Journal of Multivariate Analysis 154 (2017) 199–215

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Gaussian tree constraints applied to acoustic linguistic functional data

Nathaniel Shiers [a], John A.D. Aston [b,*], Jim Q. Smith [a], John S. Coleman [c]

[a] *Department of Statistics, University of Warwick, Coventry, UK*
[b] *Statistical Laboratory, University of Cambridge, Cambridge, UK*
[c] *Phonetics Laboratory, University of Oxford, Oxford, UK*

ARTICLE INFO

Article history:
Received 26 September 2015
Available online 11 October 2016

AMS subject classifications:
primary 62P99
secondary 62G05
13P25

Keywords:
Gaussian tree constraints
Canonical variate analysis
Functional data analysis
Phonetics
Separable covariance

ABSTRACT

Evolutionary models of languages are usually considered to take the form of trees. With the development of so-called tree constraints the plausibility of the tree model assumptions can be assessed by checking whether the moments of observed variables lie within regions consistent with Gaussian latent tree models. In our linguistic application, the data set comprises acoustic samples (audio recordings) from speakers of five Romance languages or dialects. The aim is to assess these functional data for compatibility with a hereditary tree model at the language level. A novel combination of canonical function analysis (CFA) with a separable covariance structure produces a representative basis for the data. The separable-CFA basis is formed of components which emphasize language differences whilst maintaining the integrity of the observational language-groupings. A previously unexploited Gaussian tree constraint is then applied to component-by-component projections of the data to investigate adherence to an evolutionary tree. The results highlight some aspects of Romance language speech that appear compatible with an evolutionary tree model but indicate that it would be inappropriate to model all features as such.

1. Introduction

Functional data analysis (FDA) is the statistical analysis of intrinsically infinite dimensional objects which can often be described in terms of curves or (hyper-)surfaces. The range of applications (e.g., brain imaging [69], climatology [7], and medical research [61]) for FDA is considerable, and this has resulted in considerable theoretical developments as well [35,60]. The use of FDA in statistical phonetics has recently attracted attention (e.g., [40,51]) as many phonological processes can essentially be seen as continuous, whether that be sound waves or derived objects such as spectrograms, where, in the tradition of FDA, the discretization of the signal is more properly considered as being inherently arbitrary and a result of the measurement rather than a property of the object under consideration. Analyses which involve acoustic functional data have provided particularly promising and interesting results in a diverse range of settings. For example, Grabe et al. [29] use a polynomial basis expansion to examine pitch variation in English, while Aston et al. [4] investigate Qiang, a Sino-Tibetan language, and find previously unidentified gender differences amongst speakers via a functional principal components based modeling approach.

* Corresponding author.
E-mail addresses: n.l.shiers@warwick.ac.uk (N. Shiers), j.aston@statslab.cam.ac.uk (J.A.D. Aston), j.q.smith@warwick.ac.uk (J.Q. Smith), john.coleman@phon.ox.ac.uk (J.S. Coleman).

http://dx.doi.org/10.1016/j.jmva.2016.09.015

The acoustic structure of spoken words can be used to investigate areas of linguistic interest in a similar way that discrete (alphabetic) representations of speech have been utilized to make cross-language comparisons, and more recently inferences regarding proto languages [11,33]. The differences and similarities between spoken languages suggest that any meaningful functional observations taken across languages are unlikely to be independently, identically distributed. As such, it is probable that the language relationships form a tree or network structure, which may be informative about possible historical developments of these languages. If this alternative (acoustic) approach can be used to corroborate known and uncontroversial language relationships, then our methods offer great potential for less certain language relationships. For instance, this would be useful for languages where there are few historical records but in which inference of a family tree is reasonably supported by the contemporary data (e.g., African language families), or alternatively, in cases where reconstruction of a family tree is disputed, such as Greenberg's classification of Native American languages [10].

