Computational Statistics and Data Analysis 74 (2014) 157–179

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Dimension reduction in principal component analysis for trees

Carlos A. Alfaro [a], Burcu Aydın [b,*], Carlos E. Valencia [a], Elizabeth Bullitt [c], Alim Ladha [c]

[a] Departamento de Matemáticas, Centro de Investigación y de Estudios Avanzados del IPN, Apartado Postal 14–740, 07000 Mexico City, D.F., Mexico

[b] HP Laboratories, 1501 Page Mill Rd MS 1140, Palo Alto, CA, United States

[c] Department of Neurosurgery, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

ARTICLE INFO

Article history:
Received 10 August 2012
Received in revised form 21 August 2013
Accepted 21 December 2013
Available online 10 January 2014

Keywords:
Object oriented data analysis
Combinatorial optimization
Principal component analysis
Tree-lines
Tree structured objects
Dimension reduction

ABSTRACT

The statistical analysis of tree structured data is a new topic in statistics with wide application areas. Some Principal Component Analysis (PCA) ideas have been previously developed for binary tree spaces. These ideas are extended to the more general space of rooted and ordered trees. Concepts such as tree-line and *forward* principal component tree-line are redefined for this more general space, and the optimal algorithm that finds them is generalized.

An analog of the classical dimension reduction technique in PCA for tree spaces is developed. To do this, *backward* principal components, the components that carry the least amount of information on tree data set, are defined. An optimal algorithm to find them is presented. Furthermore, the relationship of these to the forward principal components is investigated, and a path-independence property between the forward and backward techniques is proven.

These methods are applied to a brain artery data set of 98 subjects. Using these techniques, the effects of aging to the brain artery structure of males and females is investigated. A second data set of the organization structure of a large US company is also analyzed and the structural differences across different types of departments within the company are explored.

1. Introduction

In statistics, data sets that reside in high dimensional spaces are quite common. A widely used set of techniques to simplify and analyze such sets is *principal component analysis* (PCA). It was introduced by Pearson in 1901 and independently by Hotelling in 1933. A comprehensive introduction can be found in Jolliffe (2002).

The main aim of PCA is to provide a smaller subspace such that the maximum amount of information is retained when the original data points are projected onto it. This smaller subspace is expressed through components. In many contexts, one dimensional subspaces are called lines, and we follow this terminology. The line that carries the most variation present in the data set is called *first principal component* (PC1). The *second principal component* (PC2) is the line such that when

* Corresponding author.
E-mail addresses: alfaromontufar@gmail.com (C.A. Alfaro), burcuaydin@gmail.com (B. Aydın), cvalencia@math.cinvestav.edu.mx (C.E. Valencia), bullitt@med.unc.edu (E. Bullitt), alim.ladha@gmail.com (A. Ladha).

0167-9473/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.12.007

combined with PC1, the most variation that can be retained in a two-dimensional subspace is kept. One may repeat this procedure to find as many principal components as necessary to properly summarize the data set in a manageable sized subspace.

Another way to characterize the principal components is to consider the distances of the data points to a given subspace. The line which minimizes the sum of squared distances of data points onto it can be considered as PC1. Similarly, PC2 is the line that, when combined with PC1, the sum of squared distances of the data points to this combination is minimized. In Euclidean space, these two characterizations are equivalent.

An important topic within PCA is called *dimension reduction* (see Mardia et al., 1973 for dimension reduction and Jolliffe, 2002, p. 144, for *backward elimination method*). The aim of dimension reduction method is to find the components such that when they are eliminated, the projection of the data onto the remaining subspace will retain the maximum amount of variation. Or alternatively, the remaining subspace will have the minimum sum of squared distances to the data points. These are the components with least influence.

