Published in final edited form as:
IEEE Trans Big Data. 2018 June ; 4(2): 289–298. doi:10.1109/TBDATA.2017.2653815.

An Enhanced Visualization Method to Aid Behavioral Trajectory Pattern Recognition Infrastructure for Big Longitudinal Data

Hua Fang and
Department of Computer and Information Science, Department of Mathematics, University of Massachusetts Dartmouth, 285 Old Westport Rd, Dartmouth, MA, 02747, and Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, 01605

Zhaoyang Zhang
College of Engineering, University of Massachusetts Dartmouth and Department of Quantitative Health Sciences, University of Massachusetts Medical School

Abstract

Big longitudinal data provide more reliable information for decision making and are common in all kinds of fields. Trajectory pattern recognition is in an urgent need to discover important structures for such data. Developing better and more computationally-efficient visualization tool is crucial to guide this technique. This paper proposes an enhanced projection pursuit (EPP) method to better project and visualize the structures (e.g. clusters) of big high-dimensional (HD) longitudinal data on a lower-dimensional plane. Unlike classic PP methods potentially useful for longitudinal data, EPP is built upon nonlinear mapping algorithms to compute its stress (error) function by balancing the paired weights for between and within structure stress while preserving original structure membership in the high-dimensional space. Specifically, EPP solves an NP hard optimization problem by integrating gradual optimization and non-linear mapping algorithms, and automates the searching of an optimal number of iterations to display a stable structure for varying sample sizes and dimensions. Using publicized UCI and real longitudinal clinical trial datasets as well as simulation, EPP demonstrates its better performance in visualizing big HD longitudinal data.

Index Terms

Enhanced projection pursuit; Pattern recognition; Visualization; Longitudinal data

1 Introduction

Building up the infrastructure for big data visualization is a challenge but an urgent need [1], [2]. Big longitudinal data are generated every day from all kinds of fields in industry, business, government and research institutes [3]–[15]. Discovering useful information from

Correspondence to: Hua Fang.

heterogeneous data requires trajectory pattern recognition techniques [16]–[22]. However, developing visualization tools is crucial to guide this technique, which can facilitate the discovery, presentation and interpretation of important structures buried in complex high-dimensional data. Projection Pursuit (PP) is a classical technique to data visualization, first introduced by Friedman and Tukey in 1974 for exploratory analysis of multivariate data [23]. The basic idea of PP is to design and numerically optimize a projection index function to locate interesting projections from high- to low-dimensional space. From these interesting projections, revealed structures such as clusters could be analyzed [24]–[27]. PP is based on the assumption that redundancy exists in the data and the major characteristics are concentrated into clusters. For example, principle components analysis is one of the typical PP methods, widely used for dimension reduction by removing uninteresting directions of variations [23], [26], [28]–[39] and now often used as an initialization before high dimensional data mapping and clustering [26], [40]–[45].

In the present study, our newly developed PP method is compared to two typical PP methods: Andrews Curves and Grand Tour, as all three methods are potentially useful for big longitudinal data visualization where high dimensionality (HD) and repeated measures for each dimension are common. Section II introduces the involvement of Andrews Curves and Grand Tour; Section III discusses the EPP function and algorithms; Section IV includes the comparison of EPP with other methods using real datasets; Section V evaluates EPP with simulated and artificial data; Section VI concludes this study.

2 Andrews Curves and Grand Tour

Proposed in 1972, Andrews Curve has been widely utilized in many disciplines such as biology, neurology, sociology and semiconductor manufacturing. The algorithm of Andrews Curve was designed to project high dimensional data onto a predefined Fourier series [46], and if any structures exist, they may be visible via Andrews Curves. Briefly, for each case **X** = $\{x_1, x_2, \ldots, x_d\}$, which is a vector of measurements, we define a series ($\frac{1}{\sqrt{2}}$, sin(s), cos(s), sin(2s), cos(2s), …), then the Andrews Curve is calculated as

$$f_x(s) = \frac{x_1}{\sqrt{2}} + x_2 \sin(s) + x_3 \cos(s) + x_4 \sin(2s) + \ldots, \quad (1)$$

for $-\pi < s < \pi$. Each case may be viewed as a curve between $-\pi$ and π, and structures may be viewed as different clusters of curves. Since 1972, several variants of the Andrews Curve have been proposed. Andrews himself also proposed to use different integers to generalize $f_x(s)$,

$$f_x(s) = x_1 \sin(n_1 s) + x_2 \cos(n_1 s) + x_3 \sin(n_2 s) + x_4 \cos(n_2 s) + \ldots . \quad (2)$$

By testing $n_1 = 2$, $n_2 = 4$, $n_3 = 8$, …, the author concluded that Equation (2) is more space filling (ie., a curve whose range contains the entire 2-dimensional unit square, or the mapping is continuous) than Equation (1) but more difficult to interpret when used for visual inspection [46]. A three-dimensional Andrews plot was suggested by Khattree and Naik [47],

$$\sqrt{2}f_x(s) = x_1 + x_2[\sin(s) + \cos(s)] + x_3[\sin(s) - \cos(s)] + x_4[\sin(2s) + \cos(2s)] + \dots . \quad (3)$$

As every projection point is exposed to a sine function and a cosine function, the advantage in Equation (3) is that the trigonometric terms do not simultaneously vanish at any given s, which establishes an interesting relation between the Andrews Curve and the eigenvectors of a symmetric positive definite circular covariance matrix.

Different from Andrews Curve, Grand Tour proposed by Asimov [48] and Buja [49] in 1985 is an interactive visualization technique. The basic idea is to rotate the projected plane from all angles and search the interesting structures [50]–[56]. However, these methods were not ideal in terms of intensive computation, computer storage, and projection recovery turns out to be difficult. Motivated by Andrews Curve, Wegman and Shen [57] suggested an algorithm for computing an approximate two-dimensional grand tour, called pseudo grand tour which means that the tour does not visit all possible orientations of a projection plane. The method has recognized advantages, such as easy calculation, time efficiency in visiting any regions with different plane orientations, and easy recovery of projection. Briefly, assuming d is an even number without loss of generality [57], let $\mathbf{a_1(s)}$ be

$$\sqrt{\frac{2}{d}}(\sin(\lambda_1 s), \cos(\lambda_1 s), \dots, \sin(\lambda_{d/2}s), \cos(\lambda_{d/2}s)), \quad (4)$$

and $\mathbf{a_2(s)}$ be

$$\sqrt{\frac{2}{d}}(\cos(\lambda_1 s), -\sin(\lambda_1 s), \dots, \cos(\lambda_{d/2}s), -\sin(\lambda_{d/2}s)), \quad (5)$$

where λ_i has irrational values. $a_1(s)$ and $a_2(s)$ have the following properties,

$$\|\mathbf{a_1(s)}\|_2^2 = \frac{2}{d}\sum_{j=1}^{d/2}(\sin^2(\lambda_j s) + \cos^2(\lambda_j s)) = 1, \quad (6)$$

$$\|\mathbf{a_2}(\mathbf{s})\|_2^2 = \frac{2}{d}\sum_{j=1}^{d/2}\left(\cos^2(\lambda_j s) + (-\sin)^2(\lambda_j s)\right) = 1,$$

and

$$\langle \mathbf{a_1(s)}, \mathbf{a_2(s)} \rangle = \frac{2}{d} \sum_{j=1}^{d/2} (\sin(\lambda_j s)\cos(\lambda_j s) - \cos(\lambda_j s)\sin(\lambda_j s)) = 0, \quad (7)$$

where $\langle \cdot \rangle$ is the inner product of two vectors $\mathbf{a_1(s)}$ and $\mathbf{a_2(s)}$. Then, the projections of data points on the plane formed by the two basic vectors are

$$f_{x_i}(s) = \left(X'_{i_1}, X'_{i_2}\right), i = 1, 2, \ldots, N, \quad (8)$$

in which

$$X'_{i_1} = \sum_{k=1}^{d} x_k a_{1k}, \quad (9)$$

$$X'_{i_2} = \sum_{k=1}^{d} x_k a_{2k}.$$

According to (6), $\mathbf{a_1(s)}$ and $\mathbf{a_2(s)}$ form an orthonormal basis for a two dimensional plane. Because of the dependence between sin(·) and cos(·), this two-dimensional plane is not quite space filling. However, the algorithm based on (8) is much computationally convenient. By taking the inner product as in (7), a [$\mathbf{a_1(s)}$, $\mathbf{a_2(s)}$] plane is constructed on which the high dimensional data are projected.

