ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Visualizing data through curvilinear representations of matrices

Daniel K. Sewell

Department of Biostatistics, University of Iowa, 145 N. Riverside Dr., Iowa City, IA 52242, 1-319-384-1585, United States

HIGHLIGHTS

- This paper provides a novel visualization method for any arbitrary matrix.
- The proposed approach can be used to inspect the mean and the covariance structure.
- There is great flexibility for atypical data formats such as dissimilarity matrices.
- An R package implementing this visualization methodology is provided.

ARTICLE INFO

Article history:
Received 1 November 2017
Received in revised form 20 July 2018
Accepted 22 July 2018
Available online xxxx

Keywords:
Andrews curves
Basis splines
Fourier series
Singular value decomposition

ABSTRACT

Most high dimensional data visualization techniques embed or project the data onto a low dimensional space which is then used for viewing. Results are thus limited by how much of the information in the data can be conveyed in two or three dimensions. Methods[1] are described for a lossless functional representation of any real matrix that can capture key features of the data, such as distances and correlations. This approach can be used to visualize both subjects and variables as curves, allowing one to see patterns of subjects, patterns of variables, and how the subject and variable patterns relate to one another. A theoretical justification is provided for this approach, and various facets of the method's usefulness are illustrated on both synthetic and real data sets.

1. Introduction

Understanding high dimensional data is crucial to information conveyance and hypothesis formation. Visualization is a critical component of human understanding, and as such has received much attention. However, the problem of visualizing high dimensional data is, in the words of Bertini et al. (2011), "notoriously complex and cumbersome". The task of information extraction through visual processes has seen a wide array of approaches. Most of these, however, follow a fixed set of themes, building and adapting core approaches to suit the ever growing needs of researchers faced with data having high dimensionality or complex structure.

The approach proposed in this paper is based on representing any matrix as a set of curves and then visualizing the resulting curvilinear representation of the data matrix. Two existing approaches to visualize data through curves are parallel coordinates (Wegman, 1990; Inselberg and Dimsdale, 1990) and Andrews curves (Andrews, 1972). The idea behind parallel coordinates is to use parallel rather than orthogonal axes, and represent each observed data point as a line connecting the axes; hence n data points in a d-dimensional space can be represented as n sequences of $d-1$ connected line segments. An example of this is given in Fig. 1c. Parallel coordinates have been extended in a number of ways, such as angular histograms

E-mail address: daniel-sewell@uiowa.edu.
URL: http://myweb.uiowa.edu/dksewell/home.html.

