then applied FPCA to this big data matrix. FPCA gives a "data-driven" basis that is constructed from the observed data. Figure 7 shows the results of top four ranks identified by FPCA from left to right, where the feature loadings are shown in the upper panel and the eigenfunctions are shown in the lower panel. Comparing these plots with the PWs identified by PWA, one can see that the patterns identified by PWA are more consistent with the true latent patterns. We also applied PTA to this dataset. As shown in Figure 8, PTA tends to produce smooth principal trends. We also arranged the simulated data as an $(N \times T) \times P$ big matrix, and applied penalized matrix decomposition (PMA) [22], sparse principal component analysis (SPCA) [30] and independent component analysis (ICA) on this dataset. Figure 9, Figure 10 and Figure 11 show the corresponding results, which demonstrate PWA is more appropriate for the analysis.

We further performed simulation studies based on more complicated signals by mixing the above four types of signals. Specifically, we generated x_{npt} based on Equation 13

$$x_{npt} = 0.25 * (f(t) + g(t) + m(t) + l(t)) * b_{1,p} + \epsilon_{npt}, \quad (13)$$

where $n \in \{1, \ldots, N\}$, $p \in \{1, \ldots, P\}$, $t \in \{1, \ldots, T\}$, $b_{1,p} = I(0 < p \leq 50)$, $N = 10, T = 2048$, and $f(t)$, $g(t)$, $m(t)$ and $l(t)$ are based on Equations 5, 7, 9 and 11, respectively. We first set $P = 200$, and noises to be Gaussian distributed with signal-to-noise ratio (SNR) as 7 (case 1). We then increased the number features by setting $P = 2000$, and keep SNR=7 for Gaussian distributed noises (case 2). We further decreased SNR=2 for Gaussian distributed noises, and keep $P = 200$ (case 3). Finally, we incorporated correlated noises into raw data generation using the first order auto-regression model with correlation $corr(\epsilon_{npt}, \epsilon_{np(t+1)}) = 0.5$, and set SNR=7 and $P = 200$ (case 4). Figure 12 shows the simulated raw data from one subject and the corresponding PWA estimates in each case. This set of simulation studies

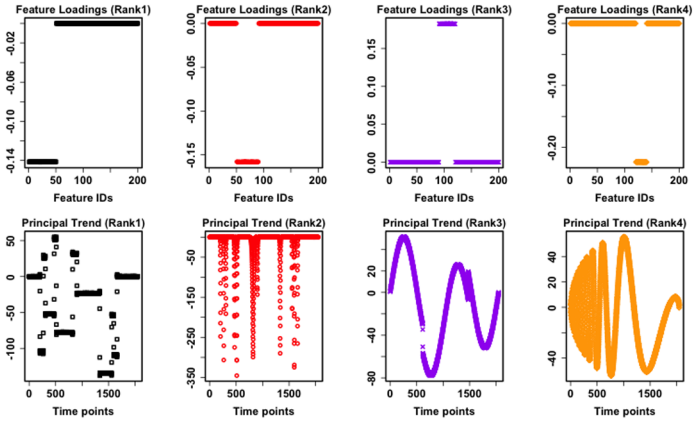

Figure 8. Simulation studies for PTA with multiple ranks. The top four ranks of feature loadings (upper panel) and principal trends (lower panel) are shown from left to right.

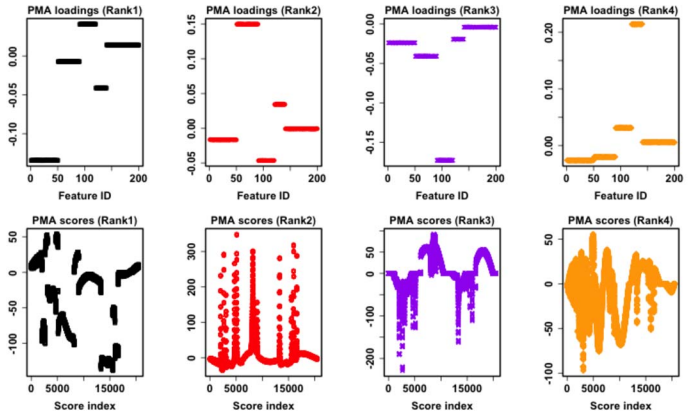

Figure 9. Simulation studies for PMA with multiple ranks. The top four ranks of loadings (upper panel) and scores (lower panel) are shown from left to right.