Different from Andrews Curve and Pseudo Grand Tour, our new enhanced projection pursuit (EPP) method was built upon Sammon Mapping, assuming not all big longitudinal data fit trigonometric functions or transformation. Sammon mapping has been one of the most successful nonlinear multidimensional scaling methods [58], [59] proposed by Sammon in 1969 [60]. It is highly effective and robust to hyper-spherical and hyper-ellipsoidal clusters [60]. The idea is to minimize the error (called “stress”) between the distances of projected points and the distances of the original data points by moving around projected data points on lower dimensional space (mostly 2-dimenstional place) to best represent those in high-dimensional space. Since its advent, much effort concentrated on improving the optimization algorithm [61]–[65] but rarely on modifying Sammon’s Stress function [64].

Our proposed EPP modified Sammon Stress Function by balancing two weights for between and within cluster errors, respectively, in order to better segment and visualize structures (e.g., clusters) on a projected two-dimensional plane while preserving their cluster membership in high-dimensional space. To this end, we developed a nonlinear algorithm to compute EPP stress. Besides, our EPP was developed to automate the searching and finding of the optimal number of iterations to display a stable structure, for varying sample sizes and dimensions. Our goal is to aid the trajectory pattern recognition of longitudinal data. To evaluate the performance of EPP, one big publicized data set and two real longitudinal random controlled trials (RCT) datasets including a large web-delivered trial data were used to compare EPP with Andrews Curve and Pseudo Grand Tour. Simulated big longitudinal data sets based on RCT data parameters were used to evaluate EPP performance at varying conditions.

3 Enhanced Projection Pursuit (EPP)

In longitudinal data analyses, repeated measures for each dimension result in inevitable high-dimensionality. Built upon Sammon Mapping [60], we proposed an Enhanced Projection Pursuit method (EPP) where the Sammon stress becomes a special case of EPP stress when there is only one cluster and the weights of within and between cluster stresses are equal. EPP is used to aid trajectory pattern recognition for such longitudinal data. The key idea of EPP is to balance the weights of between and within cluster variations in order to achieve better visualization, thus aid pattern recognition for high dimensional (HD) longitudinal data. Table 1 summarizes the notations used hereafter. First, we define our data size and high dimensional space.

Definition 1

let N be the number of cases (e.g., subjects, data points, etc.), $\mathbf{X}_i$, $1 \le i \le N$ be a vector of d variables $\{x_1, x_2, \ldots, x_d\}$, each $\mathbf{X}_i$ be repeatedly measured with t times, then the data has dt dimensional space and the entire data size is $\ell = N\,dt$. e.g, with N cases, $\mathbf{X}_i$ is a dt dimensional vector $\{x_{11}, x_{12}, \ldots, x_{1t}, x_{21}, x_{22}, \ldots, x_{2t}, \ldots, x_{d1}, x_{d2}, \ldots, x_{dt}\}$.

Then, the projection of the big longitudinal data from high-dimensional space onto a two-dimensional plane is defined as follows:

Definition 2

To project big HD longitudinal data onto a two dimensional plane and similar to [60], let the distance between any two vectors of $\mathbf{X}_i$ and $\mathbf{X}_j$ in the dt high dimensional space be defined by D_{ij}^*, $D_{ij}^* = \|X_i - X_j\|_2$, where $\|\cdot\|_2$ is the Euclidean norm.

Based on Definition 1 and 2, randomly choose an initial two-dimensional space for the N vectors of $\mathbf{X}'$ and compute all the two dimensional distances D_{ij}, $1 \le i, j \le N, i \ne j$. The Sammon Stress [60] is calculated as:

$$S_{sam} = \frac{1}{\sum_{i<j} D_{ij}^*} \sum_{i<j} \frac{\left(D_{ij}^* - D_{ij}\right)^2}{D_{ij}^*}. \quad (10)$$

Different from Equation (10), the Stress of EPP stress function S_{EPP} is expressed as the weighted sum of the within-cluster stress S_{EPP_w} and between-cluster stress S_{EPP_b},

$$S_{EPP} = \alpha S_{EPP\_w} + \beta S_{EPP\_b} \quad (11)$$

Algorithm 1(a)

Main EPP Algorithm

Input: longitudinal data $\mathbf{X}_i$, $i = 1, 2, \ldots, N$, cluster labels c_i, $0 \le i \le N$, and a range of stress error bound ε, maximum iteration number, $l_{\max}$, weight change step δ

Output: α, β, f_x and S_{EPP}

```
Initialize $\mathbf{X}'$ by PCA
Set initial values for $S_{EPP_0} \rightarrow \infty$, $l = 0$, $m = 0$, $\alpha_0$ and $\beta_0$ ($\alpha_0, \beta_0 > 0$, $\alpha_0 + \beta_0 = 1$)
for $l = 0$ to $l_{\max}$ do
  $f_{x_l} = \arg\min_{f_x} S_{EPP}(\alpha_l, \beta_l, f_x)$
  $S_{EPP_l} = S_{EPP}(\alpha_{l+1}, \beta_{l+1}, f_{xl})$
  while $\alpha_l, \beta_l > 0$, $\alpha_l + \beta_l = 1$ do
    if $S_{EPP}(\alpha_l + \delta, \beta_l - \delta, f_{xl}) < S_{EPP}(\alpha_l, \beta_l, f_{xl})$ then
      $\alpha_{l+1} = \alpha_{l+1} + \delta$, $\beta_{l+1} = \beta_{l+1} - \delta$
    else
      if $S_{EPP}(\alpha_l - \delta, \beta_l + \delta, f_{xl}) < S_{EPP}(\alpha_l, \beta_l, f_{xl})$ then
        $\alpha_{l+1} = \alpha_{l+1} - \delta$, $\beta_{l+1} = \beta_{l+1} + \delta$
      else
        break
      end if
    end if
  end while
  if $|S_{EPP_l} - S_{EPP_{l-1}}| \le \varepsilon$ then
    break
  end if
end for
```

in which

$$\begin{cases} S_{EPP\_w} = \dfrac{1}{\sum_{i<j} D_{ij}^*} \displaystyle\sum_{i<j, c_i = c_j} \dfrac{\left(D_{ij}^* - D_{w_{ij}}\right)^2}{D_{ij}^*} \\ S_{EPP\_b} = \dfrac{1}{\sum_{i<j} D_{ij}^*} \displaystyle\sum_{i<j, c_i \ne c_j} \dfrac{\left(D_{ij}^* - D_{b_{ij}}\right)^2}{D_{ij}^*} \end{cases} \quad (12)$$

where $\frac{1}{\sum_{i<j} D^*_{\boldsymbol{ij}}}$ is a constant for a given big HD longitudinal data, $\sum_{i<j, c_i = c_j} \frac{\left(D^*_{ij} - D_{w_{ij}}\right)^2}{D^*_{ij}}$ and $\sum_{i<j, c_i \neq c_j} \frac{\left(D^*_{ij} - D_{b_{ij}}\right)^2}{D^*_{ij}}$ are the within-cluster and between-cluster stress, respectively, $D_{w_{ij}}$ is the within cluster Euclidean distance between case i and j if they are in the same cluster, and $D_{b_{ij}}$ is the between cluster Euclidean distance between case i and j if they belong to different clusters; α and β are the weights of the within-cluster stress and between-cluster stress, respectively, $a, \beta > 0$ and $\alpha + \beta = 1$. Note again that the Sammon stress is a special case of EPP stress when there is only one cluster, $c_i = 1$, $i = 1, 2, \ldots, N$ and the weights of within cluster and between cluster stresses are equal, $\alpha = \beta$.

