WARWICK
THE UNIVERSITY OF WARWICK

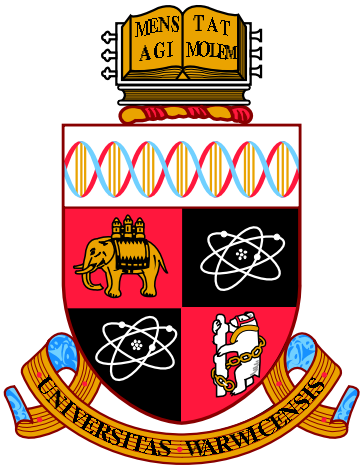
A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:
http://wrap.warwick.ac.uk/80590

Copyright and reuse:
This thesis is made available online and is protected by original copyright.
Please scroll down to view the document itself.
Please refer to the repository record for this item for information to help you to cite it.
Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

warwick.ac.uk/lib-publications



Gaussian Latent Tree Model Constraints for Linguistics and Other Applications

by

Nathaniel Shiers

A thesis submitted for the degree of

Doctor of Philosophy in Statistics

University of Warwick, Department of Statistics

June 2016

THE UNIVERSITY OF WARWICK

Contents

List of Figures

List of Tables

Declarations

I hereby declare that this thesis is the result of my own work and research, except where otherwise indicated. This thesis has not been submitted for examination to any institution other than the University of Warwick.

Some of this work has been published or is currently under review:

- *Graphical inequality diagnostics for phylogenetic trees* [Shiers and Smith, 2012] has been published in the refereed conference proceedings from the 6th European Workshop on Probabilistic Graphical Models. This paper covers the latter half of Chapter 4 which introduces graphical diagnostics for the semi-algebraic binary tree constraints.

- *Gaussian tree constraints applied to acoustic linguistic functional data* [Shiers et al., 2014] has been submitted for review. Parts of this paper make up sections in Chapter 3 detailing the Romance language data set and the functional data tools used, and in Chapter 7 as the main linguistic application but with a much more extensive analysis.

- *The correlation space of Gaussian latent tree models with applications.* [Shiers et al., 2016] has been submitted for review. Sections of the paper make up Chapter 5 and Chapter 6, plus the yeast growth example in Chapter 7.

Acknowledgements

A huge thanks to Jim Q. Smith and John A. D. Aston for providing the opportunity to study for a PhD under their supervision and for providing such an interesting area of research. They have offered extensive wisdom, patience, academic guidance and moral support. I count myself very lucky to have had such knowledgeable and attentive supervisors.

I wish to thank Piotr Zwiernik for helpful discussions and an enjoyable collaboration. A big thank you to John S. Coleman for providing his experience and linguistic expertise to provide deeper meaning to the statistical output. I am very much indebted to Pantelis Hadjipantelis for preprocessing the Romance language data set.

Thank you to the Department of Statistics at the University of Warwick, to the ESRC DTC, and to the ESRC for their financial support (grant ES/I90427/1). Thank you to the panel members and everyone else who has provided academic feedback on my work.

I am fortunate to have such loving family and friends. Thank you Alison, John, Francis, Eleanor and Sebastian. Thank you Kimmy for all your love and support that has kept me going. Thank you to Frances for your encouragement and understanding. Thank you to all my friends, housemates and colleagues over the years. Thank you to everyone in the SU and to the friends who have made me feel at home through People & Planet and Fossil Free.

To Jacquie

Abstract

The relationships between languages are often modelled as phylogenetic trees whereby there is a single shared ancestral language at the root and contemporary languages appear as leaves. These can be thought of as directed acyclic graphs with hidden variables, specifically Bayesian networks. However, from a statistical perspective there is often no formal assessment of the suitability of these latent tree models. A lot of the work that seeks to address this has focused on discrete variable models. However, when observations are instead considered as functional data, the high dimensional approximations are often better considered in a Gaussian context. The high dimensional data is often inefficiently stored and so the first challenge is to project this data to a low dimension while retaining the information of interest. One approach is to use the newly developed tool named separable-canonical variate analysis to form a basis.

Extending the techniques for assessing latent tree model compatibility to beyond discrete variables, the complete set of Gaussian tree constraints are derived for the first time. This set comprises equations and inequality statements in terms of correlations of observed variables. These statements must in theory be adhered to for a Gaussian latent tree model to be appropriate for a given data set. Using the separable-canonical variate analysis basis to obtain a truncated representation, the suitability of a phylogenetic tree can then be plainly assessed. However, in practice it is desirable to allow for some sampling error and as such probabilistic tools are developed alongside the theoretical derivation of Gaussian tree constraints.

The proposed methodology is implemented in an in-depth study of a real linguistic data set to assess the phylogenies of five Romance languages. This application is distinctive as the data set consists of acoustic recordings, these are treated as functional data, and moreover these are then being used to compare languages in a phylogenetic context. As a consequence a wide range of

theory and tools are called upon from the multivariate and functional domains, and the powerful new separable-canonical function analysis and separable-canonical variate analysis are used. Utilising the newly derived Gaussian tree constraints for hidden variable models provides a first insight into features of spoken languages that appear to be tree-compatible.

Abbreviations

BN Bayesian Network.

BOLD Barcode of Life Data.

BP base pair.

CFA canonical function analysis.

COI cytochrome c oxidase I.

CTA confirmatory tetrad analysis.

CVA canonical variate analysis.

ETA exploratory tetrad analysis.

FDA functional data analysis.

FPCA functional principal component analysis.

GLTM Gaussian latent tree model.

MSE mean squared error.

PCA principal component analysis.

Notation

$|\cdot|$ cardinality.

$\perp\!\!\!\perp$ conditionally independent.

cov covariance of a matrix.

$\det$ determinant of a matrix.

$E(\cdot)$ expectation.

$\exp$ exponential function.

$G = (V, E)$ graph G with vertex set V and edge set E.

$\overline{ab}$ path between a and b.

$\Pr$ probability.

sgn signum function.

$A|B$ split of vertex sets A and B.

$\sup$ supremum.

Δ symmetric difference of two sets.

T tree.

tr trace of a matrix.

var variance.

Chapter 1

Introduction

1.1 Phylogenetic trees and networks in linguistics and biology

Evolutionary models of languages are usually considered to take the form of trees. However some researchers have shifted away from describing the evolutionary language relationships using trees instead using networks (for example, Forster and Toth [2003], Nelson-Sathi et al. [2011]). This shift has also been seen in other areas of study such as biological phylogenetics (e.g. Rieppel [2010]). On the other hand, trees have a somewhat more natural interpretation in terms of evolutionary structure. Thus, assessing the suitability of a tree model for language data is therefore of interest to researchers in linguistics. The main application of this thesis 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 is an adequate model for describing certain features of these language relationships.

To address questions of whether data is compatible with a latent tree model, 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. Allman and Rhodes [2008], Sturmfels and Sullivant [2005]). 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. Settimi and Smith [2000]. Recently

these have become much better understood (for example Allman et al. [2009, 2014], Drton and Sullivant [2007]) and fully characterised in some cases (e.g. the binary case Zwiernik and Smith [2011, 2012]). With the development of so-called tree constraints the plausibility of the tree model assumptions can be assessed for certain random variables.

In this thesis we derive and utilise tree constraints for Gaussian latent tree models (GLTMs). This allows us to check whether the moments of observed variables lie within regions consistent with a GLTM and moreover to develop statistical tools in order to get a probabilistic assessment of compatibility of data with a GLTM. 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 this functional data set 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 emphasise language differences whilst maintaining the integrity of the observational language-groupings.

The set of Gaussian tree constraints is applied to the covariances of component-by-component projections of the data to investigate adherence to an evolutionary tree. 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-compatible and others not. 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.

1.2 Key contributions of the thesis

This thesis makes a number of contributions to the literature with the main ones being listed below:

- The first key contributions are the tools separable-canonical variate analysis (separable-CVA) and the functional counterpart separable-canonical function analysis (separable-CFA), both of which are defined in Section 3.6. These are hugely useful tools that allow for CVA and CFA to be implemented in the commonly occurring situation that the number

of variables exceeds the number of observations. Importantly the assumption of separability does not affect the validity of these tools and, in the analyses we perform, the efficiency does not appear to be significantly impaired. This new tool is described in Shiers et al. [2014, Section 3], which has been submitted for review.

- The latter half of Chapter 4 introduces a preliminary exploration of graphical diagnostics for inequality constraints for binary latent tree models. This stems from the definite need to develop more nuanced assessment of tree-compatibility. The graphical diagnostics add insight, which was previously not available, into how much credence should be given to an inequality constraint being violated (or not). Furthermore, we use multi-dimensional scaling to obtain a visual aid for distinguishing between alternative trees. These results have been published in refereed conference proceedings [Shiers and Smith, 2012].

- In Chapter 5, the first complete description of tree constraints for the GLTM is derived. This is given as a set of equations and inequality statements involving correlations of the observed variables. These statements hold if and only if the correlations are associated with a GLTM, which means that these constraints can form the foundation of techniques for assessing compatibility of data sets with GLTMs. These results can be found in Shiers et al. [2016, Section 3], which has been submitted for review.

- In Chapter 6, using the tree constraints presented in Chapter 5, a novel methodology is presented for assessing GLTM compatibility of data sets. For the inequality constraints, the Wishart and inverse-Wishart distributions are employed to obtain posterior probabilities of tree-compatibility. This is an efficient and versatile tool for assessing both general tree constraints and constraints specific to a particular tree. Additionally, the link is made between the equality constraints and vanishing tetrads which then allows for the established chi-squared test to be performed to assess tree-compatibility. Moreover, two main scenarios are given for when equality constraints can be utilised: Firstly, for smaller scale models an exploratory tetrad analysis (ETA) can be implemented to select the best fitting tree (if any). Secondly, a confirmatory tetrad analysis (CTA) can be performed when a specific tree is to be tested, which is particularly useful when one wants to test a widely accepted or a newly proposed tree model. This suite of tools, for both equality and inequality constraints, makes use of all the tree constraints, and it allows assessment of a

particular data set as to whether the class of GLTMs is suitable for specified trees. This methodology can be found in Shiers et al. [2016, Sections 4–5], which has been submitted for review.

- Using the tree constraints derived in Chapter 5 and the associated methodology developed in Chapter 6, the first implementations of this newly constructed set of tools for assessing tree-compatibility are given in Chapter 7. The primary application is the acoustic linguistic data set consisting of audio recordings of speakers from five different languages. The tree constraint testing is used to assess whether any tree in the class of GLTMs is an appropriate fit for the data. In contrast, the secondary application is biological, considering whether a growth curve data set relating to several yeast species supports an existing purported evolutionary tree from the literature. Between these two examples, the full range of the constraints and associated methodologies are demonstrated, and furthermore, preliminary findings are discussed in the context of the applications. The linguistic application is the basis of Shiers et al. [2014] and it also features in a further analysis in Shiers et al. [2016, Section 6] in addition to the biological yeast example.

1.3 Structure of the thesis

The overall aim of this thesis is to present a methodology for assessing whether features of spoken languages may be suitably modelled as 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.

Chapter 2 provides the background material regarding graphical models and more specifically Bayesian networks (BNs) for which GLTMs, the model class of interest, are a particular subclass. We also touch upon the Wishart and related distributions due to their possible use in Monte Carlo sampling for model search. These ideas are brought across in later chapters to assist with assessing the suitability of GLTMs.

In Chapter 3, a brief overview is given of quantitative and statistical linguistics before moving into the use of functional acoustic data in a linguistic context. The acoustic data set originating from Romance language speakers is then introduced so as to provide motivation for the techniques we subsequently discuss in this and future chapters. At this point, we discuss the preprocessing work that the raw data undergoes in order to get to the desired form for a functional data analysis. The rest of the chapter is devoted to introducing functional data tools for dimension reduction, namely functional principal component analysis (FPCA) and canonical function analysis (CFA). These are then associated with their multivariate counterparts principal component analysis (PCA) and CVA, which are often used as approximations to FPCA and CFA respectively. Finally, we introduce the new techniques separable-CFA and separable-CVA which prove to be very useful in situations where the number of variables exceed the number of observations.

Chapter 4 introduces the concept of tree constraints, a theoretical set of equations and inequalities that must satisfied if a data set is to be deemed compatible with a latent tree model. We then review the existing literature on binary tree constraints and illustrate their use with two examples, one from linguistics and another from phylogenetics. The latter half of the chapter provides the first look at moving away from a black-and-white assessment of tree-compatibility. By utilising graphical tools a more subtle approach can be given as to judging whether a data set is tree-compatible and even as to which tree is a better fit.

In Chapter 5, we derive the complete description of the correlation space of GLTMs, thus obtaining the full set of Gaussian tree constraints. This is achieved by making the link between tree metrics and another defined space called the space of phylogenetic oranges, and then noticing the close relationship between this space and that of GLTMs. This provides a theoretical framework for assessing whether a data set is compatible with a GLTM.

Whereas Chapter 5 provides the theoretical results required for assessing tree-compatibility, Chapter 6 details how to utilise these in practice, extending the use of constraints beyond simple binary diagnostics. The inverse-Wishart is combined with the Wishart distribution to efficiently obtain a posterior probability of tree-compatibility with respect to the inequality constraints. For the equality constraints a direct link is made with the concept of vanishing tetrads opening up the

use of distributional results and a chi-squared test statistic for ETAs and CTAs. Thus methods for utilising all of the Gaussian tree-constraints are detailed.

Chapter 7 features the core application of the thesis. The aim is to assess whether a subset of Romance languages can be adequately modelled by a GLTM. Using GLTMs to model language evolution is often performed implicitly without any check on whether this is appropriate. Here we take the relatively recent approach of considering spoken languages as functional data and using these objects as the observations of interest. This application draws upon tools detailed in earlier chapters such as the novel separable-CFA and separable-CVA in Chapter 3 and the Gaussian tree constraints and associated methodology in Chapter 5 and Chapter 6 respectively. To complement the linguistic application a biological example is also given, studying the growth curves of yeast species. As with languages, a GLTM is often implicitly adopted when modelling biological evolution. In this example, an existing phylogenetic tree for the yeast species is taken from the literature and the data set is tested against this specific tree using a CTA. Between the linguistic and biological examples, we are able to showcase the full suite of Gaussian tree constraints and methodologies proposed in the thesis.