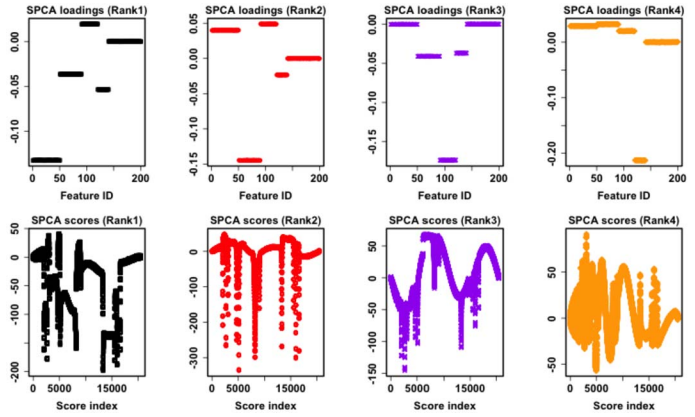

Figure 10. Simulation studies for SPCA with multiple ranks. The top four ranks of loadings (upper panel) and scores (lower panel) are shown from left to right.

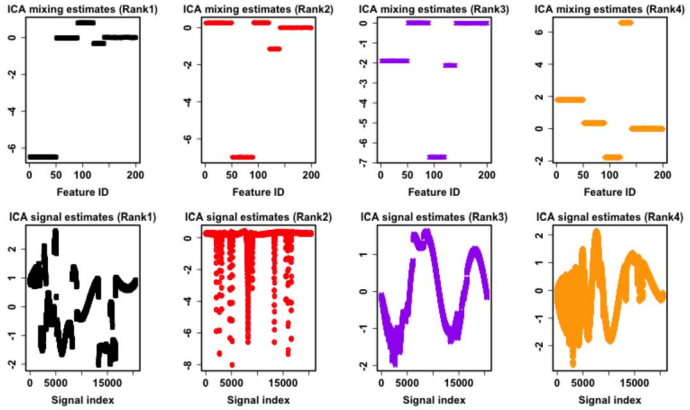

Figure 11. Simulation studies for ICA with multiple ranks. The first four columns of the estimated mixing matrix (upper panel) and the corresponding estimated signals (lower panel) are shown from left to right.

further demonstrated the practical merits of PWA in various scenarios.

4. APPLICATIONS

4.1 Epigenomic regulation in human embryonic stem cells

To demonstrate the practical merits of the proposed method, we first applied it to an epigenomic research scenario. Specifically, we focused on the alterations of chromatin states through histone modifications. The chromatin states along the genome are named as the epigenome. Epigenomes can be tissue and developmental stage specific. Notably, human embryonic stem cells (hESCs) have epigenetic remodeling characteristics, which are critical for their self-renewal and pluripotence. Thus, it is important to investigate epigenetic signatures in hESCs. High-throughput biotechnologies such as chromatin immunoprecipitation sequencing (ChIP-seq) assays can characterize profiles of histone modifications at the genome scale [15]. In our epigenome case study, we employed histone 3 lysine 4 trimethylation (H3K4me3) ChIP-seq data to demonstrate the usefulness of the proposed PWA method. Specifically, we used the ChIP-seq data for H3K4me3 epigenetic modification with two replicates in hESCs. These data were derived from the following resources available in the public domain [18]. We used top 20,000 genes with the highest variance with 3,720 bases around the transcription start site (TSS) in two replicates.