https://doi.org/10.1016/j.csda.2018.07.010
0167-9473/

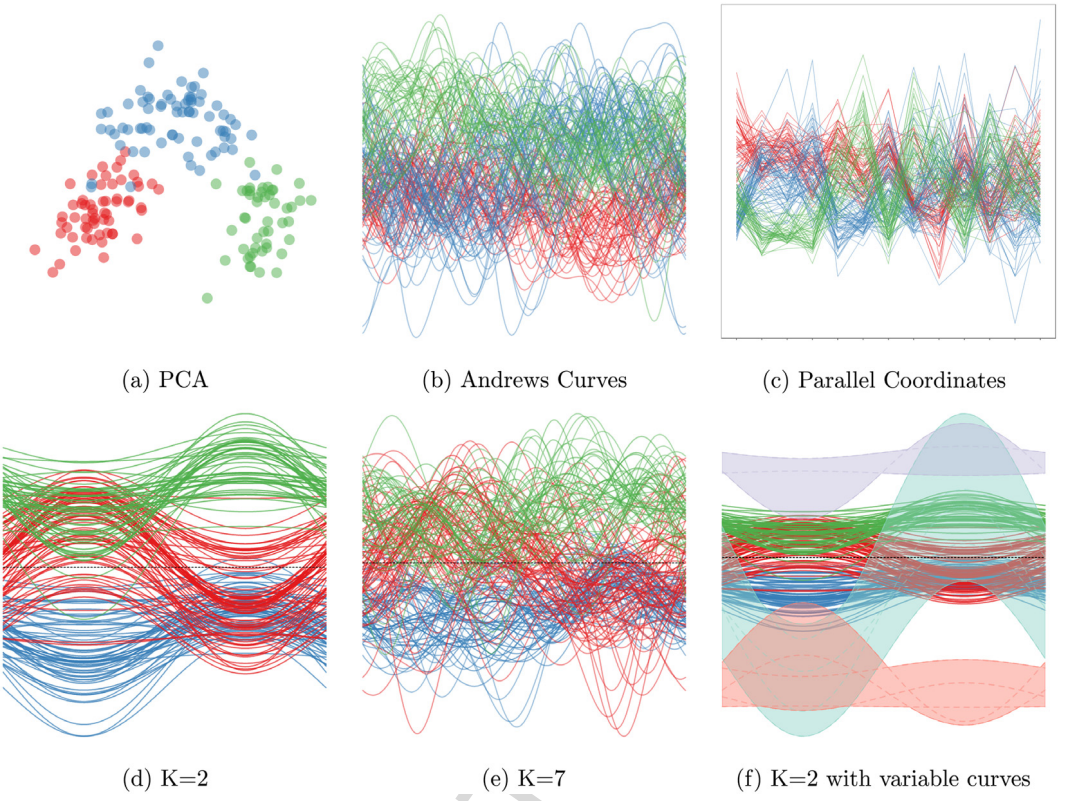

Fig. 1. Wine data using PCA, Andrews curves, and the proposed partial Fourier series (PFS) representation. Each point in (a) and line in (b) through (e) corresponds to a wine; the colors correspond to three cultivars. (d) and (e) show the PFS representations of the best rank-2 and rank-7 approximations of the data respectively, showing that for higher rank representations we are just trying to visualize the noise in the data rather than the signal. In (f), the solid lines correspond to the wines and the dotted lines/shaded regions correspond to variables; the inner product of the ith wine curve and the jth variable curve equals Y_{ij}. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Geng et al., 2011) and parallel sets (Bendix et al., 2005; Hofmann and Vendettuoli, 2013) (which extend parallel coordinates to categorical data).

The second approach that maps high dimensional points to curves, which is closer in spirit to our proposed approach, is Andrews curves. Andrews (1972) first had the idea of representing a vector of measurements as a curve; specifically, for $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iq})$ Andrews suggested plotting the curve

$$f_{Y_i}(t) = Y_{i1}/\sqrt{2} + Y_{i2}\sin(t) + Y_{i3}\cos(t) + Y_{i4}\sin(2t) + Y_{i5}\cos(2t) + \cdots .$$

This approach has the benefits of

1. mean preservation, i.e., $f_{\bar{Y}}(t) = \frac{1}{n}\sum f_{Y_i}(t)$,
2. distance preservation, $\int_{-\pi}^{\pi}\left(f_{Y_i}(s) - f_{Y_j}(s)\right)^2 ds \propto \|Y_i - Y_j\|^2$,
3. the representation corresponds to the magnitude of one-dimensional projections onto the vector $(1/\sqrt{2}, \sin(t), \cos(t), \sin(2t), \cos(2t), \ldots)$, and by plotting the Andrews curves we are assessing the continuum of these one dimensional projections, and
4. if Y_{ij} are uncorrelated with constant variance σ^2, the variance of $f_{Y_i}(t)$ is proportional to σ^2.

Koziol and Hacke (1991) extended this approach to paired multivariate data. Various tweaks on Andrews' original method have been proposed, such as those given by Gnanadesikan (1997), Kulkarni and Paranjape (1984), Khattree and Naik (2002) and García-Osorio and Fyfe (2005). Functions other than combinations of sine and cosine terms have been proposed as well, including Chebyshev polynomials, Legendre polynomials, and wavelets (Embrechts and Herzberg, 1991; Rietman and Layadi, 2000), though they have not as yet seen widespread use.

Andrews curves have their disadvantages. First, as seen by the many variations on the functional structure used, this approach is largely ad hoc with little theoretical foundation for support. Second, they are not order preserving. That is, it is critically important to place the most important variables alongside the large period terms as these are the terms that humans are most adept at inspecting. However, this has been shown to be a NP-complete optimization problem

Please cite this article in press as: Sewell D.K., Visualizing data through curvilinear representations of matrices. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2018.07.010.