Chapter 2

Basics of graphical models and Bayesian analysis

This chapter introduces the concept of graphical models with a focus on BNs. The relevant notation and definitions are given alongside some key results. The subject of graphical models is vast and thus we keep the focus of the chapter to the key elements required for understanding the concepts in the thesis. BNs form the basis of the tree constraints reviewed in Chapter 4 and the newly derived ones in Chapter 5. In the latter half of the chapter, the Wishart and inverse-Wishart distributions are introduced with an example of their use in graphical model selection [Atay-Kayis and Massam, 2005]. These are of relevance as the inverse-Wishart will be utilised in Chapter 6 as part of the methodology for implementing Gaussian tree constraints.

2.1 Graphical models

A probabilistic graphical model is a graph that is used to encode the joint probability of random variables and to visualise the (conditional) independence relationships between these random variables. It is then possible to determine the conditional and marginal distributions of the random variables. One of the main advantages of graphical models is efficiency; the representation of the joint density is usually more compact than other descriptions but also calculations and simulations are often more efficient due to the sleek representation. On top of this, it is often easier

to quickly read information (such as conditional independence) from a graphical representation and consequently graphical models can also be good ways to communicate information regarding statistical relationships between random variables. There are plenty of texts on the subject of graphical models. For example Koller et al. [2007], Lauritzen [1996], Pearl [1988], Smith [2010], Sucar [2015], Whittaker [1990]. Here we primarily follow Lauritzen [1996] though will draw upon other sources where stated.

A graphical model is defined by its graph G and by a set of local functions of random variables, denoted by $f(\cdot)$. Thus if the full set of random variables is given by $X = \{X_1, \ldots, X_n\}$ then the joint probability is given by

$$\Pr(X_1, \ldots, X_n) = K \prod_{i=1}^{M} f(Y_i) \tag{2.1.1}$$

where $Y_i \subseteq X$ and K is a constant that normalises all probabilities such that they sum to one [Koller et al., 2007]. Before considering specific types of graphical models we shall review some definitions and notation regarding graphs that we will use later in the thesis.

Definition 2.1.1 (Graph)**.** *A graph G is a pair $G = (V, E)$ whereby V is a set of vertices and E is a set of edges. The edge set E comprises ordered pairs (v_i, v_j) whereby $v_i, v_j \subseteq V$. If $i = j$ then the edge is known as a loop. If $i \neq j$ and if $(v_i, v_j) \in E$ and $(v_j, v_i) \in E$ then there is an undirected edge between vertices v_i and v_j. Otherwise, $(v_i, v_j) \in E$ indicates a directed edge from v_i to v_j, which can be denoted $v_i \to v_j$.*

In terms of visual representation, vertices are denoted by circles with edges given by lines joining the relevant vertices complete with arrow head in the case of directed edges.

Definition 2.1.2 (Parent and child)**.** *If $v_i \to v_j$ then v_i is the parent of v_j and v_j is the child of v_i. Furthermore, $pa(v_j)$ is the set of all parents of v_j and similarly $ch(v_i)$ is the set of children of v_i.*

Definition 2.1.3 (Directed graph)**.** *A graph G is said to be directed if all edges are directed, i.e. if $(v_i, v_j) \in E$ then $(v_j, v_i) \notin E$.*

Definition 2.1.4 (Subgraph)**.** *$G_1 = (V^*, E^*)$ is a subgraph of $G = (V, E)$ (written $G^* \subseteq G$) if $V^* \subseteq V$ and $E^* \subseteq E$.*

Definition 2.1.5 (Completeness)**.** *A graph $G = (V, E)$ is said to be complete if $\forall\, i, j$, $(v_i, v_j) \in E$ or $(v_j, v_i) \in E$ or both.*

Definition 2.1.6 (Clique)**.** *If a subgraph G^* is complete, it is said to be a clique.*

Definition 2.1.7 (Skeleton)**.** *The skeleton $sk(G)$ is the graph G with all directed edges replaced with undirected edges.*

Definition 2.1.8 (Path)**.** *A sequence of vertices $(v^1, \ldots, v^j, \ldots, v^k)$ is a (non-repeating) path denoted $\overline{v^1 v^k}$ if $v^i \in V$ and if v^i are unique vertices $\forall\, i = 1, \ldots, k$, and if $(v^i, v^{i+1}) \in E\ \forall$ $i = 1, \ldots k-1$.*

Definition 2.1.9 (Ancestor)**.** *For the graph $G = (V, E)$, a vertex $v_j \in V$ is an ancestor of $v_i \in V$ if path (v_j, v_i) is found in G. The ancestor set of v_i is all such v_j and is denoted $an(v_i)$.*

Definition 2.1.10 (Separation)**.** *For a graph G, if $V_A, V_B, V_C \subset V$ and are pairwise disjoint, then C separates A from B if $\forall\, v_A \in V_A$, $v_B \in V_B$ all paths from v_A to v_B contain at least one element $v_C \in V_C$. Furthermore, V_C is known as a separator.*

Definition 2.1.11 (Cycle)**.** *A graph G contains a cycle if $\exists\ v_i, v_j \in V$ such that $\overline{v^i v^j}$ and $\overline{v^j v^i}$ are paths in G and the sets of vertices comprising these two paths are not identical.*

Definition 2.1.12 (Decomposable)**.** *A graph G is decomposable if in $sk(G)$ every cycle of length 4 or more has an edge not on the path between two vertices.*

Definition 2.1.13 (Directed acyclic graph)**.** *A graph G is a directed acyclic graph (DAG) if G is a directed graph with no cycles.*

Definition 2.1.14 (Unmarried parent and unmoralised child)**.** *For a DAG G, a vertex $v_i \in V$ is an unmarried parent if $\exists\ j, k$ with i, j, k unique, such that $(v_i, v_k), (v_j, v_k) \in E$ and $(v_i, v_j), (v_j, v_i) \notin E$. Let any such v_k be called an unmoralised child.*

Definition 2.1.15 (Verma graph)**.** *The Verma graph of a DAG G is denoted $ver(G)$ and is obtained by replacing directed edges with undirected edges $\forall\ (v_i, v_j) \in E$, where v_j is not an unmoralised child.*

Definition 2.1.16 (DAG equivalence)**.** *Two DAGs G_1 and G_2 are said to be equivalent if they encode the same set of independences.*

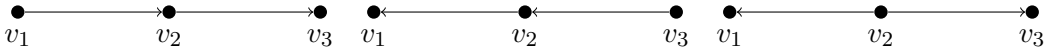


FIGURE 2.1: Equivalent DAGs.

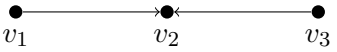


FIGURE 2.2: Non-equivalent DAG.

Definition 2.1.17 (Faithfulness)**.** *A graph G is said to be faithful in relation to a joint probability distribution P if every independence in P can be read from G.*

Theorem 1 (Equivalence of DAGs). Two DAGs G_1 and G_2 are equivalent iff $ver(G_1)$ is identical to $ver(G_2)$.

The concept of equivalent DAGs is taken from Verma and Pearl [1990]. Consider the simple DAG which is a path with three vertices (v_1, v_2, v_3). There are three graphs in the equivalent class, which are shown in Figure 2.1. However, note that the fourth combination of arrow head directions (Figure 2.2) is not equivalent to the other three DAGs. This can be observed by comparing the Verma graph which would retain the directed edges unlike the other three DAGs. One further point to note is that the direction of the arrows does not (in general) imply a causal relationship in a graphical model. The equivalence of DAGs above exemplifies why this is the case given the complete reversal of paths in Figure 2.1.

Thus far, we have introduced the basics of graphs, which make up half of the specification of a graphical model. The other half is specified by functions on the random variables as formulated by (2.1.1). We now consider the most frequently studied of the probabilistic graphical models, the BN.

A BN is a graphical model whereby for a graph $G = (V, E)$ the vertices represent random variables $X_1, \ldots, X_n$ and the edges represent the dependence relationships between the random variables. Before we give a more formal definition by stating the particular functions that embellish the graphical description to form the complete probabilistic description of a BN, we define conditional independence.

Definition 2.1.18 (Conditional independence)**.** *If X_1, X_2 and X_3 are random variables and f are probability density or mass functions, then*

$$X_1 \perp\!\!\!\perp X_2|X_3 \iff f_{X_1X_2|X_3}(x_1, x_2|x_3) = f_{X_1|X_3}(x_1|x_3)f_{X_2|X_3}(x_2|x_3)$$

where $X_1 \perp\!\!\!\perp X_2|X_3$ denotes that X_1 is conditionally independent of X_2 given $X3$.

In terms of the graphical representation of a BN the separation theorem provides us with a method for reading conditional independence statements from a BN.

Theorem 2 (Separation theorem). Consider the BN with graph $G = (V, E)$. If $V_A, V_B, V_C \subset V$ are pairwise disjoint and C separates A from B then:

$$A \perp\!\!\!\perp B|C.$$

Furthermore, we can see that

$$X_i \perp\!\!\!\perp \{an(X_i)\backslash pa(X_i)\}|pa(X_i).$$

Recall that a joint probability density can be written as a product of conditional probabilities:

$$p(x_1, \ldots, x_n) = p_n(x_n|x_1, \ldots, x_{n-1})p_{n-1}(x_{n-1}|x_1, \ldots, x_{n-2}) \ldots p_2(x_2|x_1)p_1(x_1)$$

where the vertices are numbered such that $i < j$ for each directed edge (v_i, v_j). Here each p_k can denote probability mass functions or probability density functions depending on whether X_k is discrete or continuous.

For a BN, we can make a further simplification:

$$p(x_1, \ldots, x_n) = \prod_{k=1}^{n} p_k(x_k|pa(x_k)).$$

Definition 2.1.19 (Bayesian network)**.** *Consider the random variables X_1, $\ldots$, X_n and an acyclic graph $G = (V, E)$ with n nodes whereby each node is associated with a unique random*

variable X_i. Then G is a Bayesian network for $X_1, \ldots, X_n$ if

$$\Pr(X_1, \ldots, X_n = \prod_{j=1}^{n} \Pr(X_j | pa(X_j).$$

Thus a BN for random variables $X_1, \ldots, X_n$ is defined by an acyclic graph $G = (V, E)$ and the set of probabilities $\Pr(X_i|pa(X_i))$ for each random variable X_i, $i = 1, \ldots, n$.

The thesis will pay particular attention to a specific type of BN whereby the graph G is a tree.

Definition 2.1.20 (Tree)**.** *A graph $G = (V, E)$ is a tree if in $sk(G)$ there is a unique path $\overline{v_i v_j}$ $\forall\, v_i, v_j \in V$.*

It is clear from the definition that if $sk(G)$ is a tree then its edges can be directed such that the resulting graph is a DAG by selecting a node and pointing all edges away from this vertex.

Definition 2.1.21 (Roots and leaves)**.** *If for a DAG G, $pa(v_i) = \emptyset$, then v_i is a root. If $ch(v_i) = \emptyset$, then v_i is a leaf.*

A tree can be assigned a unique root by selecting a vertex and then directing all edges away from the selected vertex. This is often seen in graphs that represent evolutionary processes where informally the directed edges represent some aspect of time. From a BN point of view, we have already seen that directed edges can be reversed under certain conditions while encoding the same conditional independences. Thus in BNs the directed edges do not necessarily represent the direction of time, though it is sometimes possible to direct them to do so. We now define a particular type of tree that is often used in evolutionary models.

Definition 2.1.22 (Binary tree)**.** *A directed tree is binary if no vertex has more than two children. If all interior vertices (i.e. non-leaves) have two children, then it is called strictly binary.*

Definition 2.1.23 (Vertex degree)**.** *The degree of a vertex v_i is the number of edges connected to the vertex and denoted $deg(v_i)$.*

Definition 2.1.24 (Trivalent)**.** *A graph G is trivalent if in $sk(G)$ the maximum degree of any vertex is 3. It is strictly trivalent if every interior vertex has degree 3.*

Binary and trivalent trees are often used to represent evolutionary models whereby the vertices represent species or languages for example. These often are explicitly assigned a single root representing the common ancestor of all the other vertices to convey the idea of evolution over time. It is often the case that random variables associated with interior nodes of a tree are unobservable, perhaps because species or languages are extinct and the relevant data unobtainable. In terms of BNs, we can represent this graphically by distinguishing between observed and unobserved random variables. If a random variable is unobserved this can be called a latent or hidden variable, in contrast to observed variables also known as manifest variables. Latent variables are represented graphically as white circles, manifest variables as black circles. Latent models are often more reflective of reality since often there are unknown but relevant random variables worth modelling. However, latent models tend to be more difficult to analyse and perform inference on and as such there is a vast dedicated literature on latent variable graphical models that is still developing (e.g. Anandkumar et al. [2014], Bentler [1980], Bollen [2014], Duncan et al. [2013], Stanghellini and Vantaggi [2013]).

Definition 2.1.25 (Latent tree model)**.** *A graph T is a latent tree model if T is a BN, a tree, and has latent interior vertices and manifest leaf vertices.*

Latent tree models are commonly used to represent evolutionary processes whereby we only observe the extant species at the leaves and so the ancestor set of all leaves is empty. This need not necessarily be a binary or trivalent tree, though often this is the case [Felsenstein, 1978]. In Settimi and Smith [2000], it is shown that the class of trivalent trees on discrete random variables contains the class of statistical models of manifest variables on all latent tree models. So in this sense we lose nothing by focusing on the former class. Binary latent tree models will be the basis of the tree constraints derived in this thesis (Chapter 5) and the focus of our analyses of Romance languages (Chapter 7). Furthermore, we consider a specific class of latent tree models that are not only binary in structure but whereby the nodes are also jointly multivariate Gaussian in distribution. We now recall the definition of the multivariate Gaussian distribution and define Gaussian latent tree models.