EPP algorithm aims to obtain an interesting two-dimensional projection of the original high dimensional data that minimizes its stress function. The optimization problem is expressed as

$$\begin{aligned} &\text{minimize} \;\; \alpha S_{EPP\_w} + \beta S_{EPP\_b} \\ &\text{subject to} \;\; \alpha, \beta > 0, \alpha + \beta = 1\,. \end{aligned} \quad (13)$$

Definition 3

To minimize $S_{EPP}(\alpha, \beta, f_x)$ where f_x stands for the projections of $D_{w_{ij}}$ and $D_{b_{ij}}$, the gradual approximation algorithm works as: Given a fixed pair of α and β, update the values of f_x where S_{EPP} has the minimum value, that is, keep updating α and β until there are no changes according to (12).

$$\begin{cases} \alpha = \alpha + \delta, \beta = \beta - \delta, \text{if } S_{EPP}(\alpha + \delta, \beta - \delta, f_x) < S_{EPP} \\ \alpha = \alpha - \delta, \beta = \beta + \delta, \text{if } S_{EPP}(\alpha - \delta, \beta + \delta, f_x) < S_{EPP} \\ \alpha = \alpha, \beta = \beta, \text{otherwise} \end{cases} \quad (14)$$

The main EPP algorithm is shown in Algorithm 1(a). The embedded gradual approximation algorithm is displayed in Algorithm 1(b) to minimize S_{EPP} given α and β; the values of f_x were retained when S_{EPP} has the minimum value. Specifically, the EPP algorithm initialize $\mathbf{X}'$ based on the results from PCA; update f_x according to Algorithm 1(b) based on Equation (15), calculate the EPP stress and update α and β, with a weight change step δ based on Equation (14). If the difference between two consecutive stress values is less than the threshold ε, the algorithm stops. Repeat this process until reaching the maximum iteration number, $l_{\max}$.

$$f_{x_l} = \underset{f_x}{\arg\min}\, S_{EPP}(\alpha_l, \beta_l, f_x)\,. \quad (15)$$

Algorithm 1(b)

Algorithm for Updating f_X

```
Input: Projections $\mathbf{X}'$, $\alpha$ and $\beta$, error bound $\varepsilon$, maximum iteration number $m_{\max}$, $S_{EPP}^{(0)} \rightarrow \infty$
Output: $S_{EPP}^{(m+1)}$ and $f_X^{(m+1)}$
for $m = 0$ to $m_{\max}$ do
  $f_X^{(m+1)} = f_X^{(m)} - \tau \cdot \Delta^{(m)}$
  $S_{EPP}^{(m+1)} = S_{EPP}(\alpha, \beta, f_X^{(m+1)})$
  if $|S_{EPP}^{(m+1)} - S_{EPP}^{(m)}| \leq \varepsilon$ then
    break
  end if
end for
```

Note that in Algorithm 1(b) when updating f_X, $f_X^{(m)}$ are the projections of the data on the two-dimensional space at the m-th iteration, τ is the iteration step size which is set at 0.3 or 0.4 according to [60], $\Delta^{(m)} = \frac{\partial S_{EPP}^{(m)}}{\partial f_X^{(m)}} / \left| \frac{\partial^2 S_{EPP}^{(m)}}{\partial (f_X^{(m)})^2} \right|$ and $w = \frac{-2}{\sum_{i<j} D_{ij}^*}$ is a constant. Then the first-order derivative with respect to f_X is shown in Equation (16) and the second-order derivative is expressed in Equation (17).

Unlike nonlinear mapping algorithm [60], the EPP algorithm further automates the searching and finds the optimal number of iterations to display a stable structure by learning the change of S_{EPP} in two consecutive iterations at a range of varying error bounds, sample size and the number of dimensions.

4 EPP Performance in Case Studies

Our EPP method was tested on 3 real datasets, including one publicized [66] and two random controlled trial (RCT) datasets [43], [67]–[69]. These data features are summarized in Table 2.

The Waveform data were generated by a clustering data generator described in [70] and published by [66], [70]. It consists of 5000 cases, each with 21 attributes ($\ell = 105,000$). There are 3 clusters of waves identified for testing algorithms. Figure 1 shows the performance of the three PP methods for waveform datasets. Clearly, Andrews Curve and grand tour were unable to visualize the three classes while the EPP demonstrated its projection power in visualizing the 3-cluster structure.

TDTA data were collected from a longitudinal culturally-tailored smoking cessation intervention for 109 Asian American smokers ($\ell = 2,180$). It contains three identified culturally-adaptive response patterns [43]. This intervention used three components: Cognitive behavioral therapy, cultural tailoring, and nicotine replacement therapy. The first two were measured by scores on Perceived Risks and Benefits, Family and Peer Norms, and Self-efficacy scales. Each scale has four repeated measures, total 20 attributes, of which only

Perceived Benefits and Family Norms were used using our multiple imputation based fuzzy clustering method discussed elsewhere [71]–[73]. As shown in Figure 2, two of the three clusters projected by Andrews Curve was completely overlapped, while Grand Tour seems to perform as good as EPP for this longitudinal dataset. The parameters of TDTA data are shown in Table 3 and Table 4.

QuitPrimo dataset includes 1320 cases ($\ell = 23,760$) with missing values about 8.4%. This study aims to evaluate an integrated informatics solution to increase access to web-delivered smoking cessation support. The data is collected via an online referral portal about three components: 1) My Mail, 2) Online Community, 3) Our Advice. Each of the first three component has 6 monthly values measured during 6 months. Figure 3 again showcases the strength of EPP over the other two methods for this big longitudinal dataset. Projected four patterns were overlapped using Andrews Curve while and the blue and green patterns were overlapped to a noticeable degree using the Grand Tour. Table 5 and 6 show the mean values and standard deviations of QuitPrimo dataset, respectively.

$$\frac{\partial S_{EPP}{}^{(m)}}{\partial f_x{}^{(m)}} = \begin{cases} w \sum_{j=1, j \neq p}^{N} \left[\frac{D_j^* - D_{w_j}}{D_j^* D_{w_j}} \right] \left(f_x{}^{(m)} - X_j{}'^{(m)} \right) \text{if } c_p = c_j, \\ w \sum_{j=1, j \neq p}^{N} \left[\frac{D_j^* - D_{b_j}}{D_j^* D_{b_j}} \right] \left(f_x{}^{(m)} - X_j{}'^{(m)} \right) \text{if } c_p \neq c_j. \end{cases} \tag{16}$$

$$\frac{\partial^2 S_{EPP}{}^{(m)}}{\partial \left(f_x{}^{(m)} \right)^2} = \tag{17}$$

$$\begin{cases} w \sum_{j=1, j \neq p}^{N} \frac{1}{D_j^* D_{w_j}} \left[(D_j^* - D_{w_j}) - \frac{(f_x{}^{(m)} - X_j{}'^{(m)})^2}{D_{w_j}} \left(1 + \frac{D_j^* - D_{w_j}}{D_{w_j}} \right) \right] \text{if } c_p = c_j, \\ w \sum_{j=1, j \neq p}^{N} \frac{1}{D_j^* D_{b_j}} \left[(D_j^* - D_{b_j}) - \frac{(f_x{}^{(m)} - X_j{}'^{(m)})^2}{D_{b_j}} \left(1 + \frac{D_j^* - D_{b_j}}{D_{b_j}} \right) \right] \text{if } c_p \neq c_j. \end{cases}$$

The optimal pairs, α and β, for included real longitudinal datasets TDTA and QuitPrimo given f_x can be detected by the following steps. Initialize a pair of values, e.g., (0.5,0.5), and calculate the stress of the proposed EPP method by Equation (10) and (11). Increase α and decrease β, or vice versa, by a boundary parameter δ, e.g., $\delta = 0.1$, to obtain a new stress value. Updating α and β until the stress values no longer decease, we can obtain the optimal weights α and β for the within and between cluster stresses. As shown in Figure 4(a) and

Figure 4(b), the optimal weights of (0.8, 0.2) were founded for TDTA and QuitPrimo data, respectively.

5 EPP Performance Using Simulated Longitudinal Data

The proposed EPP was also evaluated using simulated data. First, simulated longitudinal data were generated using parameters from the two real datasets, TDTA and QuitPrimo. The data generation procedure is described as follows:

1. Fit the multivariate normal distribution to TDTA and the zero-inflated Poisson mixture distribution to the QuitPrimo web trial data [71], respectively, and learn the parameters such as cluster mean vectors and standard deviations, the results are shown in Table 3, 4, 5 and 6;
2. Set the number of cases of each cluster according to the proportion of each cluster (Table 7);
3. Generate data for each cluster based on the model parameters from (1) and cluster size (2).
4. Randomize data from (3) to generate a complete dataset;
5. Repeat (1–4) and generate datasets with varying sample sizes, N is in {100, 200, 300, 500, 1000, 5000}, $d_{TDTA} = 20$, $d_{QuitPrimo} = 18$, and $\ell_{TDTA} = \{2000, 4000, 6000, 10000, 20000, 100000\}$, $\ell_{QuitPrimo} = \{1800, 3600, 4800, 9000, 18000, 36000\}$.