Relationships between languages have long been described as phylogenetic trees constructed using linguistic factors (e.g., Schleicher [64]) where all non-leaf variables are unobserved and represent features of the past languages before their divergence. Greenberg [30] developed some of the first quantitative methods which were used to investigate evolutionary relationships between languages. More recently there have been large scale attempts to reconstruct trees or networks of languages (e.g., Nakhleh et al. [52] for the Indo-European language family, Nicholls and Ryder [54] for the Semitic language family). Some researchers have shifted away from describing the evolutionary language relationships via trees toward using networks (for example, [26,53]). However, trees have a somewhat more natural interpretation in terms of evolutionary structure, and assessing the suitability of a tree model for language data would therefore be of interest to researchers in linguistics. The focus of this paper is to examine functional acoustic data from speakers of five Romance languages (French, Italian, Portuguese, American Spanish, and Iberian Spanish) to provide insight at an exploratory level as to whether a tree may be adequate for describing certain features of these language relationships. We will examine this through a specific acoustic representation (the spectrogram), which captures considerable acoustic information for each word.

To address questions of whether data is compatible with a latent tree model (tree-amenability), we appeal to the notion of tree constraints. The theory of tree constraints is embedded in the area of algebraic statistics, a field that has a significant recent literature related to phylogenetics (e.g., [2,70]). It has been known for some time that covariance functions of data on observed variables respecting an evolutionary tree must obey particular algebraic and semi-algebraic constraints, e.g., [65]. Recently these have become much better understood (for example [1,3,19]) and fully characterized in some cases (e.g., the binary case [75,76]). In this paper, building on developments in Shiers and Smith [67], a Gaussian analogue of the binary tree constraints is applied to the covariances of component-by-component projections of the data. By considering the data component-wise, a more realistic and nuanced analysis can be performed which permits some observed features of linguistic data to be tree-amenable and others not.

The overall aim of this paper is to present a methodology for assessing whether particular features of spoken languages may be suitably modeled by Gaussian trees with latent interior variables. To this end, it is necessary to identify phonetic features that effectively distinguish languages. This is achieved by projecting high dimensional spectrograms to a novel separable-canonical variate basis, to obtain a meaningful low dimensional representation of data. The features highlighted by this projection can then be assessed for compatibility with evolutionary trees. Throughout, a Romance language data set is used to illustrate the methodology as a proof of concept for phonetic data analysis.

In more detail, Section 2 describes the data set and the preprocessing in preparing an audio recording for FDA. In a similar spirit to the work on object data by Wang and Marron [72], for the application to Romance languages presented here, the observation units of interest are two-dimensional functional data objects known as spectrograms (time–frequency descriptions of the data). These are formed from transformations of audio recordings of people speaking single words. When regarded as functional data, observations are in fact stored as high dimensional objects. Therefore, in Section 3 tools from FDA are employed to transform high dimensional speaker data to a lower dimension. This is achieved through the novel approach of using between-language covariance as described in canonical function analysis (CFA) and combining it with a tensor decomposable covariance structure.

Having achieved the required reduction in dimension, Section 4 provides a brief summary of tree constraints. A fundamental but yet to be exploited constraint for use with Gaussian data is then introduced. Statistics associated with the violation of or adherence to the constraint are then constructed from the acoustic language data to answer the question of tree suitability. In Section 4.3, in preparation for use with this Gaussian tree constraint, we describe the construction of a between-language cross-covariance matrix using the scores (projections) of the acoustic data. In Section 4.4, the general effectiveness of the Gaussian tree constraint is investigated via simulations to assess its ability to correctly accept or reject tree-amenability. Section 4.5 entails further simulation, tailored to mimic aspects of the acoustic data, so the results are more immediately relevant to the setting. Section 5 addresses whether any of the isolated features of the spoken languages can be described by a Gaussian evolutionary tree model. These tree constraints are then used to explore tree-amenability for each of the components of the chosen basis. It is found that a subset of the components adhere to the tree constraint. This suggests that some features of the acoustic linguistic data which distinguish between languages could have evolved in a tree-like manner whilst others have not. This preliminary result is consistent with the current understanding that the development of Romance languages has been complex, involving much cross-language interaction [32] in addition to a historically well-documented common origin from Latin. Thus, attempting to fit a tree model to the entire data set would be misguided based on the empirical evidence presented here. More appropriately, a different model can be fitted for each component, using the tree constraints to indicate whether to restrict the space of models to trees.