Definition 2.1.26 (Multivariate Gaussian distribution)**.** *Random variables $X_1, \ldots, X_n$ are distributed jointly Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if the density of the distribution is given by:*

$$f_{\boldsymbol{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$$

Definition 2.1.27 (Gaussian latent tree model)**.** *A graph T is a Gaussian latent tree model if T is a latent tree model and the random variables represented by its vertices are distributed jointly Gaussian.*

Note, that only the marginal distribution on the leaves are actually observed for the GLTM. Compare with the parametrisation of the GLTM in Section 5.2 and specifically Definition 5.2.2 where the space of multivariate normal distributions is linked to the space of marginal distributions on the leaves. As described in this later section, the description forms the starting point for understanding and utilising the model space.

To close this current section, some of the terms and concepts defined thus far are now exemplified using Figure 2.3.

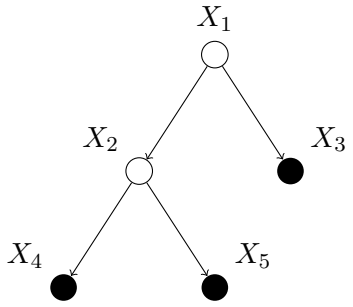


FIGURE 2.3: Example DAG, specifically a latent tree model.

Hidden nodes: X_1, X_2.

Observed nodes: X_3, X_4, X_5.

Root node: $pa(X_1) = \emptyset$ so X_1 is the root node.

Path example: Path X_1 to X_5 : (X_1, X_2, X_5).

Separation examples:

X_1 separates $A = \{X_2, X_4, X_5\}$ and $B = \{X_3\}$ so $A \perp\!\!\!\perp B | X_1$.

X_2 separates $D = \{X_1, X_3\}$ and $E = \{X_4, X_5\}$ so $D \perp\!\!\!\perp E | X_2$.

Parent, child and ancestor set examples:

$pa(X_4) = \{X_2\}, pa(X_3) = \{X_1\}$.

$ch(X_2) = \{X_4, X_5\}, ch(X_3) = \emptyset$ thus X_3 is a leaf as are X_4, X_5.

$an(X_4) = \{X_1, X_2\}$.

2.2 Learning model structure

The two main strands of learning for graphical models are estimation of parameters and model structure. In this thesis the interest is on model class (whether a tree is a suitable model) and thus learning model structure is more relevant to this work.

There are plenty of established methods for performing a model search or assessing model fit for fully observed graphs e.g. Atay-Kayis and Massam [2005], Banerjee et al. [2008], Yuan and Lin [2007], and the range of existing tools continues to expand. Loh and Wainwright [2013] make use of precision matrices to assess graph structure for discrete variable models and Schwaller et al. [2015] search the class of spanning trees under a number of distributional assumptions by considering posterior distributions across possible edges.

In Atay-Kayis and Massam [2005, Section 3], variants of the Wishart distribution, known as G-Wisharts, are used to assess model fit for decomposable models. We review the use of the G-Wishart for decomposable model search as this has some parallels to Section 6.1 where we advocate using the inverse-Wishart for assessing Gaussian tree constraints in the latent tree model setting. First we give an overview of the Wishart, inverse-Wishart and G-Wishart.

2.2.1 Wishart & inverse-Wishart

If $\hat{\Sigma}$ is based on a sample matrix X comprising n samples from $N_p(0, \Sigma)$, then the estimated scatter matrix is calculated as $\hat{S} = (n-1)\hat{\Sigma} = XX^T$ and it is well known that the scatter matrix is Wishart distributed $\hat{S} \sim \mathcal{W}_p(n, \Sigma)$ [Wishart, 1928a] with probability density function

$$f(S) = \frac{\det(S)^{\frac{n-p-1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}S))}{2^{\frac{np}{2}} \det(\Sigma)^{\frac{n}{2}} \Gamma_p(\frac{n}{2})}$$

where S is a positive definite matrix and $\Gamma_p(\cdot)$ is the multivariate gamma function which is given in its non-recursive form as:

$$\Gamma_p(a) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma(a + \frac{1-j}{2})$$

where $\Gamma(\cdot)$ is the regular gamma function. Recall that in its most general form, the gamma function is:

$$\Gamma(b) = \int_0^\infty x^{b-1} \exp(-x) dx.$$

A common prior distribution for unknown Σ is the inverse-Wishart $\mathcal{W}_p^{-1}(n_0, \Sigma_0)$, e.g. Gelman et al. [2013], Carlin and Louis [2008], Roverato [2002], with probability density function

$$f(\Sigma_0) = \frac{\det(\psi)^{\frac{n_0}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\psi\Sigma_0^{-1}))}{2^{\frac{n_0 p}{2}} \det(T)^{\frac{n_0+p+1}{2}} \Gamma_p(\frac{n_0}{2})}.$$

The inverse-Wishart is a conjugate prior, thus the posterior distribution $p(\Sigma|X)$ is the inverse-Wishart with parameters as indicated: $\mathcal{W}_p^{-1}(n_0 + n, \Sigma_0 + \hat{S})$. If $\hat{S} \sim \mathcal{W}_p(n, \Sigma)$ then $\hat{S^{-1}} \sim W_p^{-1}(n, \Sigma^{-1})$, i.e. $T = S^{-1}$ and $\psi = \Sigma^{-1}$.

Proposition 2.2.1. *The inverse-Wishart distribution is the conjugate distribution for the Wishart distribution.*

Proof. Let X be an $n \times p$ data matrix where $X \sim N_p(0, \Sigma)$ and so the scatter matrix $\hat{S} = (n-1)\hat{\Sigma} = XX^T$ and $\hat{S} \sim \mathcal{W}_p(n, \Sigma)$. Consider a prior distribution for the true covariance $\Sigma \sim \mathcal{W}_p^{-1}(n_0, \Sigma_0)$. We are interested in the posterior:

$$\begin{aligned}
f_1(\Sigma|\hat{S}) &\propto f_2(\hat{S}|\Sigma) f_3(\Sigma) \\
&= \frac{\det(\hat{S})^{\frac{n-p-1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\hat{S}\Sigma^{-1}))}{\det(\Sigma)^{\frac{n}{2}} 2^{\frac{np}{2}} \Gamma_p(\frac{n}{2})} \frac{\det(\Sigma_0)^{\frac{n_0}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma_0\Sigma^{-1})}{\det(\Sigma)^{\frac{n_0+p+1}{2}} 2^{\frac{n_0 p}{2}} \Gamma_p(\frac{n_0}{2})} \\
&\propto \frac{\exp(-\frac{1}{2}\operatorname{tr}(\hat{S}\Sigma^{-1}) \exp(-\frac{1}{2}\operatorname{tr}(\Sigma_0\Sigma^{-1}))}{\det(\Sigma)^{\frac{n}{2}} \det(\Sigma)^{\frac{n_0+p+1}{2}}} \\
&= \frac{\exp(-\frac{1}{2}\operatorname{tr}((\hat{S} + \Sigma_0)\Sigma^{-1})}{\det(\Sigma)^{\frac{n+n_0+p+1}{2}}} \\
&\implies \Sigma|\hat{S} \sim \mathcal{W}_p^{-1}(n + n_0, \hat{S} + \Sigma_0)
\end{aligned}$$

The penultimate step uses $\operatorname{tr}(A) + \operatorname{tr}(B) = \operatorname{tr}(A + B)$ (when A, B square matrices with the same dimension) and also $\det(C)^\alpha \det(C)^\beta = \det(C)^{\alpha+\beta}$. □

As in Roverato [2002], for the scale hyperparameter of the prior Σ_0 the identity matrix I_p can be used. Furthermore, by setting the degrees of freedom hyperparameter as $n_0 = p$ the prior distribution is well-defined — $n_0 \geq p$ ensures that the domain of the gamma function is respected and thus the density function of the prior is valid. Under this prior, the probability density function of the posterior inverse-Wishart distribution in Proposition 2.2.1 is given by

$$\frac{\det(\hat{S} + \Sigma_0)^{\frac{n+n_0}{2}} \exp(-\frac{1}{2}\operatorname{tr}((\hat{S} + \Sigma_0)\Sigma^{-1})}{2^{\frac{n+n_0}{2}}\Gamma_p(\frac{n+n_0}{2})\det(\Sigma)^{\frac{n+n_0+p+1}{2}}}$$

Then $\Sigma|X$ can be sampled from the posterior density. This is very efficient as the inverse-Wishart is a known distribution available in most statistical software. Alternative priors may be selected such as the scaled inverse-Wishart [O'Malley and Zaslavsky, 2008] or a strategy for modelling correlation and covariance separately [Barnard et al., 2000]. However, these alternatives bring additional computational cost and complexity. Thus, the inverse-Wishart prior is an appealing choice particularly for preliminary analyses.

2.2.2 The G-Wishart distribution for model search

Following Atay-Kayis and Massam [2005, Section 3] and adopting generally the same notation, we introduce the G-Wishart. For an undirected, decomposable graph $G = (V, E)$, let $M^+(G) = \{X \in M^+ | \text{ for } i \neq j, A(G)_{ij} = 0 \Rightarrow X_{ij} = 0\}$ where $A(G)$ is the edge adjacency matrix for G, and M^+ is the cone of positive definite matrices. Thus $M^+(G)$ is the set of positive definite matrices with the missing edges of G constrained to be zero.

Let $X_1, \ldots, X_n \sim N_p(0, \Sigma)$ where $\Sigma^{-1} \in M^+(G)$ the set of positive definite matrices restricted to the undirected graph G such that if $i \neq j$ and $(i, j) \notin E$ then $\Sigma^{-1}_{ij} = 0$.

Let $Z_i = X_i - \bar{X}_i$ where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and then let $U_i = Z_i^T Z_i$ be a $p \times p$ scatter matrix $\forall$ $i = \{1, \ldots, n\}$. Also denote $Z = (Z_1, \ldots, Z_n)$.

The joint density of the observations is thus:

$$f(Z|\Sigma^{-1}, G) = \frac{\det(\Sigma^{-1})^{\frac{n}{2}}}{(2\pi)^{\frac{np}{2}}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}U)). \tag{2.2.1}$$

We use the Diaconis-Ylvisaker conjugate prior [Diaconis and Ylvisaker, 1979] for Σ^{-1}:

$$f(\Sigma^{-1}|G) = \frac{1}{I_G(\delta, D)} \det(\Sigma^{-1})^{\frac{\delta-2}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}D)) \tag{2.2.2}$$

where $\delta \in \mathbb{R}, \delta > 2$ (prior sample size) and $D \in M(G)$ (scale parameter) with $D^{-1} \in M^+(G)$ ensures the normalising constant

$$I_G(\delta, D) = \int_{M^+(G)} \det(\Sigma^{-1})^{\frac{\delta-2}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}U))d\Sigma^{-1} \tag{2.2.3}$$

is finite. $f(\Sigma^{-1}|G)$ is known as the G-Wishart (the Wishart distribution restricted to G) written $\Sigma^{-1}|G \sim \mathcal{W}_G(\delta, D)$. In Appendix A, a detailed look at the likelihood is given that demonstrates that obtaining the maximum likelihood estimator of δ is intractable and requires numerical solutions.

We can put a uniform prior on G:

$$\pi(G) = \frac{1}{|\mathcal{G}|}$$

where $G \in \mathcal{G}$ and $|\mathcal{G}|$ the cardinality of the set of graphs of interest $\mathcal{G}$. It would then follow that

$$\begin{aligned} f(Z, \Sigma^{-1}, G) &= f(Z|\Sigma^{-1}, G) f(\Sigma^{-1}|G)\pi(G) \\ &= \frac{\det(\Sigma^{-1})^{\frac{n}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}U))}{(2\pi)^{\frac{np}{2}}} \frac{\det(\Sigma^{-1})^{\frac{\delta-2}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}D))}{I_G(\delta, D)} \frac{1}{|\mathcal{G}|} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}}|\mathcal{G}|} \frac{1}{I_G(\delta, D)} \det(\Sigma^{-1})^{\frac{\delta+n-2}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}(D+U))). \end{aligned} \tag{2.2.4}$$

Marginalising out the Σ^{-1} gives

$$\begin{aligned} p(Z, G) &= \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{1}{|\mathcal{G}|} \frac{1}{I_G(\delta, D)} \int_{M^+(G)} \det(\Sigma^{-1})^{\frac{\delta+n-2}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}(D+U)))d\Sigma^{-1} \\ &= \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{1}{|\mathcal{G}|} \frac{I_G(\delta+n, D+U)}{I_G(\delta, D)} \end{aligned} \tag{2.2.5}$$

and

$$p(Z|G) = \frac{p(Z, G)}{p(Z)} = \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{I_G(\delta+n, D+U)}{I_G(\delta, D)}. \tag{2.2.6}$$

The posterior of $G|Z$ is thus:

$$p(G|Z) = \frac{p(Z|G)\pi(G)}{\pi(Z)} \propto p(Z|G)\pi(G) \propto p(Z|G) \tag{2.2.7}$$

since $\pi(G) = \frac{1}{|\mathcal{G}|}$ is a constant. Thus

$$\begin{aligned} p(G|Z) &= \frac{p(Z|G)}{\sum_{G' \in \mathcal{G}} p(Z|G')} \\ &= \frac{J_G(\delta, n, D, U)}{\sum_{G' \in \mathcal{G}} J_{G'}(\delta, n, D, U)} \end{aligned} \tag{2.2.8}$$

where

$$J_G(\delta, n, D, U) = \frac{I_G(\delta + n, D + U)}{I_G(\delta, D)}. \tag{2.2.9}$$

Given a G that is decomposable, there is a perfect ordering of cliques and separators: $(C_1, \ldots, C_n)$ and $(S_2, \ldots, S_n)$, and so

$$p(G|Z) = \frac{\prod_{i=1}^{k} P(G_{C_i}|Z)}{\prod_{i=2}^{k} P(G_{S_i}|Z)}. \tag{2.2.10}$$

If G is complete or if considering G_{C_i} or G_{S_i} (since they are complete) then it follows that per Atay-Kayis and Massam [2005], $f(\Sigma^{-1}|G)$ has Wishart density $W(\delta, D)$ with

$$I_G(\delta, D) = \frac{2^{\frac{(\delta+p-1)p}{2}} \Gamma_p(\frac{\delta+p-1}{2})}{\det(D)^{\frac{\delta+p-1}{2}}}. \tag{2.2.11}$$

Thus

$$I_G(\delta, D) = \frac{\prod_{i=1}^{k} I_{G_{C_i}}(\delta, D_{C_j})}{\prod_{i=2}^{k} I_{G_{S_i}}(\delta, D_{S_i})} \tag{2.2.12}$$

which is closely related to the hyper-inverse-Wishart distribution [Dawid and Lauritzen, 1993]. The posterior probabilities for each G can finally be calculated by substituting (2.2.11) into (2.2.9), and then (2.2.9) into (2.2.8). Note, that $2^{\frac{(\delta+p-1)p}{2}}$ (c.f. Roverato [2002]) in (2.2.11) is the correct parameter as opposed to $2^{\frac{np}{2}}$ (c.f. Atay-Kayis and Massam [2005]). If G is not decomposable then there is no closed-form expression of the normalising constants — thus a numerical approach is required as detailed in Atay-Kayis and Massam [2005, Sections 4–5].