Figure 5 displays the EPP projection based on the TDTA parameters using different sample sizes. From $N = 100$ to $N = 5000$, the clusters are clearly projected. With smaller sample sizes, the data points are more spread within the cluster. The red and green clusters are closer to each other compared to the blue cluster.

Based on the QuitPrimo parameters, EPP again clearly projected the four clusters across a range of data size ℓ. The blue cluster is always far apart from the red cluster; the other three clusters always touch each other as shown in Figure 6.

Using the same simulated data sets, the optimal number of iterations were tested for the proposed EPP method using a different number of sample sizes or dimensions. In Figure 7 (a), the number of dimensions was fixed at 20, and the data sizes ℓ were varied from 2,000 to 100,000. In Figure 7 (b), the data sizes ℓ was fixed at 100,000, and the number of dimensions d were varied from 2 to 100. For all conditions, the change between iterations (ε) was varied from 10^{-3} to 10^{-6}.

The findings indicate that across different sample sizes or dimensions or the change of stresses between iterations (ε), the optimal number of iteration seem to be always below 350.

Furthermore, using the same data generation procedure, an artificial longitudinal dataset was generated with standardized mean and variance-covariance matrices to evaluate the EPP performance. The mean vector was set as 0.2, 0.5, and 0.8 for three clusters [74], [75], the correlation matrix (standardized variance-covariance matrix) was set with 1 at the diagonal and other matrix elements were randomly selected from {0.1, 0.3, 0.5} [74], [75]. The data size was varied from 1,000 to 500,000 and dimensions were changed from 10 to 100. The different colored planes stand for the four settings for the change of stresses between iterations (ε), 10^{-3}, 10^{-4}, 10^{-5}, and 10^{-6}. As shown in Figure 8, the optimal number of iterations seem to be always below 500 across different sample sizes, dimensions and error bounds (ε) for the change between iterations. Using 500 iterations could be an empirical rule for setting the iterations for EPP. Overall, in terms of computational time, EPP cost 11 and 22 seconds for projecting real TDTA and QuitPrimo data while up to 9 minutes assuming the worst scenario of N = 20,000 and dt = 100.

6 Conclusion

Pattern visualization is a challenging field. A robust projection pursuit method could enormously ease pattern recognition. Our enhanced projection pursuit (EPP), a variant of classic Sammon Mapping, balances the weights of between and within cluster variations and better project big high dimensional longitudinal data onto two-dimensional plane using nonlinear mapping algorithms. Compared to classical Andrews Curve and Grand Tour, our EPP method seems to perform consistently well and was more robust to such data. Different from the two methods, EPP was not built upon trigonometric functions as not all longitudinal datasets follow this assumption, especially those longitudinal random controlled trial (RCT) or observational data [40]–[45], [67], [74], [76]. Using the publicized UCI dataset, real longitudinal RCT datasets and a number of simulated big longitudinal data, EPP showcases its clear and better projection power with respect to high-dimensionality, sample sizes and error bounds for the change between iterations with satisfactory computational costs. Embedding EPP into different trajectory pattern recognition systems and further reducing computational time for bigger data would be future tasks. Testing EPP on more big longitudinal data could further warrant its robustness.

Acknowledgments

This project was supported by National Institute of Health (NIH) grants 1R01DA033323-01, and NIH National Center for Advancing Translational Sciences 5UL1TR000161-04 pilot study award to Dr. Fang.

References

1. Fang H, Zhang Z, Wang CJ, Daneshmand M, Wang C, Wang H. A survey of big data research. IEEE network. 2015; 29(5):6. [PubMed: 26504265]

2. Fox P, Hendler J. Changing the equation on scientific data visualization. Science(Washington). 2011; 331(6018):705–708. [PubMed: 21311008]

3. Kumagai M, Kim J, Itoh R, Itoh T. Tasuke: a web-based visualization program for large-scale resequencing data. Bioinformatics. 2013; 29(14):1806–1808. [PubMed: 23749962]

4. Keller, M., Beutel, J., Saukh, O., Thiele, L. Local Computer Networks Workshops (LCN Workshops), 2012 IEEE 37th Conference on. IEEE; 2012. Visualizing large sensor network data sets in space and time with vizzly; p. 925-933.

5. Pires DE, de Melo-Minardi RC, da Silveira CH, Campos FF, Meira W. acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics. 2013; 29(7):855–861. [PubMed: 23396119]

6. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics. 2008; 92(5):255–264. [PubMed: 18703132]

7. Bartel DP. Micrornas: genomics, biogenesis, mechanism, and function. cell. 2004; 116(2):281–297. [PubMed: 14744438]

8. Lynch C. Big data: How do your data grow? Nature. 2008; 455(7209):28–29. [PubMed: 18769419]

9. Brinkmann BH, Bower MR, Stengel KA, Worrell GA, Stead M. Large-scale electrophysiology: acquisition, compression, encryption, and storage of big data. Journal of neuroscience methods. 2009; 180(1):185–192. [PubMed: 19427545]

10. Frankel F, Reid R. Big data: Distilling meaning from data. Nature. 2008; 455(7209):30–30.

11. Waldrop M. Big data: wikiomics. Nature News. 2008; 455(7209):22–25.

12. McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D. Big data. The management revolution. Harvard Bus Rev. 2012; 90(10):61–67.

13. Chang F, Dean J, Ghemawat S, Hsieh WC, Wallach DA, Burrows M, Chandra T, Fikes A, Gruber RE. Bigtable: A distributed storage system for structured data. ACM Transactions on Computer Systems (TOCS). 2008; 26(2):4.

14. Tan W, Blake MB, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Computing. 2013; (5):62–69.

15. Tracey, D., Sreenan, C. Cluster, Cloud and Grid Computing (CCGrid), 2013 13th IEEE/ACM International Symposium on. IEEE; 2013. A holistic architecture for the internet of things, sensing services and big data; p. 546-553.

16. Fang, H., Wang, H., Wang, C., Daneshmand, M. Big Data (Big Data), 2015 IEEE International Conference on. IEEE; 2015. Using probabilistic approach to joint clustering and statistical inference: Analytics for big investment data; p. 2916-2918.

17. Zhang, Z., Fang, H., Wang, H. Visualization aided engagement pattern validation for big longitudinal web behavior intervention data; IEEE 17th international Conference on E-health Networking, Application & Services; 2015.

18. Fang H, Johnson C, Stopp C, Espy KA. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicology and teratology. 2011; 33(1):155–165. [PubMed: 21256430]

19. Fang H, Dukic V, Pickett KE, Wakschlag L, Espy KA. Detecting graded exposure effects: A report on an east boston pregnancy cohort. Nicotine & Tobacco Research. 2012:ntr272.

20. Zhang Z, Fang H. Multiple imputation based clustering validation (miv) for big longitudinal trial data with missing values in ehealth. Journal of Medical System. 2016

21. Fang H, Espy KA, Rizzo ML, Stopp C, Wiebe SA, Stroup WW. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. International journal of information technology & decision making. 2009; 8(03):491–513. [PubMed: 20336179]

22. Espy KA, Fang H, Charak D, Minich N, Taylor HG. Growth mixture modeling of academic achievement in children of varying birth weight risk. Neuropsychology. 2009; 23(4):460. [PubMed: 19586210]

23. Friedman J, Tukey J. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers. Sep; 1974 C-23(9):881–890.

24. Kruskal, JB. Statistical Computation. Academic Press; New York: 1969. Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new index of condensation; p. 427-440.

25. R-M E, Goulermas JY, Mu T, Ralph JF. Automatic induction of projection pursuit indices. Neural Networks, IEEE Transactions on. 2010; 21(8):1281–1295.

26. Jones MC, Sibson R. What is projection pursuit? Journal of the Royal Statistical Society. Series A (General). 1987:1–37.

27. Eslava G, Marriott FHC. Some criteria for projection pursuit. Statistics and Computing. 1994; 4(1): 13–20.

28. Dauxois J, Pousse A, Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. Journal of multivariate analysis. 1982; 12(1):136–154.

