The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1597–1615
DOI: 10.1214/09-AOAS263

A PRINCIPAL COMPONENT ANALYSIS FOR TREES

BY BURCU AYDIN[1], GÁBOR PATAKI, HAONAN WANG[2],
ELIZABETH BULLITT[3] AND J. S. MARRON[4]

University of North Carolina, University of North Carolina, Colorado State University, University of North Carolina and University of North Carolina

The active field of Functional Data Analysis (about understanding the variation in a set of curves) has been recently extended to Object Oriented Data Analysis, which considers populations of more general objects. A particularly challenging extension of this set of ideas is to populations of tree-structured objects. We develop an analog of Principal Component Analysis for trees, based on the notion of tree-lines, and propose numerically fast (linear time) algorithms to solve the resulting problems to proven optimality. The solutions we obtain are used in the analysis of a data set of 73 individuals, where each data object is a tree of blood vessels in one person's brain. Our analysis revealed a significant relation between the age of the individuals and their brain vessel structure.

1. Introduction.

Functional data analysis has been a recent active research area: we refer the reader to Ramsay and Silverman (2002, 2005) for a good introduction and overview, and Ferraty and Vieu (2006) for a more recent viewpoint. A major difference between this approach and more classical statistical methods is that curves are viewed as the *atoms* of the analysis, that is, the goal is the statistical analysis of a *population of curves*.

Wang and Marron (2007) recently extended functional data analysis to Object Oriented Data Analysis (OODA), where the atoms of the analysis are allowed to be more general data objects. Examples studied there include images, shapes and tree structures as the atoms, that is, the basic data elements of the population of interest. Other recent examples are populations of movies, as in functional magnetic resonance imaging. A major contribution of Wang and Marron (2007) was the development of a set of tree-population analogs of standard functional data analysis techniques, such as Principal Component Analysis (PCA). The foundations were laid via the formulation of particular optimization problems, whose solution resulted in that analysis method (in the same spirit in which ordinary PCA can be formulated in terms of an optimization problem).

Received October 2008; revised March 2009.

[1] Supported in part by NSF Grant DMS-06-06577 and NIH Grant RFA-ES-04-008.

[2] Supported in part by NSF Grant DMS-07-06761.

[3] Supported in part by NIH Grants R01EB000219-NIH-NIBIB and R01CA124608-NIH-NCI.

[4] Supported in part by NSF Grant DMS-06-06577 and NIH Grant RFA-ES-04-008.

Key words and phrases. Object oriented data analysis, population structure, principal component analysis, tree-lines, tree-structured objects.