In practical terms, this means that the relative posterior probabilities of graphs can be calculated

in order to identify those which appear to fit the data well, and if desired an optimum graph can be selected. However, if G also incorporates latent variables then the challenge of model search is more complex. In this thesis the main results concern whether the class of latent tree models is suitable for a given set of data, though we do also provide methodology for assessing specific G as well. We are able to draw upon the Wishart and inverse-Wishart distributions to develop similar Monte-Carlo simulation techniques as above for the GLTM.

Chapter 3

Quantitative linguistics: a functional data perspective

This chapter will be used to give a background to the statistical work in linguistics particularly from an acoustic perspective (e.g. Aston et al. [2012], Bouchard-Côté et al. [2013]). The main data set of the thesis shall be described from both a statistical and acoustic perspective. This section will outline the preprocessing which has been applied to the audio data in order to obtain the spectrograms for use in the analyses. The content of this chapter is quite different to that of Chapter 2, but both are required for the type of applications we wish to perform later on.

This chapter will go on to describe the details of the main tools that will be utilised for the linguistic application. This will include presenting both functional and multivariate versions of techniques, covering tools such as PCA and FPCA, and the less often implemented technique of CVA and the functional counterpart CFA. Finally, a novel contribution of the thesis is presented: separable covariance versions of CVA and CFA denoted separable-CVA and separable-CFA respectively. These form highly effective and practically useful tools that overcome the common problem of observation-poor, variable-rich data sets (i.e. small n, large p problems).

3.1 Selected history of quantitative linguistics

Linguistics is the scientific study of languages. Under this broad definition it can be argued that there is evidence of linguistics being practised in the time of antiquity in locations such as Greece, India and China, and then developing in Arabia in the Middle Ages [Allan, 2013, Syal and Jindal, 2007]. However, the development of modern linguistics is not considered to have begun until the late 18th and early 19th centuries. Some notable scholars in the Western world were Schlegel and Bopp who both pioneered the concept of comparative grammar, and Schleicher who led on the idea of reconstructing proto-languages [Koerner, 1999]. In the 20th century Saussure is considered to have been one of the most influential thinkers on linguistics. Two notable contributions were his principle of studying languages synchronically (at a fixed time) rather than diachronically (historically) and his structuralist approach to linguistics which significantly changed the way that languages were studied [Malmkjaer, 2009]. While general linguistic theory evolved, quantitative and statistical linguistics also began to develop. This started in earnest a few decades after modern statistics had begun to flourish and the direction was generally guided by wider linguistic research interests.

Some of the earliest quantitative work was by Ernst Förstemann in 1852 who measured how similar languages were based on grammatical and phonological (sound) characteristics [Těšitelová, 1992]. Sampson [2003] discusses other early quantitative work in the branch of linguistics called stylometry whereby the style of an individual's writing was quantified in order to assess authenticity of literary works or analyse stylistic change over an author's career. Without a quantitative approach these analyses would not detect subtle linguistic features. A well-known result that was applied to textual analysis is Zipf's Law [Lüdeling, 2009, Chapter 3] which in theory governs the relative frequency of words in a corpus. Under mild assumptions, Zipf's Law is asymptotically equivalent to Herdan's Law that relates the length of a text to the number of unique words [Egghe, 2007, Herdan, 1960].

In 1913, Markov gave the first use of Markov chains which just happened to be an application to linguistics. In his analysis of poems by Pushkin he had noticed that the proportion of vowels and consonants in a word changed based upon the location of the letter (e.g. beginning or end of a word). Furthermore, he noted that these proportions were heavily affected by whether the

preceding letter was a vowel or consonant and so could be modelled by what is now known as a Markov chain [Basharin et al., 2004]. Since then Markov chains and hidden Markov models have been used for more advanced purposes such as automatic part-of-speech identification (e.g. verbs, nouns, adjectives) with high success rates (e.g. Kupiec [1992]).

Relationships between languages have long been described as phylogenetic trees constructed using linguistic factors (e.g. Schleicher [1860]) where all non-leaf variables are unobserved and represent features of the past languages before their divergence. Greenberg [1954] developed some of the first quantitative methods that were used to investigate evolutionary relationships between languages. Other tree reconstruction methods have used cladistic techniques [Platnick and Cameron, 1977, Rexová et al., 2003]. As with cladistic techniques that were in part borrowed from biology, cluster-analysis techniques from population genetics were applied to grammatical structures by Nichols [1992] to investigate macro migration patterns. Even more recently there have been large-scale attempts to reconstruct trees or networks of languages using statistical methods (e.g. Nakhleh et al. [2005] for the Indo-European language family, Nicholls and Ryder [2011] for the Semitic language family).

The above has only touched upon the range of quantitative methods with many other mathematical and statistical tools being utilised for analysis across the study of linguistics. For example, corpus linguistics makes extensive use of quantitative tools an statistical summaries to analyse and summarise real-world texts (see Kučera and Francis [1967] for an early modern work, and Gries [2009], Oakes [1998] for overviews of more recent work). Some more thorough surveys of quantitative linguistics can be found in Köhler et al. [2005], Sampson [2003], Těšitelová [1992]. The use of quantitative and moreover statistical techniques for linguistic inference has made it possible to get additional insights and depth of understanding regarding questions of linguistic interest. Such tools alone are no substitute for linguistic theory but together are robust.

In this thesis, the statistical linguistic analyses are notable on three counts: Firstly, the data set is acoustic, taking the form of audio recordings of speakers of different languages. Secondly, these audio recordings are treated as functions rather than vectors or other forms. Thirdly, these functional audio recordings are used to compare languages in a phylogenetic context. The treatment of acoustic data as functional in a linguistic context is relatively new and its application to phylogenetic language studies even more so. This approach provides opportunities to study

and identify features of languages that might otherwise be obscured and to use an array of tools that are not available when considering some traditional quantitative analyses. In the next section we introduce the concept of functional data and consider its existing use in a linguistic context.

3.2 Functional data analysis

Functional data analysis (FDA) is a statistical approach whereby the statistical units of interest are functions defined on a continuum. Often these objects are curves or surfaces but equivalent objects can be studied in higher dimensions. There is usually a stipulation that the underlying functions are assumed to be smooth to the extent that derivatives of the functions exist up to a given order. For example, the existence of first and second order derivatives reflects the belief that data is suitably smooth but also has practical implications by enabling functional data tools that require derivatives (e.g. roughness penalisation) to be implemented. FDA can also be viewed as a specific type of object oriented data analysis whereby the objects of interest are smooth continuous functions [Wang et al., 2007].

3.2.1 Basis functions

In practice we record discretised observations even when we believe that there is an underlying continuous function. In addition to the discretisation these observations usually include added noise. Thus, often one of the first steps in an FDA is to represent the data by basis functions.

Let us first consider the theoretical setting whereby the function $x(t)$ can be expressed in terms of a linearly weighted (by c) sum of functions $\phi(t)$ known collectively as a basis

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

and where K is potentially ∞. In practice, by selecting a sensible set of basis functions and sufficiently large K we can suitably approximate infinite dimensional functional data by a finite number of basis functions. That is:

$$x(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$$

for finite K. Now considering the case of discretised data, the aim is to find such a finite K and an appropriate basis that provides a good fit for the data.

There are many classical choices for basis functions such as the monomial power series and the Fourier series. To illustrate the use of basis function approximations, data was simulated by first evaluating the polynomial $x(t) = t^3 - 5t^2 + 2t + 4$ at 100 equally spaced values $t \in \{-2, -1.9192 \ldots, \ldots, 6\}$. At each of these 100 points some Gaussian noise was added to $x(t)$ generated independently from $\varepsilon \sim N(0, 10)$ to obtain the 100 observations of the curve $\tilde{x}(t) = x(t) + \varepsilon$.

In Figure 3.1 we illustrate the third-order monomial basis $(1, t, t^2, t^3)$ that provides the least-squares fit to the simulated data. Similarly in Figure 3.2, a second-order Fourier basis $(1, \sin(\omega t), \cos(\omega t), \sin(\omega 2t), \cos(\omega 2t))$ has been fitted to the same data. In both instances there is a good fit to the true curve, which is to be expected given the data generating process is quite simple. The key point is that a 100 observation data set has been suitably summarised by 4 and 5 basis functions respectively which is an efficient description of the data. Furthermore, the functional description matches the belief that the data is obtained from an underlying curve.

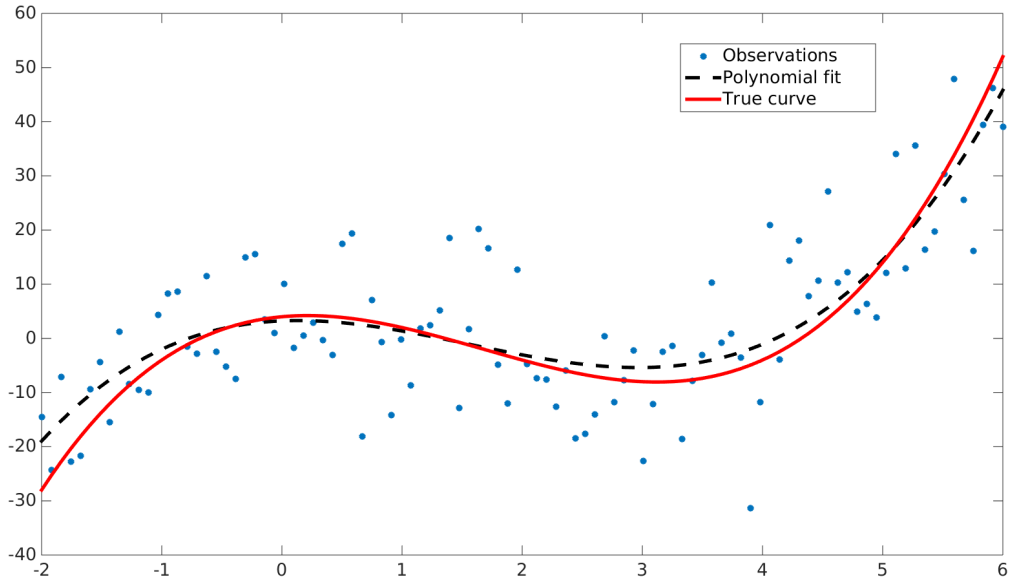


FIGURE 3.1: Least-squares fit of a third-order polynomial basis to simulated data.

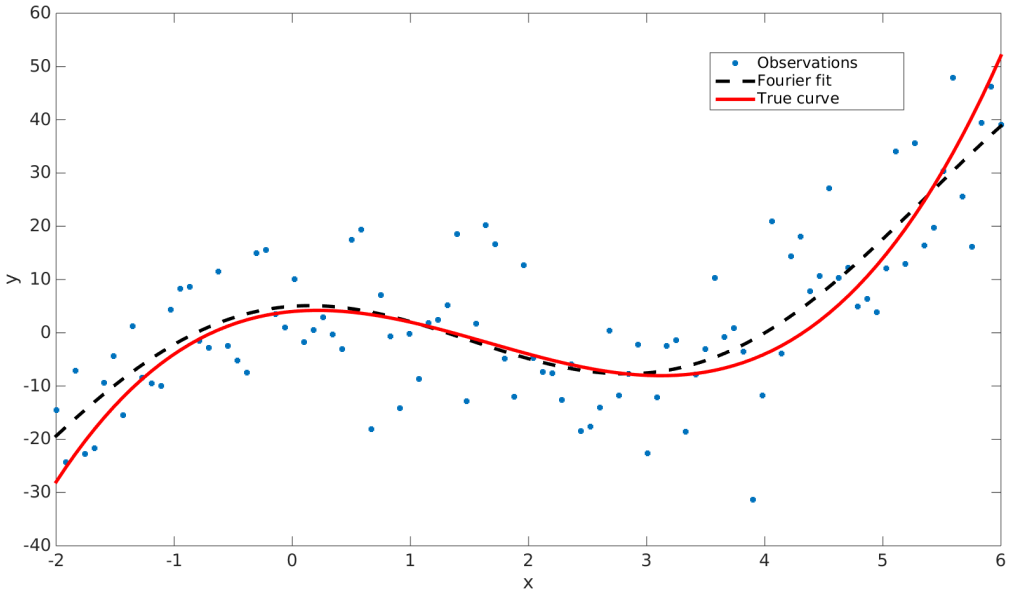


FIGURE 3.2: Least-squares fit of a second-order Fourier basis to simulated data.

3.2.2 Spline basis

A more advanced and flexible basis is based on spline functions. A spline is constructed by dividing a function's domain into subintervals and within each subinterval approximating the function with an order m polynomial (where $m - 1$ is thus the polynomial degree). The end points of the range of each interval, $\tau_0, \ldots, \tau_L$ say, are called breakpoints, and polynomials within adjacent subintervals are constrained to match at these breakpoints. Furthermore, up to order $m - 2$ derivative functions must also observe these constraints. Spline functions are currently one of the most popular bases with the B-spline [de Boor, 2001] often being the preferred variant. In general, spline functions have the advantage of being more flexible than a generic polynomial basis (sometimes even on very few component functions), and importantly, significantly faster to compute.