29. Rice JA, Silverman BW. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society. Series B (Methodological). 1991:233–243.
30. Pezzulli S, Silverman B. Some properties of smoothed principal components analysis for functional data. Computational Statistics. 1993; 8:1–1.
31. Silverman BW, et al. Smoothed functional principal components analysis by choice of norm. The Annals of Statistics. 1996; 24(1):1–24.
32. Boente G, Fraiman R. Kernel-based functional principal components. Statistics & probability letters. 2000; 48(4):335–345.
33. Ramsay, JO. Functional data analysis. Wiley Online Library; 2006.
34. Hall P, H-N M. On properties of functional principal components analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):109–126.
35. Yao F, Lee T. Penalized spline models for functional principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(1):3–25.
36. Gervini D. Free-knot spline smoothing for functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(4):671–687.
37. Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K, Boente G, Fraiman R, Brumback B, Croux C, et al. Robust principal component analysis for functional data. Test. 1999; 8(1):1–73.
38. Hyndman RJ, Ullah S, et al. Robust forecasting of mortality and fertility rates: a functional data approach. Computational Statistics & Data Analysis. 2007; 51(10):4942–4956.
39. Gervini D. Robust functional estimation using the median and spherical principal components. Biometrika. 2008; 95(3):587–600.
40. Fang H, Dukic V, Pickett KE, Wakschlag L, Espy KA. Detecting graded exposure effects: A report on an east boston pregnancy cohort. Nicotine & Tobacco Research. 2012:ntr272.
41. Fang H, Johnson C, Stopp C, Espy KA. A new look at quantifying tobacco exposure during pregnancy using fuzzy clustering. Neurotoxicology and teratology. 2011; 33(1):155–165. [PubMed: 21256430]
42. Fang H, Rizzo ML, Wang H, Espy KA, Wang Z. A new nonlinear classifier with a penalized signed fuzzy measure using effective genetic algorithm. Pattern recognition. 2010; 43(4):1393–1401. [PubMed: 20300543]
43. Fang H, DiFranza S, Zhang Z, Ziedonis D, Allison J. Pattern recognition approach to culturally-tailored behavioral interventions for smoking cessation: Dose and timing. Society for Research on Nicotine and Tobacco. 2014
44. Fang H, Espy KA, Rizzo ML, Stopp C, Wiebe SA, Stroup WW. Pattern recognition of longitudinal trial data with nonignorable missingness: An empirical case study. International journal of information technology & decision making. 2009; 8(03):491–513. [PubMed: 20336179]
45. Fang H, Allison J, Barton B, Zhang Z, Olendzki G, Ma Y. Pattern recognition approach for behavioral interventions: An application to a dietary trial. Society of Behavioral Medicine. Ann Behav Med. 2014
46. Andrews DF. Plots of high-dimensional data. Biometrics. 1972:125–136.
47. Khattree R, Naik DN. Andrews plots for multivariate data: some new suggestions and applications. Journal of statistical planning and inference. 2002; 100(2):411–425.
48. Asimov D. The grand tour: a tool for viewing multidimensional data. SIAM Journal on Scientific and Statistical Computing. 1985; 6(1):128–143.
49. Buja A, Asimov D. Grand tour methods: an outline. Computing Science and Statistics. 1986; 17:63–67.
50. Buja A, Hurley C, Mcdonald J. A data viewer for multivariate data. Colorado State Univ, Computer Science and Statistics. Proceedings of the 18 th Symposium on the Interface p 171–174(SEE N 89-13901 05-60). 1987
51. Cook D, Buja A, Cabrera J. Direction and motion control in the grand tour. Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. 1991:180–183.
52. Cook D, Buja A, Hurley C. Grand tour and projection pursuit(a video). ASA Statistical Graphics Video Lending Library. 1993

53. Cook D, Buja A, Cabrera J, Hurley C. Grand tour and projection pursuit. Journal of Computational and Graphical Statistics. 1995; 4(3):155–172.

54. Cook D, Buja A. Manual controls for high-dimensional data projections. Journal of computational and Graphical Statistics. 1997; 6(4):464–480.

55. Furnas GW, Buja A. Prosection views: Dimensional inference through sections and projections. Journal of Computational and Graphical Statistics. 1994; 3(4):323–353.

56. Hurley C, Buja A. Analyzing high-dimensional data with motion graphics. SIAM Journal on Scientific and Statistical Computing. 1990; 11(6):1193–1211.

57. Wegman E, Shen J. Three-dimensional andrews plots and the grand tour. Computing Science and Statistics. 1993:284–284.

58. Dayanik A. Feature interval learning algorithms for classification. Knowledge-Based Systems. 2010; 23(5):402–417.

59. Hu J, Deng W, Guo J, Xu W. Learning a locality discriminating projection for classification. Knowledge-Based Systems. 2009; 22(8):562–568.

60. Sammon JW. A nonlinear mapping for data structure analysis. IEEE Transactions on computers. 1969; 18(5):401–409.

61. Mao J, Jain AK. Artificial neural networks for feature extraction and multivariate data projection. Neural Networks, IEEE Transactions on. 1995; 6(2):296–317.

62. Lee RCT, Slagle JR, Blum H. A triangulation method for the sequential mapping of points from n-space to two-space. Computers, IEEE Transactions on. 1977; 100(3):288–292.

63. Pekalska E, de Ridder D, Duin RP, Kraaijveld MA. A new method of generalizing sammon mapping with application to algorithm speed-up. ASCI. 1999; 99:221–228.

64. Tenenbaum JB, De Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000; 290(5500):2319–2323. [PubMed: 11125149]

65. Yang, L. Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. Vol. 2. IEEE; 2004. Sammon's nonlinear mapping using geodesic distances; p. 303-306.

66. Asuncion A, Newman D. Uci machine learning repository. 2007

67. Kim SS, Kim S-H, Fang H, Kwon S, Shelley D, Ziedonis D. A culturally adapted smoking cessation intervention for korean americans: A mediating effect of perceived family norm toward quitting. Journal of Immigrant and Minority Health. 2014:1–10. [PubMed: 23054547]

68. Houston TK, Sadasivam RS, Ford DE, Richman J, Ray MN, Allison JJ. The quit-primo provider-patient internet-delivered smoking cessation referral intervention: a cluster-randomized comparative effectiveness trial: study protocol. Implement Sci. 2010; 5:87. [PubMed: 21080972]

69. Houston TK, Sadasivam RS, Allison JJ, Ash AS, Ray MN, English TM, Hogan TP, Ford DE. Evaluating the quit-primo clinical practice eportal to increase smoker engagement with online cessation interventions: a national hybrid type 2 implementation study. Implementation Science. 2015; 10(1):154. [PubMed: 26525410]

70. Breiman, L., Friedman, J., Stone, CJ., Olshen, RA. Classification and regression trees. CRC press; 1984.

71. Zhang, Z., Fang, H. Multiple-vs non-or single-imputation based fuzzy clustering for incomplete longitudinal behavioral intervention data; 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE); Jun, 2016 p. 219-228.

72. Zhang Z, Fang H, Wang H. Multiple imputation based clustering validation (miv) for big longitudinal trial data with missing values in ehealth. J. Med. Syst. Jun.2016 40(6):1–9. [Online]. Available: http://dx.doi.org/10.1007/s10916-016-0499-0. [PubMed: 26573639]

73. Zhang Z, Fang H, Wang H. A new mi-based visualization aided validation index for mining big longitudinal web trial data. IEEE Access. 2016; 4:2272–2280. [PubMed: 27482473]

74. Fang H, Brooks GP, Rizzo ML, Espy KA, Barcikowski RS. Power of models in longitudinal study: Findings from a full-crossed simulation design. The Journal of Experimental Education. 2009; 77(3):215–254. [PubMed: 19946462]

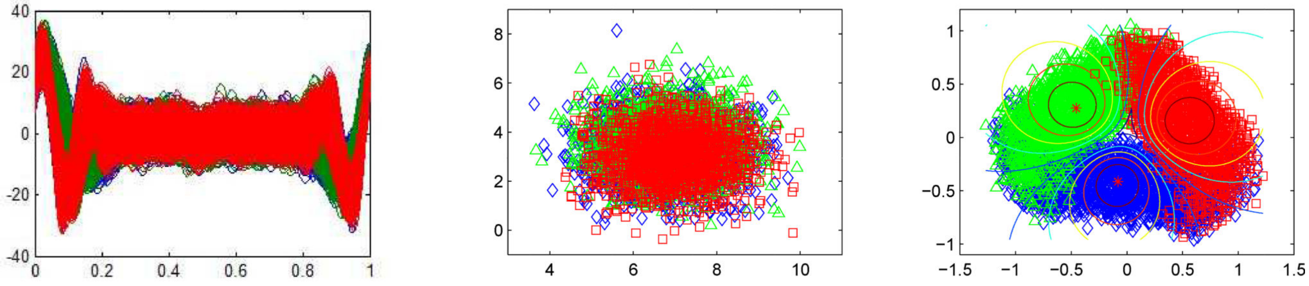
75. Fang, H. Proceedings of the Thirty-First Annual SAS Users Group Conference. SAS Institute Inc.; Cary, NC: 2006. hlmdata and hlmrmpower: Traditional repeated measures vs. hlm for multilevel longitudinal data analysis-power and type i error rate comparison.