We applied PWA to this dataset and identified 14,000 genes with nonzero loadings. We plot the feature loadings in Figure 13(a) and the estimated data heatmap in Figure 13(b). Each row in the heatmap indicates one gene. In Figure 13(a), genes are ordered according to the amplitudes of feature loadings. Each column in the estimated data

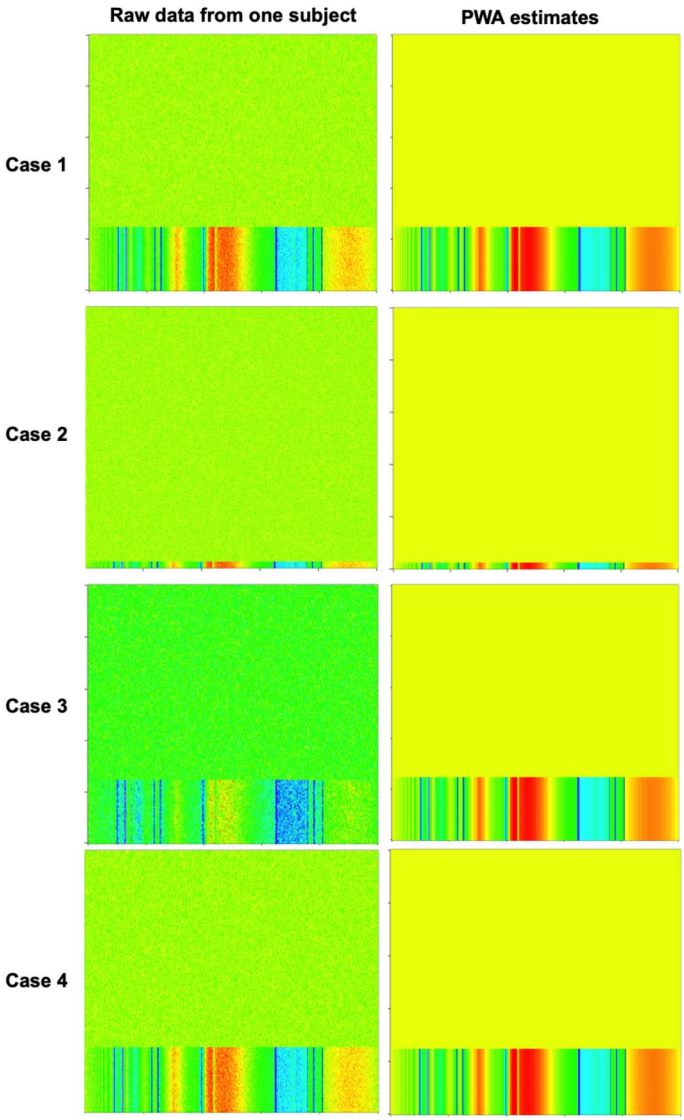

Figure 12. Simulation studies for PWA with more complicated mixed signals based on Equation 13. The subfigures in the left panel illustrate raw data from one subject for each case. The subfigures in the right panel show the corresponding PWA estimates for each case. Case 1: The number of features $P = 200$, and Gaussian distributed noises with signal-to-noise ratio SNR=7. Case 2: $P = 2000$, and Gaussian distributed noises with signal-to-noise ratio SNR=7. Case 3: $P = 200$, and Gaussian distributed noises with signal-to-noise ratio SNR=2. Case 4: $P = 200$, correlated noises using the first order auto-regression model with correlation $corr(\epsilon_{npt}, \epsilon_{np(t+1)}) = 0.5$, and SNR=7.