The spline function based on B-splines is constructed as follows:

$$S(t) = \sum_{i=1}^{n} c_i B_{i,m-1}(t)$$

where c_i are B-spline coefficients or control points which are used to guide the curves. In general the j^{th} B-spline of degree d (say) is found recursively as

$$B_{j,d}(t) = \frac{t - \tau_j}{\tau_{j+d} - \tau_j} B_{j,d-1}(t) + \frac{\tau_{j+1+d} - t}{\tau_{j+1+d} - \tau_{j+1}} B_{j+1,d-1}(t)$$

which is initialised by

$$B_{j,0}(t) = \begin{cases} 1 & \text{if } \tau_j \leq t < \tau_{j+1} \\ 0 & \text{if otherwise.} \end{cases}$$

Further details can be found in de Boor [2001]. From a practical point of view there are many pieces of software for implementing B-spline fitting often based on the objective function of minimising squared distance. In Section 7.2.1 we use the `MATLAB` package `csaps` to fit a cubic spline to some biological growth curves. In fact, we go one step further than B-splines by incorporating a smoothing parameter. These smoothed splines are known as penalised B-splines, or P-Splines [Eilers and Marx, 1996] and allow more deviation from the data than a pure least-squares fit meaning the resulting spline functions are smoother. The effect of different levels of smoothing are considered in Section 7.2.1 and displayed graphically in Figure 7.25.

3.2.3 Use of functional data in statistical linguistics

As interest in FDA has increased, so has the range of applications (e.g. brain imaging [Sørensen et al., 2013], climatology [Besse et al., 2000], and medical research [Ratcliffe et al., 2002]). This has been partly facilitated by better access to functional data, a wider range of FDA tools [Horváth and Kokoszka, 2012, Ramsay and Silverman, 2005] and through greater availability of computational power for analyses. The use of FDA in statistical phonetics has recently attracted attention (e.g. Koenig et al. [2008], Mooshammer [2010]). Such analyses, which involve acoustic functional data, have provided particularly promising and interesting results in a diverse range of settings. The acoustic structure of spoken words can be used to investigate areas of linguistic interest in a similar way that orthographic representations of speech have been utilised. For example, Aston et al. [2010] investigate Qiang, a Sino-Tibetan language and Grabe et al. [2007] use a polynomial basis expansion to examine pitch variation in English.

In more detail, the data set in Aston et al. [2010] consists of recordings of 8 speakers of Luobuzhai Qiang speaking 19 different nouns. The phonetic feature of interest in the study is known as the fundamental frequency or F0 which can broadly be thought of as pitch. Many previous studies of F0 have relied upon point estimates of recordings but this loses a lot of the information available in the full pitch contour. To overcome this the pitch contours for each of the speakers and words is modelled using FPCA (see Section 3.4.2) so that a large amount of the pitch variability can be encoded using a small number of basis functions. In practice, the pitch contours for each syllable were recorded at eleven equidistant time points and these were then used with the FPCA. A projection of the data to three dimensions was deemed sufficient and the FPC scores (projections) were then treated as the response for a linear mixed effects model with speaker and word characteristics. The results of which found previously unidentified gender differences amongst speakers. The use of a functional data approach allowed the complex pitch contours to be modelled by low order functions. Moreover, by associating FPC scores with meaningful phonetic covariates the analysis provides an interpretable linguistic conclusion.

In Grabe et al. [2007], once again F0 is treated as functional data and in this instance modelled using a third-order polynomial basis. The data set consisted of 710 sentences spoken by 42 speakers of 7 English dialects but the feature of interest is the intonation of the 7 accents that are found across these 7 dialects. For each accent, the 4 coefficients associated with each third-order fitted basis were identified. The aim was to investigate whether the accents could be distinguished by the coefficient summary. By implementing a multivariate analysis of variance on the coefficients associated with each accent, 19 of the 21 pairwise accent comparisons were found to be statistically significant. This raises possibilities of utilising accent and intonation more readily in speech technology, demonstrating the usefulness of using a functional basis representation of acoustic data for identifying particular phonetic features.

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 [Bolnick et al., 2004].

3.3 The acoustic functional data set

The core application of the thesis is an investigation into the conditional independence relationships between 5 Romance languages. The main analysis is given in Section 7.1 though the description of the data set is given here along with the preprocessing steps. The data set comprises audio recordings originating from speakers of one of five different Romance languages: French, Italian, Portuguese, Spanish (American), and Spanish (Iberian) — while two dialects of Spanish are being used in this study, they are treated as different spoken languages in this analysis as the interest is in pronunciation rather than textual representation, the difference between "dialect" and "language" being a matter of degree of difference rather than an absolute quantitative difference. Each recording is of some individual saying an integer from 'one' to 'ten' in their particular language. In total there are 219 word recordings and each can be classified by the language, the gender of the speaker and the number being spoken. Observations of the same word being spoken in different languages are treated as sharing the same word attribute. For example the word 'four' includes recordings of 'quatre' (French) and 'quattro' (Italian) as well the word 'four' in other languages. Integers were chosen because these have no ambiguity in terms of translation making comparison of their use across languages straightforward. Furthermore, the cardinals 'one' to 'ten' of Romance languages (among many other words) stem from shared Latin forms [Price, 1992]. This suggests that these words might also be suitable when comparing languages acoustically.

As mentioned, the observations are modelled as functional data as is becoming increasingly common in studies involving sound recordings (e.g. Holan et al. [2010]). Such models make the reasonable assumption that the data have been obtained by observing an underlying function at finitely many discrete points along a continuum, and that this underlying function is smooth (i.e. a certain number of derivatives exist).

3.3.1 Romance data set pre-processing

The data set used in the final analysis had already been preprocessed with the full description given in Hadjipantelis [2013, Chapter 6]. The original acoustic data set originated from a number of sources and the specifications of the recordings differed across these sources. Therefore, the audio recordings were resampled at a rate of 16000 samples per second to make the observations comparable for processing. A short-time (10ms window) Fourier transform was taken of each audio recording to produce a spectrogram. A spectrogram is a two-dimensional representation of audio signal energy intensity in frequency-time space [Fulop, 2011]. Spectrograms are a natural choice for representing power with functional data [Holan et al., 2010, Martinez et al., 2013], though approaches such as Mel-frequency cepstra can provide possible alternative representations [Davis and Mermelstein, 1980]. The value stored at a frequency-time point is a function of the power (or amplitude). A $10\log_{10}(\cdot)$ transformation of the original power was taken so that units of power are decibels.

In Holan et al. [2010], spectrograms of mating calls are used as predictors of mating success of treehoppers. Martinez et al. [2013] investigate regional differences in bat chirps by considering a functional mixed model with spectrograms as the image response. In contrast, the emphasis of this analysis will not be to seek a model that acts as the data generating process. Instead it aims to identify meaningful low dimensional representations of spectrograms that highlight differences between languages, and subsequently assess whether these distinctive acoustic features are compatible with the class of GLTM.

Frequencies were binned every 100Hz up to the Nyquist frequency of 8000Hz. The resulting spectrograms were stored as matrices of 81 frequency by 100 time points. These spectrograms were still distorted in two mains ways: firstly, the data was undoubtedly noisy (amplitude distortions) and secondly there were phase distortions. The amplitude distortion is a common feature of many data sets and can be considered as an error term having been added to the power recording at every frequency-time point. The time distortion was the result of the overall duration of a word varying significantly per speaker and furthermore the timings of intra-word elements (for instance syllables). To adjust for amplitude distortion the spectrograms underwent a smoothing algorithm aimed at removing noise. This is consistent with the smoothness

assumption inherent to the functional data framework. A penalised least squares filtering approach was used to smooth the data. Roughness was penalised using second-order difference and the unsmoothed data underwent a discrete cosine transformation following the algorithm in Garcia [2010]. Having smoothed the data, the remaining unadjusted distortion was in the time dimension. The available techniques to deal with differences in the phase of curves are known as curve alignment, curve registration or warping (see Lucero and Koenig [2000], Ramsay and Silverman [2005]). The method used on this data set was based on the pairwise synchronisation as described in Tang and Müller [2008] with adjustments for the two dimensional nature of the data. Although the warping of the spectrograms was only occurring in the time dimension, the frequency information was required for calculating discrepancy between spectrograms. These time-phase adjustments were performed on a word-gender basis as there are known differences in frequency ranges spoken by male and female speakers. As part of this process, the word durations were all standardised to have the same arbitrary length. In our analysis we refer to this as "standardised time" across a range of 0 to 100. Figure 3.3 is a spectrogram (post pre-processing) of a female French speaker saying the word 'quatre'. Broadly, this interpolated plot indicates that there is greater power in the lower frequencies, and that the beginning and the end portions of the standardised time period are quieter.

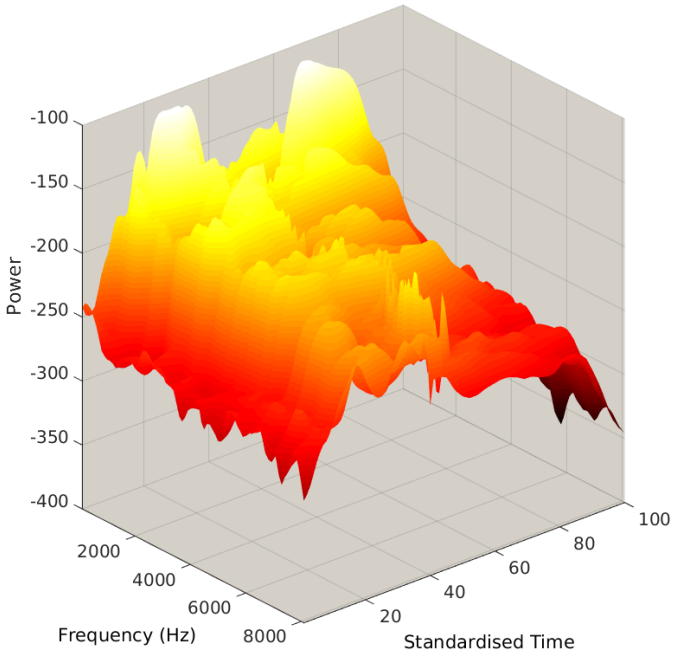


FIGURE 3.3: Post-registration spectrogram of female French speaker saying 'quatre'. It can be seen that there is greater power in the lower frequencies, and that the very beginning and end of the word are unsurprisingly two of the quietest regions.

3.3.2 Notation

The underlying function of each spectrogram is denoted $x^{d,g}_{l,m}(f,t)$ with the two dimensions f and t referring to frequency and time respectively. Recall that each spectrogram is derived from a spoken word — the subscripts and superscripts encode observational information: $l = 1,\dots,n_l$ denotes the language being spoken; $d = 1,\dots,n_d$ indicates the word being spoken; $m = 1,\dots,m_{ld}$ is a counter where m_{ld} is the number of observations of word d from language l; g refers to the gender of the speaker.

It is well documented that there are differences in the acoustics of male and female speakers which go beyond a simple shift in the spoken frequencies (for instance Nittrouer et al. [1990], Pépiot [2013]). Parris and Carey [1996] present a statistical method for discriminating between speaker gender of short acoustic recordings. In their analysis of seven Indo-European languages (of which Romance is a subset), gender was correctly identified on average 98% of the time. This suggests that there are commonalities in acoustic gender differences across Indo-European languages. In light of this result, it is judged that gender should be adjusted for at the macro level:

$$x^{d}_{l,m}(f,t) = x^{d,g}_{l,m}(f,t) + \tilde{x}^{g}(f,t)$$

where $\tilde{x}^g$ is the difference between the mean of all samples with gender g and the mean of all samples. Henceforth it will be the gender adjusted function that will be the object of interest in the thesis.

The mean spectrograms for language l, word d are defined in (3.3.1), for language l in (3.3.2), and the grand mean spectrogram in (3.3.3).

$$\bar{x}^{d}_{l}(f,t) = \frac{1}{m_{ld}} \sum_{m=1}^{m_{ld}} x^{d}_{l,m}(f,t) \tag{3.3.1}$$

$$\bar{x}_{l}(f,t) = \frac{1}{m_{l\bullet}} \sum_{d=1}^{n_d} m_{ld} \bar{x}^{d}_{l}(f,t) \tag{3.3.2}$$

$$\bar{x}(f,t) = \frac{1}{m_{\bullet\bullet}} \sum_{l=1}^{n_l} m_{l\bullet} \bar{x}_{l}(f,t) \tag{3.3.3}$$

where $m_{l\bullet} = \sum_{d=1}^{n_d} m_{ld}$, $m_{\bullet\bullet} = n = \sum_{l=1}^{n_l} m_{l\bullet}$, and for $t \in \mathcal{T}$, $f \in \mathcal{F}$. The parameters $m_{\bullet\bullet}$ and n will be used interchangeably depending on whether summation is being emphasised.

To get a further feel for the data, we plot the mean spectrograms by gender (Figure 3.4). Note that although it appears that the higher frequencies have more power for males, this is not suggesting males speak at a higher pitch, just that the male recordings tend to be louder in general. Recall that this gender effect is adjusted for throughout the rest of the analyses.

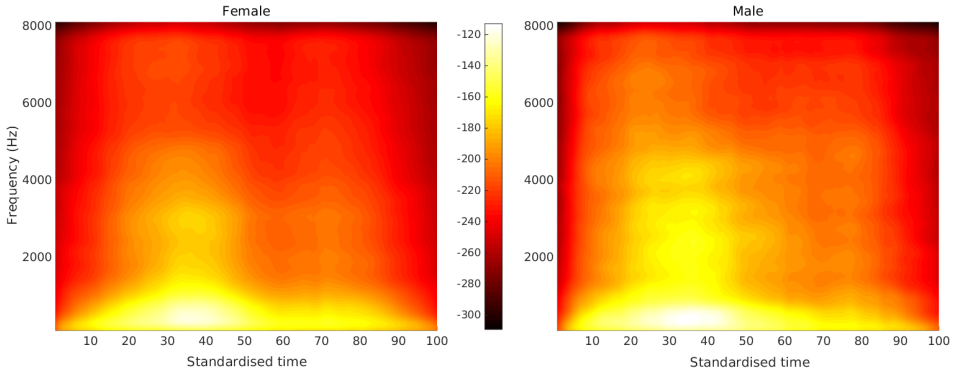


FIGURE 3.4: Mean spectrograms by gender.