76. Ma Y, Olendzki B, Wang J, Persuitte G, Li W, Fang H, Merriam P, Wedick N, Ockene I, Culver A, Schneider K, Olendzki G, Zhang Z, Ge T, Carmody J, Pagoto S. Randomized trial of single- versus multi-component 1 dietary goals on weight loss and diet quality in individuals with metabolic syndrome. Annals of Internal Medicine. 2014

Biographies

Hua Fang is an Associate Professor in the Department of Computer and Information Science, Department of Mathematics, University of Massachusetts Dartmouth, 285 Old Westport Rd, Dartmouth, MA, 02747; Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, 01605; Division of Biostatistics and Health Services Research, Department of Quantitative Health Sciences, University of Massachusetts Medical School. Dr. Fang's research interests include computational statistics, research design, statistical modeling and analyses in clinical and translational research. She is interested in developing novel methods and applying emerging robust techniques to enable and improve the studies that can have broad impact on the treatment or prevention of human disease.

Zhaoyang Zhang received the B.S. degree in science and the M. S. degree in electrical engineering from Xidian University, Xian, China, in 2007 and 2010, respectively. He is currently pursuing his Ph.D. degree in the College of Engineering, University of Massachusetts, Dartmouth, MA, USA. His current research interests include wireless healthcare, wireless body area networks, big data and cyber-physical systems.



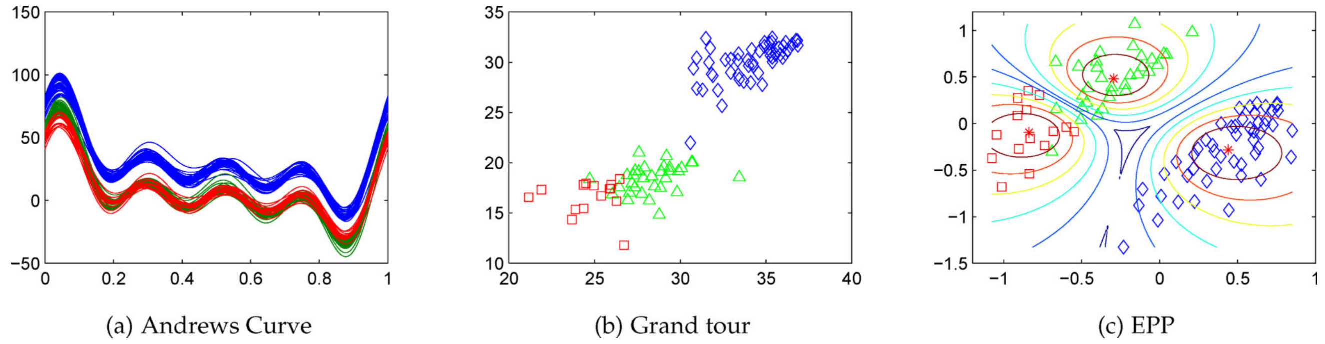
(a) Andrews Curve

(b) Grand tour

(c) EPP

Fig. 1.

Projection Pursuit of Waveform data using Andrews Curve, grand tour and proposed EPP



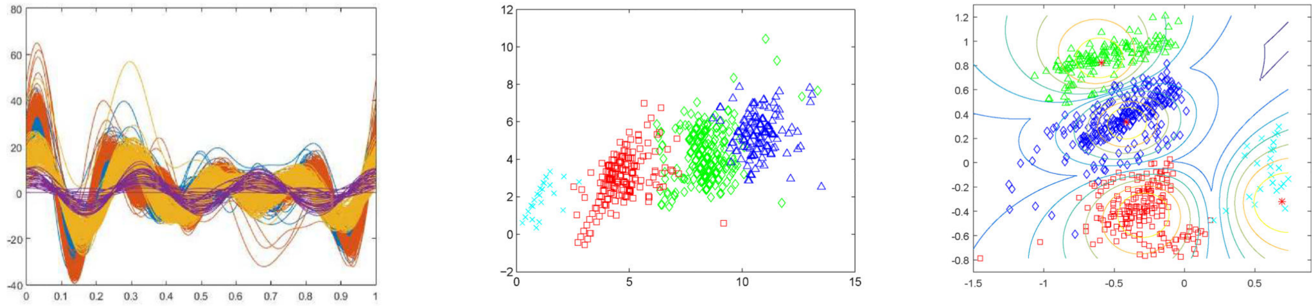
(a) Andrews Curve

(b) Grand tour

(c) EPP

Fig. 2.

Projection Pursuit of TDTA data using Andrews Curve, grand tour and proposed EPP



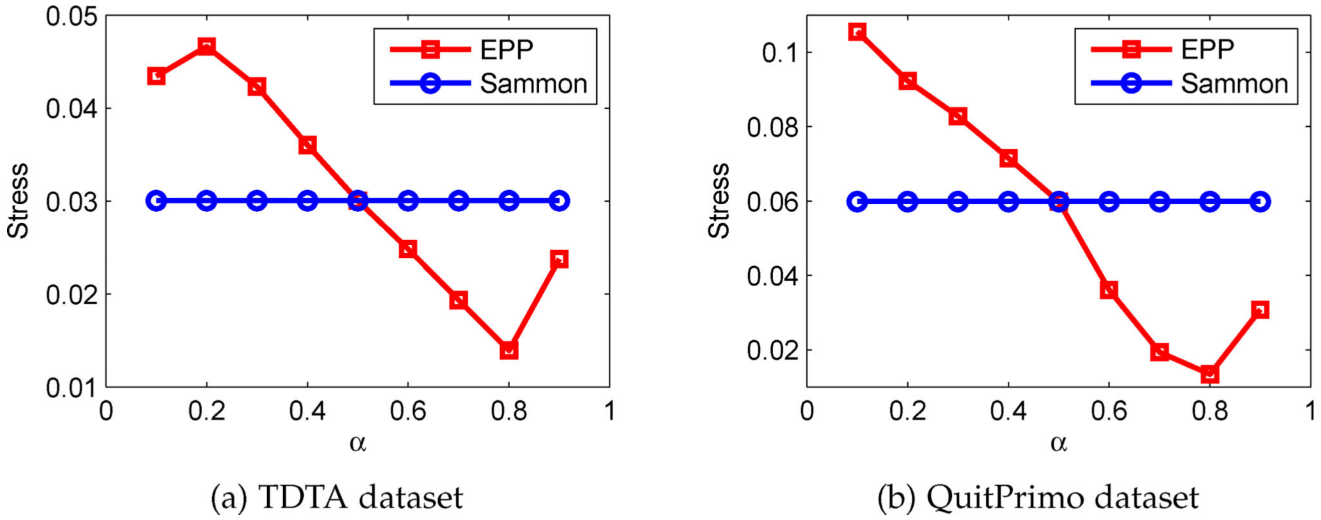
(a) Andrews Curve

(b) Grand tour

(c) EPP

Fig. 3.

Projection Pursuit of QuitPrimo data using Andrews Curve, grand tour and proposed EPP



(a) TDTA dataset

(b) QuitPrimo dataset

Fig. 4.