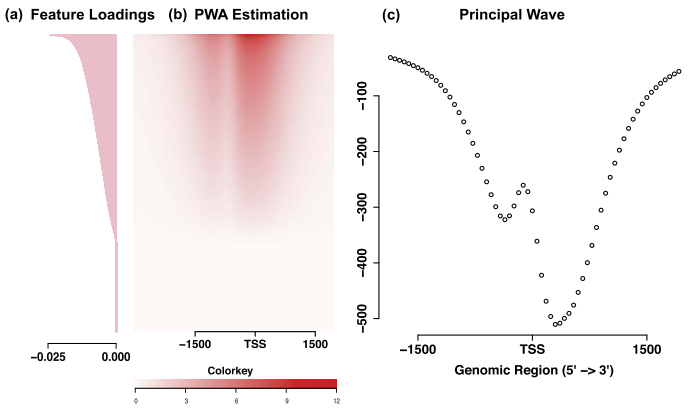

Figure 13. The application of PWA to H3K4me3 ChIP-seq data in human embryonic stem cells. (a) The PWA identified gene loadings. (b) The estimated epigenetic signatures of H3K4me3 in human embryonic stem cells. Each row in the heatmap indicate one gene. Genes are ordered according to the amplitudes of feature loadings in (a). Each column in the estimated data heatmap in Figure 13 indicates one genome locus. TSS: transcription start site. The upstream 1500bp and downstream 1500bp genomic loci are marked in the figures. (c) The principal wave (PW) identified by PWA.

heatmap in Figure 13 indicates one genome locus around the TSS. The estimated PW is shown in Figure 13(c). One can see that a smaller peak of histone modification signals exists before the TSS, and one larger peak occurs behind the TSS. H3K4me3 is a well-known regulatory hallmark which is highly enriched at active promoters near TSS of up-regulated genes [3]. Furthermore, the bimodal property of the identified PW provides insight on the structural features of the genomic region and reveals the existence of promoters for divergent transcriptions. Divergent transcriptions are common in eukaryotic genomes, including those actively transcribed genes regulated by H3K4me3 [5]. The biological signal property and regulatory behaviour can result in co-regulated gene expression. In PWA, the genes with nonzero loadings can have divergent transcriptional properties given the identified bimodal PW. We also further investigated the functions of genes with large loading amplitudes. For instance, gene ID1 has the largest loading amplitude. Existing biological literature reported that ID1 involves epigenomic regulations related to H3K4me3 in hESCs [10]. Applying PWA to the H3K4me3 ChIP-seq data from hESCs, we are able to obtain consistent findings with exiting biological knowledge.

4.2 Neuroimaging data application

We applied the PWA to an alcoholism case study on Electroencephalography (EEG). The data that support the findings of this study are openly available in https://kdd.ics.uci.edu/databases/eeg/eeg.data.html. This dataset contains 45 healthy controls and 77 alcoholics [25]. Single-stimulus EEG signals were recorded for 64 channels and 256 time points. The EEG channel means an electrode placement on a brain scalp. Each electrode placement site has a letter and number label to identify the area of brain. A human brain consists of

the cerebrum, cerebellum, and brainstem. From functional perspectives, cerebrum is responsible for interpreting touch, vision and hearing. It also controls speech, reasoning, emotions, learning and movement. The cerebrum is divided into four lobes, which are frontal, parietal, occipital and temporal. In terms of the functions for these specific subareas of a human brain, frontal lobe has functions related to emotions, behavior, judgment, problem solving, speaking, writing, body movement, concentration and self awareness. Parietal lobe can interpret signals from vision and hearing, as well as sense touch, pain and temperature. Occipital lobe interprets vision such as color, light and movement. Temporal lobe is responsible for understanding language, memory, hearing, sequencing and organization. Cerebellum is located under the cerebrum, and its function is to coordinate muscle movements, maintain posture and balance. Brainstem connects the cerebrum and cerebellum to the spinal cord. It is responsible for many automatic functions such as heart rate, breathing, digestion, wake and sleep cycles.