For even more detail, we plot the mean spectrograms for each language-number combination (Figure 3.5). Observe that it can be seen that certain words have two clear syllables whereas others just one (e.g. the number seven: French “sept” versus Italian “sette”).

3.4 Functional data tools

The development of a functional data framework has demanded functional tools equivalent to those used in other areas of data analysis. The theory of these tools is rooted in the functional context, yet in practice it is often necessary for numerical or discrete approximations to be used and thus multivariate tools often have a role to play in FDA. We now introduce two functional tools that project data to an alternative basis often with the aim of reducing dimensionality while retaining the majority of the relevant variability in the data. The multivariate counterparts are discussed in Section 3.5.

3.4.1 Group based projections of functional data

Dimension reduction is a well-studied area of statistics with tools such as principal component analysis (PCA) and multidimensional scaling (MDS) (e.g. see Cox and Cox [2010], Jolliffe

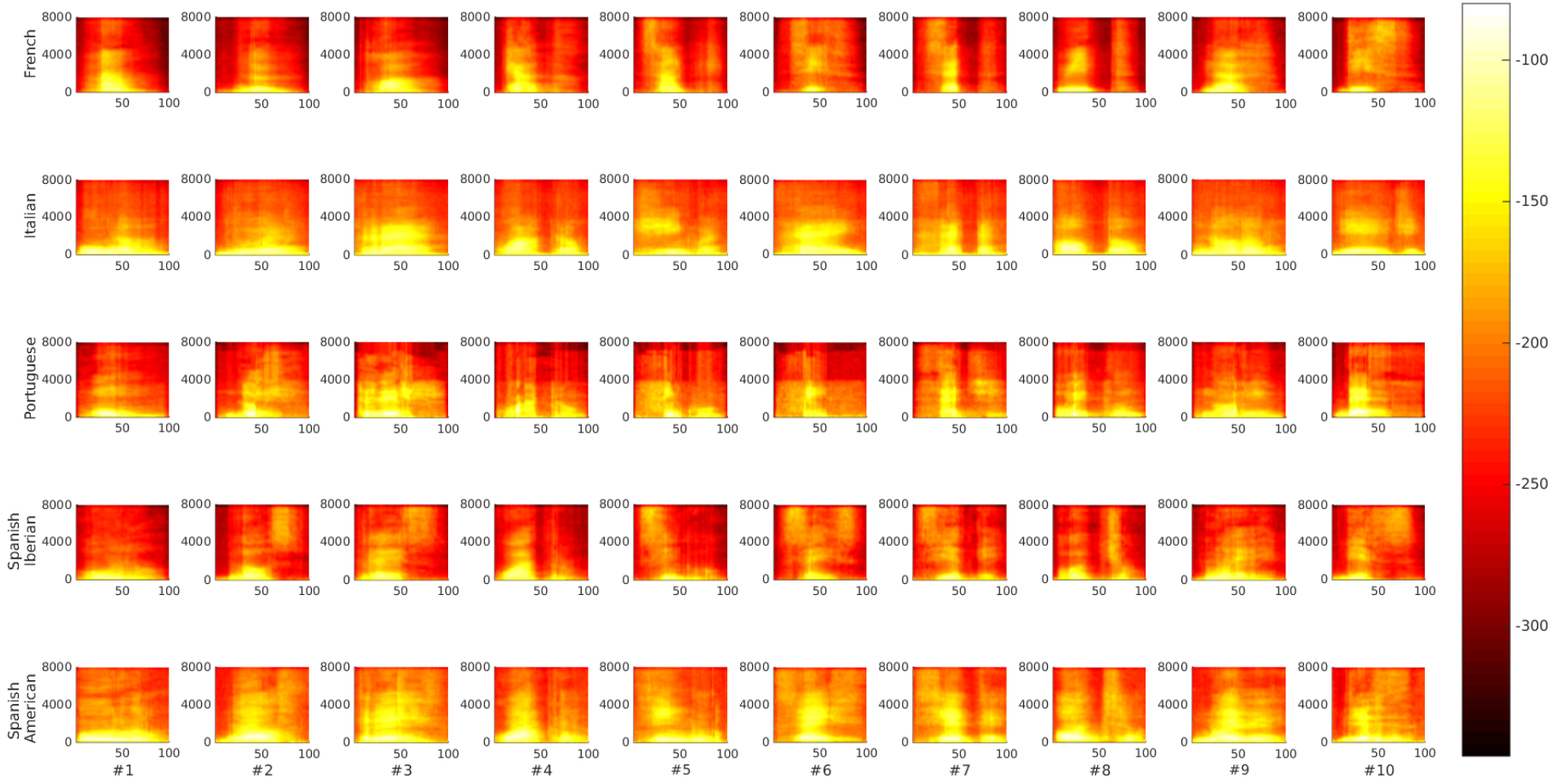


FIGURE 3.5: Mean spectrograms on a grid with rows representing languages and columns representing numbers. As with other plots vertical axes are frequency (Hz), horizontal axes are standardised time and colour represents power.

[2002]) having widespread use. Functional counterparts of such techniques have also been formulated, for example functional principal component analysis (FPCA) Castro et al. [1986], Rice and Silverman [1991], Yao and Lee [2006], and functional MDS Mizuta [2006].

The Romance data set presented as the main application of the thesis can benefit from a dimension reduction in two main ways. First and foremost, dimension reduction provides a route to feature extraction whilst also reducing unwanted noise. Second, if subsequent to the reduction it is found that $n \geq p$ then techniques which make use of inverse covariances can be implemented straightforwardly. If this is not so, standard estimates will produce singular sample covariance matrices. Of course these benefits must be balanced against potential information loss from the data reduction. One approach to feature extraction that mitigates against this loss is to find an ordered basis which prioritises one or more characteristics of interest. Thus by projecting data onto the first few components of such a basis the most prominent aspects of the data are retained whilst what remains is treated as noise. In the cases of PCA and FPCA, the dimension reduction is optimised so as to efficiently capture modes of variation. Such techniques are often used in linguistic and semantic analyses, for example Lee et al. [2001], Wenyin et al. [2010]. However, as our focus is on macro-language comparisons, it can be argued that the feature of interest is the between- to within-language covariance, and it is this which should directly inform the method selected to construct a basis. When data is known a priori to be grouped then CFA and its multivariate analogue CVA are standard techniques implemented to select variables to discriminate between groups. These tools are therefore the starting points for our analyses.

3.4.2 Functional principal component analysis

The aim of FPCA is to identify dominant directions of variability for functional data. These directions are known as functional principal components and a projection to the first r basis functions (ordered by variability) produces an efficient representation of the original (centred) data in a lower dimensional space. More precisely, define $X(t)$, $t \in T$ to be a random function whereby $\int_T |X(t)|^2 dt < \infty$, that is $X(t)$ is almost surely square-integrable. The mean function is given by

$$\mu(t) = \mathbb{E}(X(t))$$

with associated covariance function

$$C(t,u) = \text{cov}(X(t), X(u)) = \sum_{k=1}^{\infty} \lambda_k \psi_k(t) \psi_k(u) \tag{3.4.1}$$

where $u, t \in T$, and λ and ψ are the eigenvalues and eigenfunctions of the orthonormal basis associated with $C(t,u)$. Denote the centred function

$$\tilde{X}(t) = X(t) - \mu(t) = \sum_{k=1}^{\infty} \xi_k \psi_k(t)$$

where

$$\xi_k = \int_T \tilde{X}(t) \psi_k(t) dt$$

is the k^{th} functional principal component and is subject to the following constraints:

$$\int_T \psi_k^2(t) dt = 1$$

and

$$\int_T \psi_k(t) \psi_j(t) dt = 0 \quad \forall j = 1, \ldots, k-1$$

i.e. the eigenfunctions are orthonormal and mutually orthogonal.

By the Karhunen–Loève theorem (e.g. Lindgren [2012, Chapter 5]), if the domain T is finite on a range $[a,b]$ say, then FPCA produces the optimal linear basis expansion with respect of minimising mean-squared error (MSE). Thus, for the FPCA basis ordered decreasingly by eigenvalue, the truncation of this basis captures more of the variability in the data than any other set of basis functions. This result has been proven in many ways [Algazi, 1969, Brown, 1960, Kramer, 1960], here we provide a sketch proof.

Sketch proof.

$$\epsilon_N(t) = \sum_{j=N+1}^{\infty} \xi_j \psi(t)$$

Thus the mean squared error associated with a truncation to N basis functions is given by

$$\begin{aligned}\mathbb{E}(\epsilon_N^2(t)) &= \mathbb{E}\big(\sum_{i=N+1}^{\infty}\sum_{j=N+1}^{\infty}\xi_i\xi_j\psi_i(t)\psi_j(t)\big)\\ &= \sum_{i=N+1}^{\infty}\sum_{j=N+1}^{\infty}\mathbb{E}\big(\int_a^b\int_a^b\tilde{X}(r)\tilde{X}(s)\psi_i(r)\psi_j(s)drds\big)\psi_i(t)\psi_j(t)\\ &= \sum_{i=N+1}^{\infty}\sum_{j=N+1}^{\infty}\psi_i(t)\psi_j(t)\int_a^b\int_a^b C(r,s)\psi_i(r)\psi_j(s)drds\end{aligned}$$

Recall that ψ functions are orthonormal and so

$$\int_a^b\epsilon_N^2(t)dt = \sum_{j=N+1}^{\infty}\int_a^b\int_a^b C(r,s)\psi_j(r)\psi_j(s)drds$$

To minimise total MSE subject to mutual orthogonality of ψ functions, a Lagrangian multiplier β_j can be introduced. Lagrangian multiplier techniques utilise partial differentiation to find maxima and minima of functions subject to constraints. For an overview of the subject see a text such as Bertsekas and Rheinboldt [2014].

For the total MSE, by setting the derivative with respect to $\psi_j(s)$ to 0, solutions to the Lagrangian functions are given by

$$\int_a^b C(r,s)\psi_i(r)ds = \beta_i\psi_i(s)$$

which gives the required result. □

Corollary 3.4.1. *From Brown [1960], the minimum MSE for basis expansion of length n is given explicitly as:*

$$\epsilon_N^2(t) = \int_a^b\mathbb{E}(\tilde{X}^2(t))dt - \sum_{j=1}^{N}\frac{1}{\lambda_j}$$

It is clear that FPCA is a powerful tool for highlighting the dominant modes of variability but, as a consequence, also for efficient dimension reduction. To implement FPCA a numerical approximation is required. In Section 3.5.2 PCA is presented as the multivariate counterpart to FPCA. However, beyond that PCA can be used as an approximation to FPCA by discretising the functional data.

3.4.3 Canonical function analysis

Here we present CFA as a tool for FDA to produce a basis which maximises between- to within-group variation (subject to the basis component functions being orthogonal) with the intention of achieving an efficient dimension reduction. The first detailed account of CFA in the literature is given in Kiiveri [1992]. Here we present the fundamentals of the tool.

The aim of CFA is to identify canonical functions $f_q(u)$ such that between-group variation is maximised relative to within-group variation under the restriction that each canonical function is uncorrelated to every other.

The functional covariance operators required for CFA are given by:

$$B(t,u) = \frac{1}{g-1}\sum_{i=1}^{g} n_i(m_i(t) - \bar{m}(t))(m_i(u) - \bar{m}(u)) \tag{3.4.2}$$

$$W(t,u) = \frac{1}{n-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}(t) - m_i(t))(x_{ij}(u) - m_i(u)) \tag{3.4.3}$$

where in (3.4.2)

$$\bar{m}(t) = \sum_{i=1}^{g}\frac{n_i m_i(t)}{n}.$$

Additionally (3.4.2) and (3.4.3) are assumed to be bounded and piecewise continuous.

CFA was first motivated as an approximation to CVA (see Section 3.5.2) and thus the optimisation problem is derived from the optimality equation of CVA and be seen as a limiting form of the discrete case. Thus for CFA we are looking for solutions to (3.4.4) which has the equivalent functional form of the CVA eigenanalysis, c.f. (3.5.7).

$$\int_c^e (B(t,u) - \mu_q W(t,u)) f_q(u) du = 0 \tag{3.4.4}$$

with $B(t,u)$ and $W(t,u)$ defined in (3.4.2) and (3.4.3) respectively and whereby μ_q and f_q are eigenvalues and eigenfunctions respectively. To obtain subsequent solutions that are orthogonal to one another (3.4.5) is maximised subject to the restriction in (3.4.6):

$$\int_c^e \int_c^e f_q(t) B(t,u) f_q(u) dt du \tag{3.4.5}$$

$$\int_c^e \int_c^e f_{q_1}(t) W(t,u) f_{q_2}(u) dt du = \delta_{q_1 q_2} \tag{3.4.6}$$

where $\delta_{q_1 q_2}$ is the standard Kronecker delta:

$$\delta_{q_1 q_2} = \begin{cases} 1 & q_1 = q_2 \\ 0 & q_1 \neq q_2 \end{cases}$$

Note that (3.4.5) and (3.4.6) together lead to he derivation of (eq:optfunc2) so in words CFA is looking for the $f_q(u)$ that maximises between-group to within-group variation under the orthogonality conditions. We denote as CFA pairs $(f_1(t), \mu_1), \ldots, (f_s(t), \mu_s), \ldots$ which solve (3.4.4). There may be countably infinite solutions to this equation but only a maximum of s will have non-zero μ_q.

In the discretised estimation, only a maximum of s (say) will have non-zero μ_q. Pairs of canonical functions and real numbers $(h_1(t), \mu_1), \ldots, (h_s(t), \mu_s)$ can be found by solving (3.4.4) numerically, where $\mu_1, \ldots, \mu_s$ is a monotone decreasing sequence. These are solutions to the generalised eigen-equation. Furthermore, an r-dimensional projection of the data is obtained using the first r canonical functions ($r \leq s$), and this projection is such that the between- to within-group covariance is maximally retained.

3.5 Multivariate analysis

We now discuss the multivariate PCA and CVA, which are the analogues of FPCA and CFA. These multivariate equivalents are not only useful in a multivariate framework but can also be used to provide approximate solutions to their functional counterparts.