Finding an optimal pair of weights that balance the between and within stresses for TDTA and QuitPrimo using EPP (blue line is reference line from Sammon's Stress)

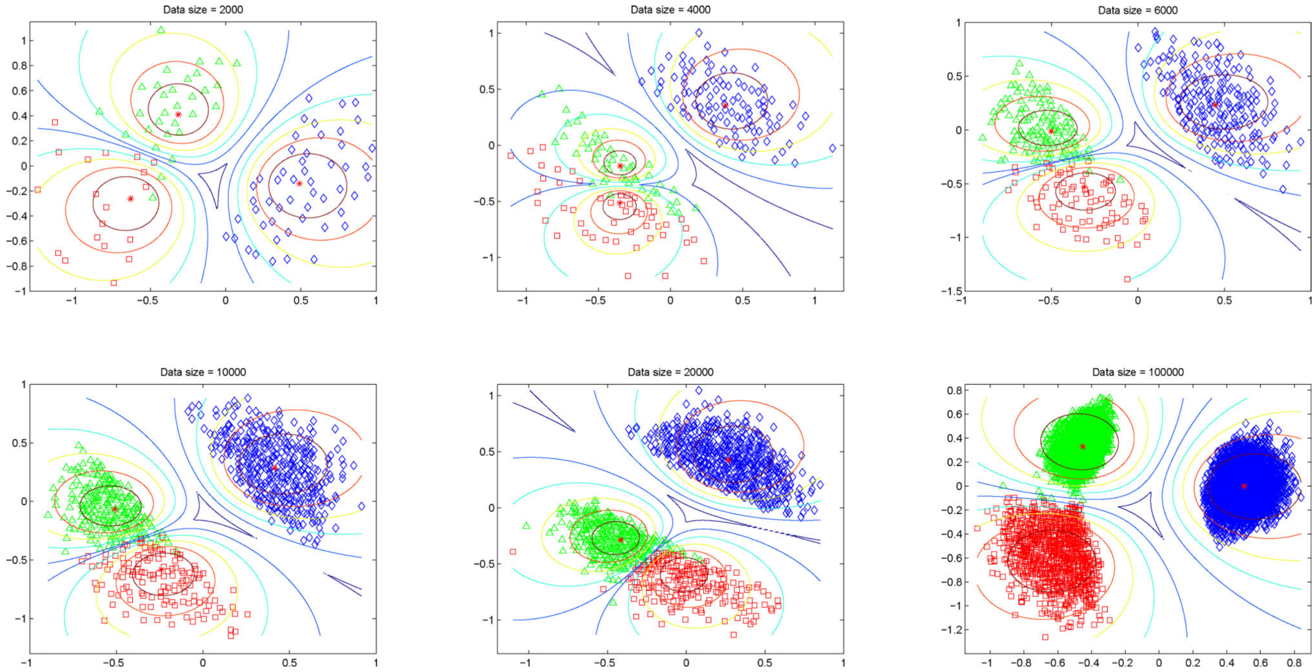


Fig. 5.

EPP for simulated longitudinal data using TDTA parameters and ℓ is from 2000 to 100000

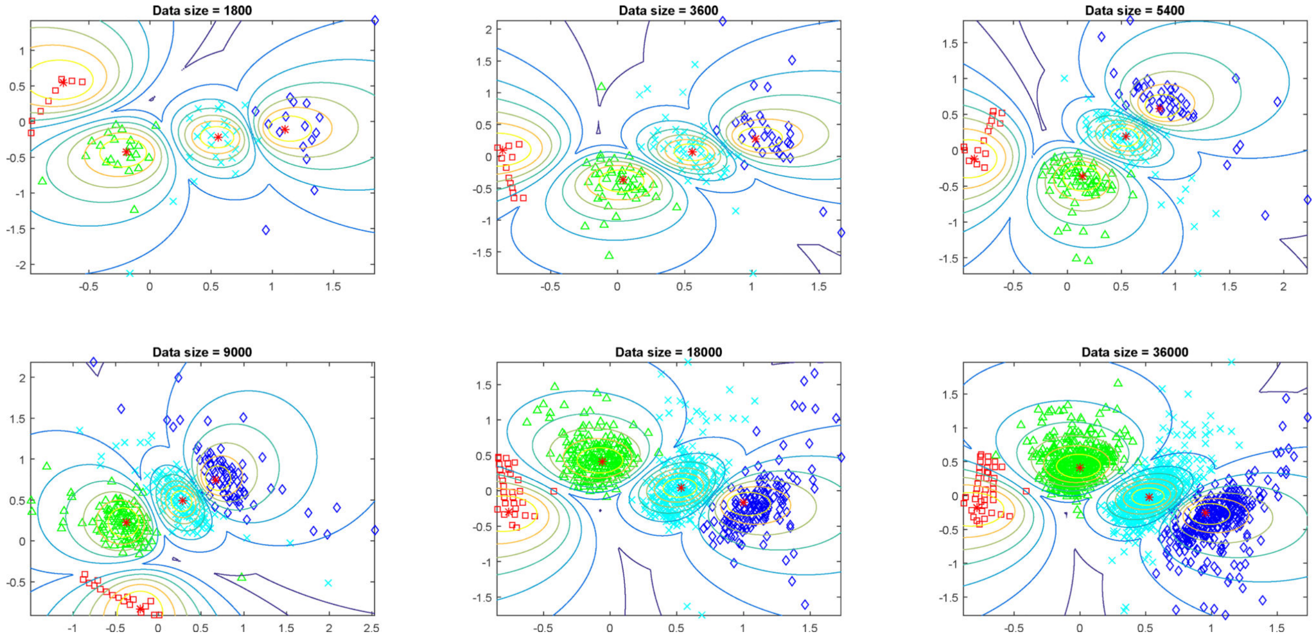
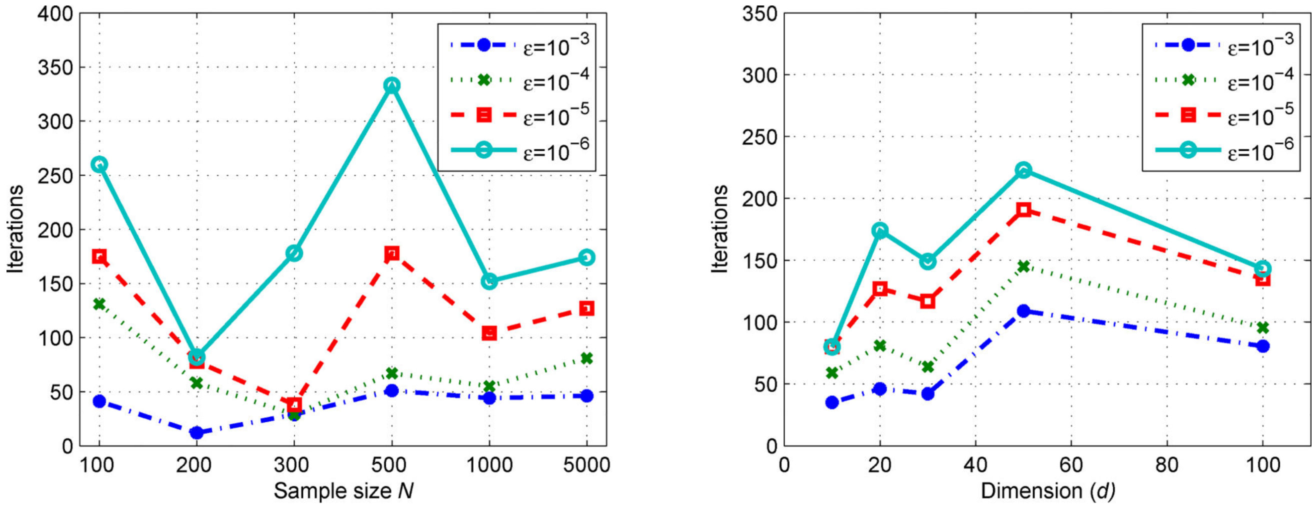


Fig. 6.

EPP for simulated longitudinal data using QuitPrimo parameters and from 1800 to 36000



(a) Simulated data, ℓ from 2000 to 100000

(b) Simulated data, ℓ fixed at 100,000

Fig. 7.