We performed the proposed PWA method on the EEG data from control and alcoholic groups, respectively. Then, we investigated the denoised EEG signals estimated by the PWA through the formula $\hat{\boldsymbol{a}}\hat{d}\hat{\boldsymbol{\theta}}^{\mathsf{T}}\boldsymbol{W}$. The estimated signals for the control group are shown in Figure 14(a), and those for the alcoholic group are shown in Figure 14(b). In Figure 14, brain areas are shown by the electrode labels. The mapping between the electrode labels and the brain areas are as follows: FP means prefrontal, F means frontal, T means temporal, P means parietal, O means occipital. The electrode label C is reading from central, however there is no central lobe. Based on the placement and individual, the "C" electrodes can exhibit frontal, temporal and parietal-occipital type of activities in their EEG signals. AF means intermediate electrode placement between FP and F, FC means between F and C, FT means between F and T, CP means between C and P, TP means between T and P, PO means between P and O. Odd numbers (1, 3, 5, 7) in the electrode labels refer to electrode placement on the left hemisphere, whereas even numbered electrodes (2, 4, 6, 8) refer to those on the right. A letter "Z" in the electrode labels means zero, referring to an electrode placed on the midline sagittal plane of the skull, as seen in FPZ, FZ, CZ and OZ. By looking at Figure 14(a) and Figure 14(b) simultaneously, one can see that brain activities in the areas of prefrontal, parietal and occipital are more time-varying than those in frontal, temporal and central areas. In addition, the PWA estimated signals tend to be similar to each other for the nearby brain locations. By comparing Figure 14(a) with Figure 14(b), we found that the estimated EEG signals in alcoholic group exhibit differently in the prefrontal, parietal and occipital areas compared to the control group. As aforementioned, these areas have functions related to interpreting signals corresponding to vision, hearing, movement, pain and temperature. These findings based on the PWA estimated signals are consistent with existing studies on alcoholism [1, 14].

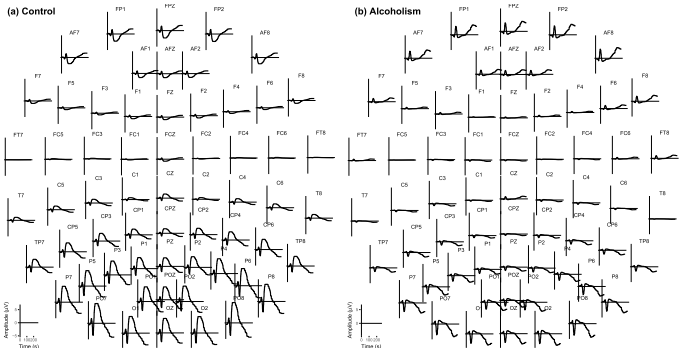

Figure 14. PWA estimated EEG signals across 64 brain areas and 256 time points. (a) PWA estimation for healthy control EEG data. (b) PWA estimation for alcoholism EEG data.

5. CONCLUSION

We have developed a new statistical method, named as PWA, motivated by the problem of extracting underlying patterns and identifying the corresponding features for high-dimensional structured data. The proposed PWA does not assume the embedded patterns are entirely smooth. We demonstrated the practical merits of PWA through simulation studies with different types of signal phenomena embedded in the data, such as jump discontinuities, spikes, varying frequency behaviour, and smooth signals containing jumps/spikes. We also applied PWA to epigenomic research in human embryonic stem cells and an alcoholism case study on EEG data. The real applications demonstrated the proposed PWA method can reveal biologically meaningful underlying signal patterns and identify important biological features.

APPENDIX

A.1 Proof of the triconvex property

The loss function is defined as $L(\boldsymbol{a}, d, \boldsymbol{\theta}|\boldsymbol{X}) = \sum_{n=1}^{N} \|\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\|_F^2$.

The squared error loss can be shown as

$$
\begin{aligned}
& L(\boldsymbol{a}, d, \boldsymbol{\theta}|\boldsymbol{X}) = \sum_{n=1}^{N} \mathrm{tr}(\boldsymbol{E}_n^{\mathsf{T}}\boldsymbol{E}_n) \\
= & \sum_{n=1}^{N} \mathrm{tr}\left[(\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W})^{\mathsf{T}}(\boldsymbol{X}_n - \boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W})\right] \\
= & \sum_{n=1}^{N} \mathrm{tr}\left[\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{X}_n - 2\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W} + \boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] \\
= & \sum_{n=1}^{N} \mathrm{tr}\left[\boldsymbol{X}_n^{\mathsf{T}}\boldsymbol{X}_n\right] - 2\mathrm{tr}\left[\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] \\
& + N \cdot \mathrm{tr}\left[\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right]
\end{aligned}
$$