3.5.1 Principal component analysis

As with FPCA, the aim of PCA is to find the direction of maximum variability of a particular data set. With PCA the data is usually multivariate or a discretised approximation to functional data which in practice can be treated in the same way as multivariate data.

Given a centred $n \times p$ data matrix $\boldsymbol{X}$ with n observations and p variables we seek a vector $\boldsymbol{a}^T$ of length p such that the linear projection $\boldsymbol{y} = \boldsymbol{a}\boldsymbol{X}^T$ retains the maximum amount of variability possible in the data. To obtain the $\boldsymbol{a}$ that satisfies this property consider the covariance of the projected data.

$$\begin{aligned}
\mathbb{E}((\boldsymbol{y}^T - \mathbb{E}(\boldsymbol{y}^T))^2) &= \mathbb{E}((\boldsymbol{a}^T\boldsymbol{X} - \mathbb{E}((\boldsymbol{a}^T\boldsymbol{X}))^2) \\
&= \mathbb{E}(\boldsymbol{a}^2(\boldsymbol{X} - \mathbb{E}(\boldsymbol{X}))^2) \\
&= \boldsymbol{a}\mathbb{E}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{a}^T \\
&= \boldsymbol{a}\,\mathrm{cov}(\boldsymbol{X})\boldsymbol{a}^T \\
&= \boldsymbol{a}\Sigma_X\boldsymbol{a}^T
\end{aligned}$$

The $\boldsymbol{a}$ that maximises $\boldsymbol{a}\,\mathrm{cov}(\boldsymbol{X})\boldsymbol{a}^T$ subject to $\boldsymbol{a}\boldsymbol{a}^T = 1$ can be obtained using a Lagrange multiplier λ. This leads to re-expressing the optimisation problem as $\Sigma_X\boldsymbol{a}^T = \lambda\boldsymbol{a}^T$ which implies that $(\Sigma_X - \lambda\boldsymbol{I})\boldsymbol{a}^T = \boldsymbol{0}$. Thus we seek non-trivial solutions (i.e. $\boldsymbol{a} \neq \boldsymbol{0}$) to $\det(\Sigma_X - \lambda\boldsymbol{I}) = 0$. The solutions are eigenvectors of the sample covariance Σ_X and λ the corresponding eigenvalues. To determine which of the solutions provides the direction of greatest variance, note that $\boldsymbol{a}\Sigma_X\boldsymbol{a}^T = \lambda$ and thus λ represents the sample variance. It then follows that the desired vector $\boldsymbol{a}$ must correspond to the largest eigenvalue λ. This $\boldsymbol{a}$ is called the first principal component.

To obtain subsequent principal components it can be shown that these are also solutions to the characteristic polynomial [Krzanowski, 1990, Chapter 2] subject to the condition that all principal components are orthogonal to one another, i.e. $\boldsymbol{a}_i\boldsymbol{a}_j^T = 0 \ \forall i \neq j$. Thus to project the data from p dimensions to r dimensions, the first r principal component pairs $(\lambda_1, \boldsymbol{a}_1), \ldots, (\lambda_r, \boldsymbol{a}_r)$ are required where $\lambda_1 \geq \ldots \geq \lambda_r \geq 0$.

In the linguistic application we have five languages and so our data has a group structure that would be desirable to incorporate during dimension reduction. Thus, although PCA is the standard dimension reduction technique and can be used with discretised functional data, we instead turn to CVA as an approximation to CFA so as to include language information in our analysis and consider a different type of variability in the data set. The similarities between CFA and

CVA become clear in Section 3.5.2. Moreover, the justification for the use of CVA in relation to CFA is outlined in Kiiveri [1992] but in short the numerical solutions to the CFA optimality equation are none other than the solutions to a CVA.

3.5.2 Canonical variate analysis

Although less frequently implemented than PCA, the theory of CVA as a multivariate tool has been well developed in Krzanowski [1990, Chapter 11]. However, beyond being a purely multivariate technique, CVA can also be used with functional data as an approximation to CFA as is presented in Kiiveri [1992]. The technicalities of implementing CVA do not significantly differ whether in functional or multivariate settings. However, it is sometimes necessary to interpret their outputs differently, for example referring back to the smoothness of functional data. This is encountered in other instances and with other multivariate techniques that are used with functional data (e.g. Fervaha and Remington [2012]). As CVA will form the foundation of the eventual dimension reduction technique applied to the linguistic data, we shall go into a bit more detail than the previously mentioned techniques.

With regard to the linguistic application, in practice spectrograms are often discretised representations of underlying functions, and so each function $x^d_{l,m}$ is instead given by a matrix $\boldsymbol{X}^d_{l,m}$ with time-frequency dimensions $n_f \times n_t$ (i.e. the number of sample points of the frequency and time). These finite approximations tend to be high dimensional and so the question of dimension reduction is pertinent. The rows of these matrix representations of spectrograms can be concatenated to present the data as vectors and thus the data can be used with the standard vector-description of CVA. As CVA considers each covariance entry independently of its adjacent values, this does not affect the implementation of CVA. The only notable downside of concatenation is that it can obscure visual representation and description of the data.

In a similar fashion to PCA, the aim of CVA is to find successive uncorrelated vectors $\boldsymbol{a}$ that form linear combinations $y = \boldsymbol{a}\boldsymbol{x}^T$ (where $\boldsymbol{x}$ is p-dimensional data) that maximise the ratio of the between-groups covariance $\boldsymbol{B}$ to the within-groups covariance $\boldsymbol{W}$. In the context of the linguistic application, $\boldsymbol{B}$ describes the variation between the per-language mean spectrograms and the grand mean spectrogram, whereas the $\boldsymbol{W}$ describes the variation between individual observations and the associated per-language mean spectrograms.

3.5.2.1 Stating and solving the optimality problem

This section presents the background and workings of CVA as a technique to reduce data dimensionality whilst highlighting accurately true differences between g groups.

The aim of CVA is to find $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)$ which forms the linear combination $y_{ij} = \boldsymbol{a}\boldsymbol{x}_{ij}^T$ that (by some definition) optimises the separation between groups and within groups. More precisely, it maximises the between-groups sum of squares $SSB(\boldsymbol{a})$ and minimises the within-group sum of squares $SSB(\boldsymbol{a})$ through maximisation of:

$$F = \frac{\frac{1}{g-1} SSB(\boldsymbol{a})}{\frac{1}{n-g} SSW(\boldsymbol{a})} \tag{3.5.1}$$

where $SSB(\boldsymbol{a}) = \boldsymbol{a}\boldsymbol{B}_0\boldsymbol{a}^T$ and $SSW(\boldsymbol{a}) = \boldsymbol{a}\boldsymbol{W}_0\boldsymbol{a}^T$ and where $\boldsymbol{B}_0$ and $\boldsymbol{W}_0$ are respectively the between- and within-groups sum of squares and products matrices of dimensions $p \times p$:

$$\boldsymbol{B}_0 = \sum_{i=1}^{g} n_i(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}})^T(\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}) \tag{3.5.2}$$

$$\boldsymbol{W}_0 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i)^T(\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i) \tag{3.5.3}$$

A hypothesis test can be performed on the ratio F in (3.5.1), to investigate whether there are differences in the g group means. The ratio follows an F distribution on $g-1$ and $n-g$ degrees of freedom. See Section 3.5.2.4 for the assumptions required for CVA.

The between-groups variance-covariance matrix $\boldsymbol{B}$ and the within-groups variance-covariance matrix $\boldsymbol{W}$ are related to $\boldsymbol{B}_0$ and $\boldsymbol{W}_0$ as follows:

$$\boldsymbol{B} = \frac{1}{g-1}\boldsymbol{B}_0 \tag{3.5.4}$$

$$\boldsymbol{W} = \frac{1}{n-g}\boldsymbol{W}_0 \tag{3.5.5}$$

This allows us to rewrite (3.5.1):

$$F = \frac{\boldsymbol{a}\boldsymbol{B}\boldsymbol{a}^T}{\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T} \tag{3.5.6}$$

so the optimal $\boldsymbol{a}$ for the specified aim is found by maximising F with respect to $\boldsymbol{a}$, which can be determined through differentiation with respect to $\boldsymbol{a}$ and setting F' to $\mathbf{0}$.

$$F' = \frac{2\boldsymbol{B}\boldsymbol{a}^T(\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T) - 2\boldsymbol{W}\boldsymbol{a}^T(\boldsymbol{a}\boldsymbol{B}\boldsymbol{a}^T)}{(\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T)(\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T)} = \mathbf{0}$$

$$\Rightarrow \boldsymbol{B}\boldsymbol{a}^T = \frac{\boldsymbol{W}\boldsymbol{a}^T(\boldsymbol{a}\boldsymbol{B}\boldsymbol{a}^T)}{(\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T)}$$

Let the $\boldsymbol{a}$ that maximises F be denoted $\boldsymbol{a}^*$, and observe that for a constant $\lambda \in \mathbb{R}$:

$$\left.\frac{\boldsymbol{a}\boldsymbol{B}\boldsymbol{a}^T}{\boldsymbol{a}\boldsymbol{W}\boldsymbol{a}^T}\right|_{\boldsymbol{a}=\boldsymbol{a}^*} = \lambda$$

then

$$(\boldsymbol{B} - \lambda\boldsymbol{W})\boldsymbol{a}^T = \mathbf{0} \Rightarrow (\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda\boldsymbol{I})\boldsymbol{a}^T = \mathbf{0} \tag{3.5.7}$$

which is of a familiar form; λ is an eigenvalue and $\boldsymbol{a}^T$ an eigenvector of $\boldsymbol{W}^{-1}\boldsymbol{B}$. Furthermore, since λ is the supremum of F, $\boldsymbol{a}^T$ is the largest eigenvector of $\boldsymbol{W}^{-1}\boldsymbol{B}$. The larger the value of λ, the clearer the separation of the groups under the linear transformation involving $\boldsymbol{a}^T$.

Sometimes $\hat{\boldsymbol{W}}^{-1}$ does not exist — for example, if $p > n$ then $\hat{\boldsymbol{W}}$ is singular. This is a common occurrence for data from underlying functions with suspected high curvature, as to produce a suitably accurate approximation to the functional data, observations may be recorded with many data points to the extent that $p \gg n$. On the other hand if the data is assumed to have some underlying level of smoothness, then in effect the dimension could be thought of as much smaller. In the language data $\boldsymbol{X}$, $n = 23$ and $p = 8100$. Therefore obtaining a solution to non-invertible $\hat{\boldsymbol{W}}$ is not just of interest, but necessary for CVA due to data structure dimensions. A method of constructing an invertible $\hat{\boldsymbol{W}}$ is addressed in Section 3.6.1.

In summary, assuming that $\boldsymbol{W}$ is non-singular, finding the optimal $\boldsymbol{a}$ is equivalent to solving $(\boldsymbol{W}^{-1}\boldsymbol{B} - \lambda\boldsymbol{I})\boldsymbol{a}^T = \mathbf{0}$ where $\lambda \in \mathbb{R}$. This reduces to performing an eigenanalysis on $\boldsymbol{W}^{-1}\boldsymbol{B}$, whereby the eigenvector corresponding to the largest eigenvalue is the optimal $\boldsymbol{a}$.

3.5.2.2 Determining and selecting multiple canonical variates

In most practical situations more than one pair of eigenvalues and eigenvectors is required to study the differences between groups. The first pair can be renamed as $(\boldsymbol{a}_1^T, \lambda_1)$ and $\boldsymbol{y}_q = \boldsymbol{a}_q \boldsymbol{X}^T$ can be designated as the length n row vector relating to q^{th} canonical variate.

To obtain subsequent pairs $(\boldsymbol{a}_2^T, \lambda_2), \ldots, (\boldsymbol{a}_s^T, \lambda_s)$, consider eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}$ which are ordered $\lambda_1 > \lambda_2 > \ldots > \lambda_s > 0$ and corresponding eigenvectors $\boldsymbol{a}_1^T, \boldsymbol{a}_2^T, \ldots, \boldsymbol{a}_s^T$. Each subsequent $\boldsymbol{a}_j$ is uncorrelated with respect to the within-groups covariance matrix and so satisfies the property that $\boldsymbol{a}_i \boldsymbol{W} \boldsymbol{a}_j^T = 0 \ \forall i \neq j$. This means that $\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T$ is a diagonal matrix. The diagonal entries can be scaled in an arbitrary manner though the most common normalisation is that $\boldsymbol{a}_i \boldsymbol{W} \boldsymbol{a}_i^T = 1$ and so $\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T = \boldsymbol{I}$ where $\boldsymbol{I}$ is the identity matrix. Note that each decreasing eigenvalue accounts for less relative variability than the previous. It follows that $(\boldsymbol{a}_r^T, \lambda_r)$ has r^{th} greatest ratio of between-group to within-group variability and thus the optimal result in r directions uses $(\boldsymbol{a}_1^T, \lambda_1), \ldots, (\boldsymbol{a}_r^T, \lambda_r)$ (i.e. the sample individuals are plotted using the first r canonical variates as the axes).

Canonical variates can be determined for up to $s = \min(p, g-1)$ as (in general) this is the number of non-zero eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}$. When utilising multiple canonical variates, matrix notation is cleaner. Thus $\boldsymbol{\Lambda}$ is the $s \times s$ diagonal matrix with i^{th} diagonal entry λ_i, and $\boldsymbol{A}$ is the $s \times p$ matrix with i^{th} row $\boldsymbol{a}_i$. The canonical variate space is fully described by the collection of canonical variates $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}^T$. The canonical variates are uncorrelated (not orthogonal as in PCA) - this is a consequence of the optimality equations rather than an additional constraint. The group means in this new space are calculated as $\bar{\boldsymbol{y}}_i = \boldsymbol{A}\bar{\boldsymbol{x}}_i^T$ for group i.