The optimal number of iterations for EPP at different number of sample sizes or dimensions for simulated data

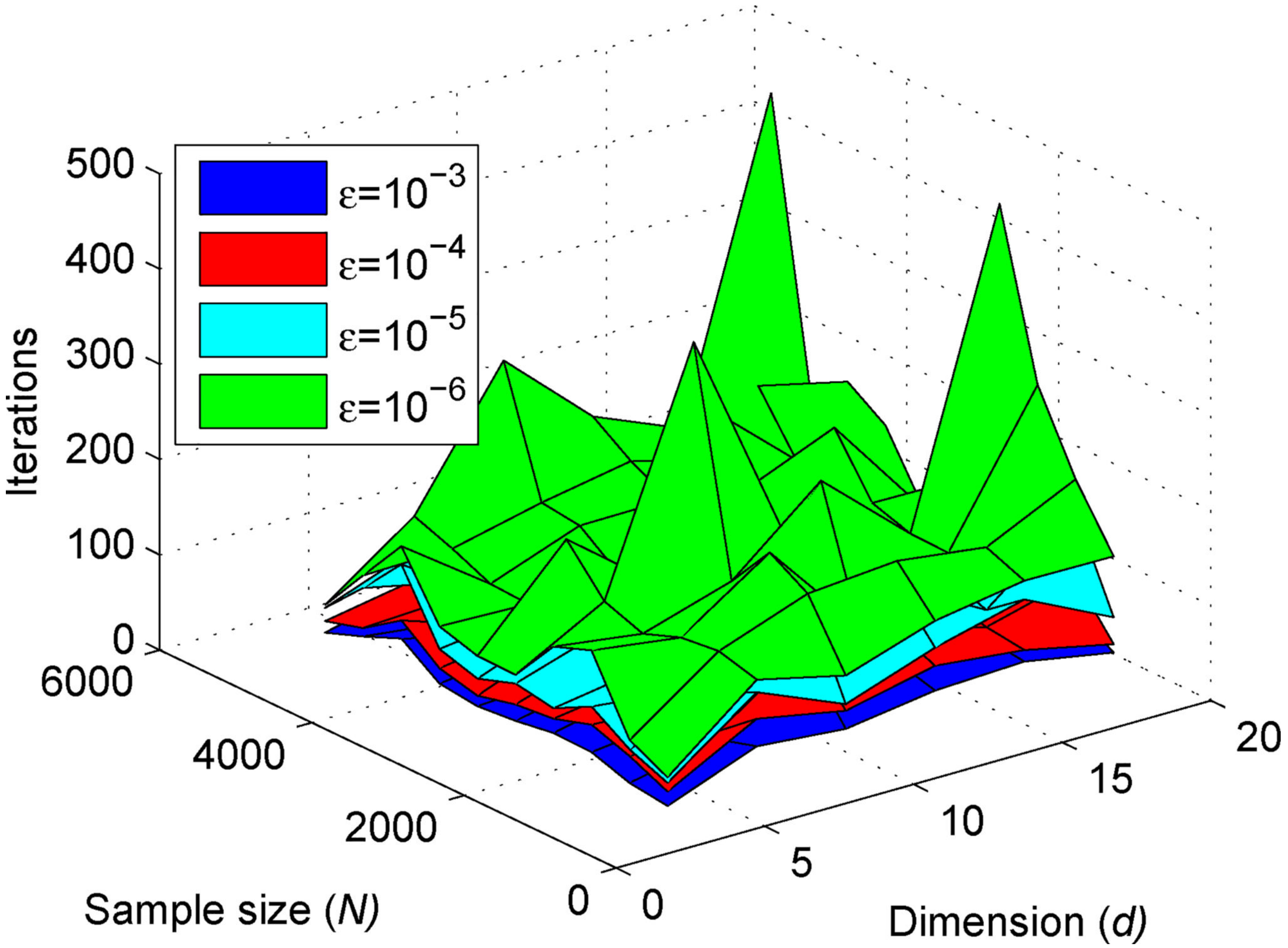


Fig. 8.

The optimal number of iterations for EPP algorithm for the artificial longitudinal data with varied sample sizes, dimensions and error bounds (ε) for the change between iterations

TABLE 1

Notations

Symbols	Definitions
$\mathbf{X}$	a vector of measurements
$\mathbf{X}_i$,$\mathbf{X}_j$	The i-th and j-th cases
$\mathbf{X}_i'$,$\mathbf{X}_j'$	The projections in a 2D space
s	angle, $0 < s < \pi$
λ	Linearly independent over the rational
$\mathbf{a}_1$ (s), $\mathbf{a}_2$ (s)	Orthonormal basis for a 2D plane
N	Number of cases
T	Sample times
d	Number of dimensional
p	Number of components
D_{ij}	Distance between $\mathbf{X}_i'$ and $\mathbf{X}_j'$
D_{ij}^{*}	Distance between $\mathbf{X}_i$ and $\mathbf{X}_j$
S	Stress
c_i	Cluster label of case i
k	The optimal number of clusters
$\bar{D}$	Average distance
ℓ	Total data size
α	Weight of the within-cluster stress
β	Weight of the between-cluster stress
S_{EPP}	Total EPP stress
f_x	Low-dimensional projections of data
ℓ	Size of the simulated data

TABLE 2

Real Data Description

Name	**Waveform**	**TDTA**	**QuitPrimo**
Cases(N)	5000	109	1320
Components(p)	21	5	3
Time points(t)	1	4	6
Total data size(ℓ)	105,000	2,180	23,760
Clusters(c)	3	3	4

TABLE 3

Mean values of TDTA Data

		t_1	t_2	t_3	t_4
Benefits	C_1	133	128	127	127
	C_2	138	127	133	134
	C_3	113	112	115	112
Family Norm	C_1	116	116	113	111
	C_2	116	115	114	115
	C_3	101	102	100	99

TABLE 4

Standard Deviation of TDTA Data

		t_1	t_2	t_3	t_4
	C_1	13.89	21.15	17.35	22.60
	C_2	14.88	25.80	16.21	14.54
Benefits	C_3	26.38	16.11	16.59	19.95
	C_1	9.94	7.19	9.88	12.20
	C_2	7.22	7.98	9.14	9.63
Family Norm	C_3	12.96	10.81	12.17	9.47

TABLE 5

Mean values of QuitPrimo Data

		t_1	t_2	t_3	t_4	t_5	t_6
MM	C_1	0.747	0.154	0.017	0.025	0.006	0.000
	C_2	1.091	0.465	0.139	0.080	0.139	0.043
	C_3	0.047	0.000	0.000	0.000	0.000	0.000
	C_4	0.659	0.157	0.003	0.000	0.000	0.000
OA	C_1	5.708	8.601	8.736	6.902	3.997	3.638
	C_2	5.708	8.601	8.736	6.902	3.997	3.638
	C_3	0.888	0.100	0.000	0.000	0.000	0.000
	C_4	6.345	8.686	5.857	1.213	0.007	0.000
OC	C_1	0.284	0.020	0.006	0.006	0.003	0.006
	C_2	0.455	0.080	0.011	0.021	0.021	0.000
	C_3	0.006	0.000	0.000	0.000	0.000	0.000
	C_4	0.275	0.031	0.014	0.000	0.000	0.000

TABLE 6

Standard Deviation of QuitPrimo Data

		t_1	t_2	t_3	t_4	t_5	t_6
MM	C_1	1.718	1.124	0.237	0.339	0.106	0.000
	C_2	1.595	2.437	0.979	0.732	1.079	0.462
	C_3	0.384	0.000	0.000	0.000	0.000	0.000
	C_4	1.972	1.246	0.059	0.000	0.000	0.000
OA	C_1	1.972	1.246	0.059	0.000	0.000	0.000
	C_2	2.431	0.875	0.893	1.394	1.172	1.484
	C_3	2.249	0.457	0.000	0.000	0.000	0.000
	C_4	2.490	1.067	3.384	1.797	0.083	0.000
OC	C_1	2.490	1.067	3.384	1.797	0.083	0.000
	C_2	0.996	0.463	0.103	0.178	0.206	0.000
	C_3	0.078	0.000	0.000	0.000	0.000	0.000
	C_4	0.783	0.194	0.186	0.000	0.000	0.000

TABLE 7

Cluster Information for TDTA and QuitPrimo

	cluster	1	2	3	4
TDTA	# of cases	50	31	16	-
	proportions	0.52	0.32	0.16	-
QuitPrimo	# of cases	356	187	490	287
	proportions	0.27	0.14	0.37	0.22