For fixed $\boldsymbol{a}$ and d, the second and third terms are respectively linear and quadratic in $\boldsymbol{\theta}$, since

$$\begin{aligned}\operatorname{tr}\left[\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] &= \operatorname{tr}\left[\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\right]\\ &= \left(\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\boldsymbol{a}d\right)^{\mathsf{T}}\boldsymbol{\theta},\end{aligned}$$

and

$$\begin{aligned}\operatorname{tr}\left[\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] &= \operatorname{tr}\left[\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\right]\\ &= \operatorname{tr}\left[\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\right]\\ &= \boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d.\end{aligned}$$

Similarly, for fixed $\boldsymbol{\theta}$ and d, the second and third terms are respectively linear and quadratic in $\boldsymbol{a}$, since

$$\begin{aligned}\operatorname{tr}\left[(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}})\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] &= \operatorname{tr}\left[\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\right]\\ &= \left(d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\right)\boldsymbol{a}.\end{aligned}$$

and

$$\begin{aligned}\operatorname{tr}\left[\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\right] &= \operatorname{tr}\left[d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}\right]\\ &= \operatorname{tr}\left[d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}\right]\\ &= d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}.\end{aligned}$$

Convexity follows for each parameter with the others fixed, since $\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}$ and $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}$ are nonnegative. Thus, the original optimization problem with the appropriate convex relaxation $\|\boldsymbol{\theta}\|_2^2 \le 1$ and $\|\boldsymbol{a}\|_2^2 \le 1$ is triconvex. This suggests an iterative algorithm.

A.2 Derivation of the PWA algorithm

With $\boldsymbol{a}$ and d fixed, we minimize the following criterion:

$$N\cdot\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d-2\cdot\left(\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\boldsymbol{a}d\right)^{\mathsf{T}}\boldsymbol{\theta}+2\lambda_1\|\boldsymbol{\theta}\|_1+\lambda_3\|\boldsymbol{\theta}\|_2^2,$$

and we differentiate, set the derivative to 0, and solve for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}=\frac{\operatorname{soft}\left(\frac{1}{N}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n\right)^{\mathsf{T}}\boldsymbol{a}d,\lambda_1\right)}{\left\|\operatorname{soft}\left(\frac{1}{N}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n\right)^{\mathsf{T}}\boldsymbol{a}d,\lambda_1\right)\right\|_2},$$

where $\operatorname{soft}(t,s)$ is a soft-thresholding function.

With $\boldsymbol{\theta}$ and d fixed, we minimize the following criterion:

$$N\cdot\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}d\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}d-2\cdot\left(d\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\right)\boldsymbol{a}+2\lambda_2\|\boldsymbol{a}\|_1+\lambda_4\|\boldsymbol{a}\|_2^2,$$

and we differentiate, set the derivative to 0, and solve for $\boldsymbol{a}$:

$$\boldsymbol{a}=\frac{\operatorname{soft}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_n\boldsymbol{W}\boldsymbol{\theta}d,\lambda_2\right)}{\left\|\operatorname{soft}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{X}_n\boldsymbol{W}\boldsymbol{\theta}d,\lambda_2\right)\right\|_2}.$$

With $\boldsymbol{\theta}$ and $\boldsymbol{a}$ fixed, we minimize the following criterion:

$$-2d\left(\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}\left(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}\right)\right)\boldsymbol{a}+Nd^2\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}.$$

Thus, $d = \frac{1}{N}(\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a})^{-1}(\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}}))\boldsymbol{a} = \frac{1}{N}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{W}(\sum_{n=1}^{N}\boldsymbol{X}_n^{\mathsf{T}})\boldsymbol{a}$

ACKNOWLEDGEMENTS

The author acknowledges the Faculty Research Excellence Program Award of the University of Connecticut. The author would also like to thank anonymous reviewers for their insightful comments and helpful suggestions.

Received 30 October 2019

REFERENCES

[1] Abernathy, K., Chandler, L. J., and Woodward, J. J. (2010). Alcohol and the prefrontal cortex. In *International review of neurobiology* **91**, 289–320. Elsevier.