We would like to note that, in the general sense, any PCA method can be regarded as a dimension reduction process. However, Mardia et al. (1973) reserves the term dimension reduction specifically for this method, which some other resources also refer as backward elimination, or backward PCA. In this paper we will follow Mardia et al. (1973)'s convention, together with "backward PCA" terminology. The original approach will be called *forward PCA*.

In Euclidean space, the choice of which technique to use depends on the needs of the end user: If only the few principal components with most variation in them are needed, then the forward approach is more suitable. If the aim is to eliminate only the few least useful components, then the backward approach would be the appropriate choice.

The historically most common space used in statistics is the Euclidean space ($\mathbb{R}^n$) and the PCA ideas were first developed in this context. In $\mathbb{R}^n$, the two definitions of PCs (maximum variation and minimum distance) are equivalent, and the components are all orthogonal to each other. In Euclidean space, applying forward or backward PCA n times for a data set in $\mathbb{R}^n$ would provide an orthogonal basis for the whole space.

Moreover, in this context, the set of components obtained with the backward approach is the same as the one obtained by the classical forward approach, only the order of the components is reversed. This is a direct result of orthogonality properties in Euclidean space. This phenomenon can be referred as *path independence* and it is very rare in non-Euclidean spaces. In fact, to the best of the authors' knowledge, this paper is presenting the first known example of path independence in non-Euclidean spaces.

With the advancement of technology, more and more data sets that do not fit into the Euclidean framework became available to researchers. A major source of these is biological sciences; collecting detailed images of their objects of interest using advanced imaging technologies. The need to statistically analyze such non-traditional data sets gave rise to many innovations in statistics. The type of non-traditional setting we will be focusing in this paper is sets of trees as data. Such sets arise in many contexts, such as blood vessel trees (Aylward and Bullitt, 2002), lung airways trees (Tschirren et al., 2002), and phylogenetic trees (Billera et al., 2001).

A first starting point in PCA analysis for trees is Wang and Marron (2007), who attacked the problem of analyzing the brain artery structures obtained through a set of Magnetic Resonance Angiography (MRA) images. They modeled the brain artery system of each subject as a binary tree and developed an analog of the forward PCA in the binary tree space. They provided appropriate definitions of concepts such as distance, projection and line in binary tree space. They gave formulations of first, second, etc. principal components for binary tree data sets based on these definitions. This work has been the first study in adapting classical PCA ideas from Euclidean space to the new binary tree space. Wang and Marron (2007)'s definitions involve a vector of attributes for each node.

The PCA formulations of Wang and Marron (2007) gave rise to interesting combinatorial optimization problems. Aydın et al. (2009) provided an algorithm to find the optimal principal components in binary tree space in linear time. This work however used the simplified versions of Wang and Marron (2007)'s definitions where attributes are not considered. Only topology information is included in the analysis. This development enabled a numerical analysis on a full-size data set of brain arteries, revealing a correlation between their structure and age.

In the context of PCA in non-Euclidean spaces, Jung et al. (2010) gave a backward PCA interpretation in image analysis. They focus on *mildly non-Euclidean*, or manifold data, and propose the use of Principal Nested Spheres as a backward step-wise approach.

Marron et al. (2010) provided a concise overview of backward and forward PCA ideas and their applications in various non-classical contexts. They also mention the possibility of backwards PCA for trees: "…The notion of backwards PCA can also generate new approaches to tree line PCA. In particular, following the backwards PCA principal in full suggests first optimizing over a number of lines together, and then iteratively reducing the number of lines". This quote essentially summarizes one of our goals in this paper.

In this work, our first goal is to define and discuss the subject of rooted ordered tree spaces. We will elaborate on the correspondence concept, which is at the heart of any numerical analysis for ordered tree data. We will also suggest some indexing methods, and provide the generalized versions of some basic definitions such as distance, projection, etc.

Secondly, we will extend the definitions and results of Wang and Marron (2007) and Aydın et al. (2009) on forward PCA from binary tree space to the more general rooted ordered tree space and proceed with providing optimal algorithms for