[2] Bali, J. L., Boente, G., Tyler, D. E., Wang, J.-L., et al. (2011). Robust functional principal components: A projection-pursuit approach. *The Annals of Statistics* **39**(6) 2852–2882. MR3012394

[3] Bibikova, M., Chudin, E., Wu, B., Zhou, L., Garcia, E. W., Liu, Y., Shin, S., Plaia, T. W., Auerbach, J. M., Arking, D. E., et al. (2006). Human embryonic stem cells have a unique epigenetic signature. *Genome research* **16**(9) 1075–1083.

[4] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research* **33**(2) 261–304. MR2086350

[5] Core, L. J., Waterfall, J. J., and Lis, J. T. (2008). Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**(5909) 1845–1848.

[6] Daubechies, I. (1992). *Ten lectures on wavelets*, volume 61. MR1162107

[7] Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**(3) 425–455. MR1311089

[8] Gupta, M. R. and Jacobson, N. P. (2006). Wavelet principal component analysis and its application to hyperspectral images. In *2006 International Conference on Image Processing*, 1585–1588. IEEE.

[9] Haar, A. (1910). Zur theorie der orthogonalen funktionensysteme. *Mathematische Annalen* **69**(3) 331–371. MR1511592

[10] Hammoud, S. S., Nix, D. A., Zhang, H., Purwar, J., Carrell, D. T., and Cairns, B. R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**(7254) 473–478.

[11] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics* **12**(3) 531–547. MR2002634

[12] Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* **41**(2) 495–502. MR0254999

[13] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51**(3) 455–500. MR2535056
[14] Li, Z., Coles, C. D., Lynch, M. E., Ma, X., Peltier, S., and Hu, X. (2008). Occipital-temporal reduction and sustained visual attention deficit in prenatal alcohol exposed adults. *Brain Imaging and Behavior* **2**(1) 39–48.
[15] O'Geen, H., Echipare, L., and Farnham, P. J. (2011). Using chip-seq technology to generate high-resolution profiles of histone modifications. In *Epigenetics Protocols*, 265–286. Springer.
[16] Pollock, D. S. G. et al. (1993). Smoothing with cubic splines.
[17] Ramsay, J. O. and Silverman, B. W. (2007). *Applied functional data analysis: methods and case studies*. Springer. MR2168993
[18] Shen, L., Shao, N., Liu, X., and Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC genomics* **15**(1) 284.
[19] Wahba, G. (1990). *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA. MR1045442
[20] Wang, J.-L., Chiou, J.-M., and Müller, H.-G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application* **3** 257–295.
[21] Wang, Y. (2011). *Smoothing splines: methods and applications*. Chapman and Hall/CRC. MR2814838
[22] Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10**(3) 515. ISSN 1465-4644. MR4172296
[23] Xiong, L., Chen, X., Huang, T.-K., Schneider, J., and Carbonell, J. G. (2010). Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, 211–222. SIAM.
[24] Yao, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statistica Sinica*, pages 965–983. MR2408647
[25] Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin* **38**(6) 531–538.
[26] Zhang, Y. (2018). Lagged principal trend analysis for longitudinal high-dimensional data. *Stat* **7** e213. MR3910688
[27] Zhang, Y. and Davis, R. (2013). Principal trend analysis for time-course data with applications in genomic medicine. *The Annals of Applied Statistics* **7**(4) 2205–2228. MR3161719
[28] Zhang, Y. and Ouyang, Z. (2018). Joint principal trend analysis for longitudinal high-dimensional data. *Biometrics* **74**(2) 430–438. MR3825329
[29] Zhao, X., Barber, S., Taylor, C. C., and Milan, Z. (2018). Classification tree methods for panel data using wavelet-transformed time series. *Computational Statistics & Data Analysis* **127** 204–216. MR3820318
[30] Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics* **15**(2) 265–286. ISSN 1061-8600. MR2252527

Yuping Zhang
Department of Statistics
University of Connecticut
Storrs, CT
USA 06269
E-mail address: yuping.zhang@uconn.edu