It is convenient if $r \leq 3$ (where r is the number of canonical variates required for adequate representation), as this allows full representation using traditional plots. This is the case if $g \leq 4$ or $p \leq 3$ as then $s \leq 3$. Often $s > 3$ and so being able to select an appropriate r is important. However, as in PCA, it is rare to have anything but an arbitrary (albeit sensible) method for choosing r. Common methods are analogous to those used in PCA. For example, by considering the proportion of variability taken up by the first r canonical variates, $\frac{\lambda_1+\lambda_2+\ldots+\lambda_r}{\lambda_1+\lambda_2+\ldots+\lambda_s}$, a threshold can be specified to determine a value for r. For a detailed account of selection methods for r refer to Jolliffe [2002, Chapter 6].

3.5.2.3 Interpretation

Once canonical variates have been determined, interpretation of a canonical variate y_i can be performed. Although a similar situation is encountered in PCA, it is incorrect to rank canonical variates by the magnitude of their coefficients $\boldsymbol{a}_i$. In CVA a large coefficient indicates large between-group variability or a small within-group variability - the latter is of less interest in most analyses. Thus, the within-group variability component can be removed through standardisation to unit within-group variance:

$$\boldsymbol{a}_{ij}^{*} = \boldsymbol{a}_{ij}\sqrt{w_{jj}} \tag{3.5.8}$$

where w_{jj} is the i^{th} diagonal element of $\boldsymbol{W}$ the within-groups covariance matrix as defined through (3.5.5) and (3.5.3). Only after this transformation may magnitude be used for identifying variates which explain significant differences between groups.

3.5.2.4 Assumptions required for CVA

The first assumption is that the data set is jointly multivariate Gaussian, or in reality, sufficiently close to being so. Thus the first two moments are sufficient for characterising the data, and allows us to focus on the covariance of the data. The use of the F distribution introduces a second assumption; the variance ratio used in CVA follows an F distribution only if the pooled variance-covariance matrix $\boldsymbol{W}$ is representative of each of the groups, i.e. the true variance-covariance matrix for each group is the same (or sufficiently similar). If this is not the case, then CVA may be unsuitable and could give misleading results and explanations. These are discussed in more detail in Chapter 7.

3.5.2.5 Link between PCA and CVA

Campbell and Atchley [1981] give an elegant description of CVA as a twice applied PCA. The first PCA is performed on the original data whereby the full projection into principal component space means that the new axes are in the direction of the principal components. In terms of CVA, this relates to within-group variance and emphasises the reason for homogeneity of within-group variance. By scaling the data by one over the square root of the relevant eigenvalues, the

scaled projected data now has unit variances in all principal component directions i.e. this has performed a transformation from orthogonality to orthonormality. The second PCA is performed on the group means as projected into orthonormalised principal component space — the group means are typically weighted by the number of observations associated with each group. This relates to the between-group variance. This produces principal components but to obtain the desired canonical variates the scaling and rotation from the first PCA must be reversed. This returns the data to the original space as expected but now the principal components identified by the second PCA have now become canonical variates as desired.

When deciding whether to use CVA or PCA (if either), the first aspect to consider is whether the observations you are studying are grouped. If they are not grouped then CVA is not an option, whereas PCA does not demand any group structure. If the data set contains groups, then the second consideration is whether the differences and similarities in the variability of groups is considered of interest, in which case CVA as a data reduction tool or a tool for investigation modes of between to within covariance would be well suited. Conversely if the covariance structure is of more interest irrelevant of the grouping, then PCA will act across groups to find dominant directions of variability and would be more suited. In Section 3.5.3 this key difference between CVA and PCA is illustrated using simulated data. Of course, before using CVA, the required assumptions such as joint multivariate Gaussianity and common within-group covariances, both of which are cautioned in Section 3.5.2.4. Thus, validity of assumptions may end up being a key factor in whether to use CVA, PCA or see out an alternative tool.

3.5.3 Illustration of PCA and CVA

To illustrate the differences between PCA and CVA the two tools are applied to simulated data. We generate 400 samples split into two groups of 200 a follows:

$$X \sim Uni(-10, 10) \quad Y \sim Uni(0, 8)$$

Group 1 contains 200 samples from (X, Y) and Group 2 another 200 samples from $(X, -Y)$ where X and Y are sampled independently. Alternatively a Gaussian simulation could have been performed in a similar fashion, but for such a contrived (though informative) example this

is not necessary. A plot of this two dimensional data (after centring) can be seen in Figure 3.6 where Group 1 is coded as blues squares and Group 2 as red circles. We can then use PCA and CVA in turn to illustrate the difference in projection from two dimensions to one dimension. The first principal component (PC1) and canonical variate (CV1) are plotted on this same figure (though the direction of the arrows could be reversed as the eigenvectors are arbitrary by a factor -1). PCA ignores the group nature of the data so in this simulation as the number of samples

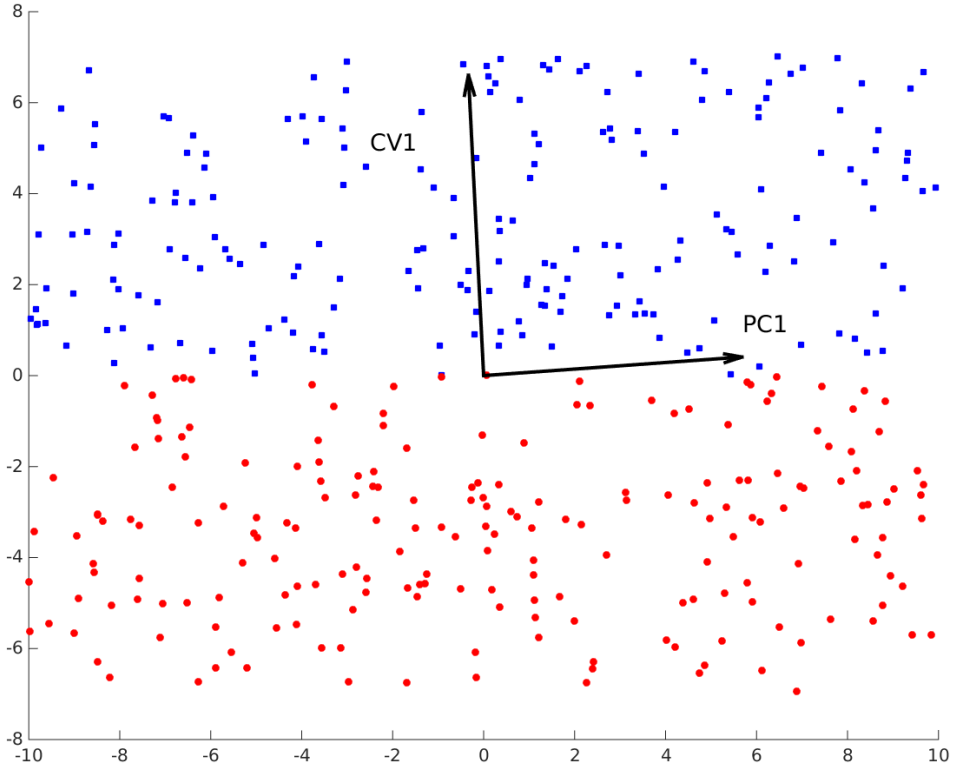


FIGURE 3.6: Centred simulated data where colour and shape indicate one of two nominal groups. The arrows indicate directions of the first principal component (PC1) and canonical variate (CV1).

tends to infinity the dominant direction of variability is parallel to horizontal axis. CVA on the other hand considers the group information aiming to find the dominant direction of between group variation to within group variation; CV1 turns out to be almost parallel to the vertical axis since the contrived simulations separate the samples efficiently in this manner. Figure 3.7 and Figure 3.8 show the one dimensional projections using dominant directions of variation specific to PCA and CVA respectively where the fainter points are the pre-projection data points. A slight offset is added to the projected data so that the groupings do not overlap visually — in practice all points lie on a single line.

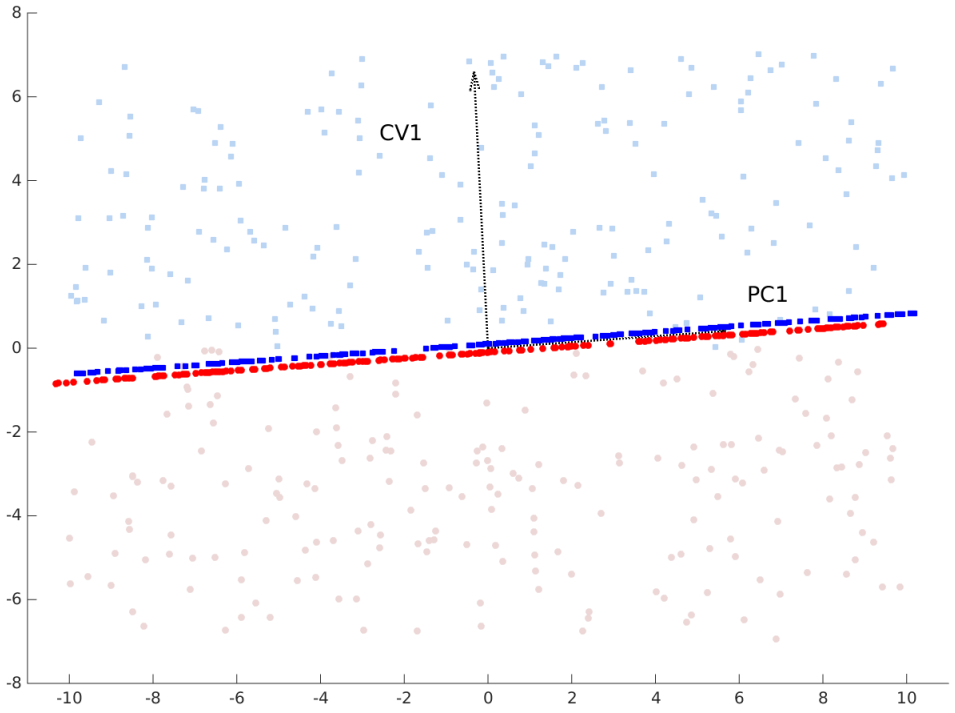


FIGURE 3.7: Centred simulated data projected using the first canonical variate and subsequently, for clarity, blue group points and red group points have had 0.1 added and subtracted respectively in the first co-ordinate.

Recall that principal components are orthogonal and hence perpendicular to one another and thus in this example the second principal component would be very similar to the first canonical variate. This is due to the form of the simulated data but in general this is not the case.

3.6 Separable covariance structure

As mentioned previously, if the number of variables is higher than the number of observations (i.e. $p > n$) then $\hat{\boldsymbol{W}}^{-1}$ does not exist because $\hat{\boldsymbol{W}}$ is singular. This is the case with the acoustic data set where $p = 8100$ and $n = 219$ but is a common occurrence in data analyses. In this section we adopt a novel approach which we refer to as separable-CFA and its discretised analogue separable-CVA which acts to approximate separable-CFA. This approach is very powerful because it overcomes the problem $p > n$ for all but the most extreme cases.

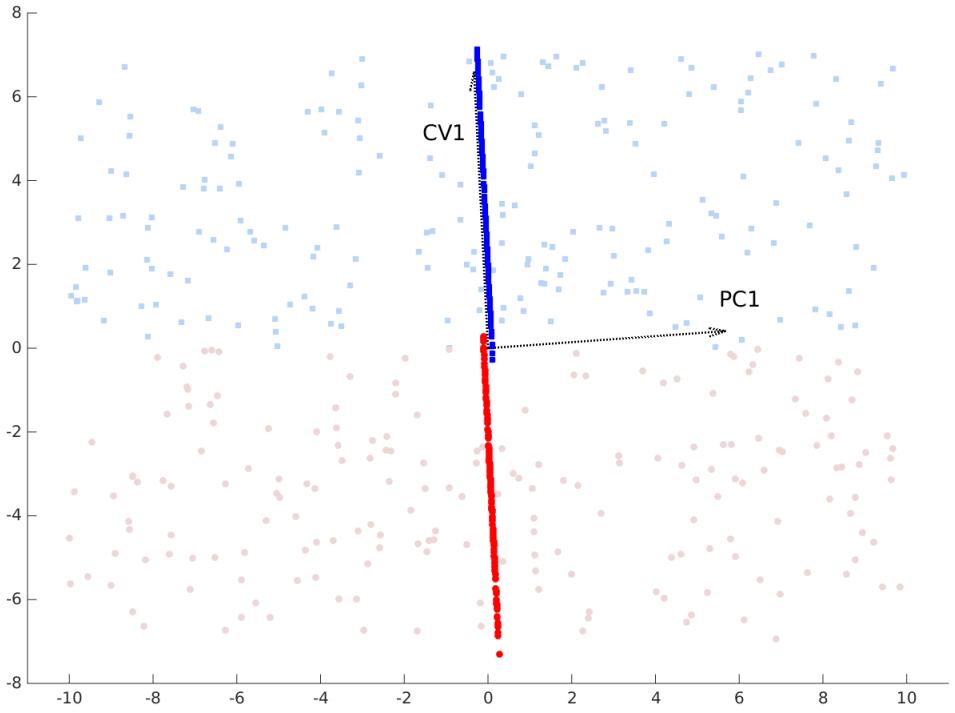


FIGURE 3.8: Centred simulated data projected using the first canonical variate and subsequently, for clarity, blue group points and red group points have had 0.1 added and subtracted respectively in the second co-ordinate.

3.6.1 Separable-CFA

Recall that the Romance data set comprises spectrograms which have time and frequency directions. A straightforward model for describing how these directions interact is that of separable covariance. The assumption underpinning this model can be encapsulated as there being no dependency between the (standardised) time and frequency of the data. This is, of course, a significant simplification of the likely underlying model. However, there are circumstances where the assumption can be made with little penalty and can prove to be particularly useful computationally. Recall that a covariance function C is said to be separable if:

$$C((f_1, t_1), (f_2, t_2)) = C_f(f_1, f_2)C_t(t_1, t_2) \tag{3.6.1}$$

where C_f and C_t are functions only of their arguments. The factored covariances provide an understanding of how frequency or time dimensions of the spectrograms vary when the other has been averaged out. Perhaps even clearer, under the separable covariance

$$corr((f_1, t_1), (f_1, t_2)) = corr_t(t_1, t_2).$$

It is useful to consider the notion of covariance separability in terms of statistical independence and uncorrelation. Rougier provides the necessary results. If two random processes G_f and G_t are probabilistically independent then $C = C_f C_t$ has a separable covariance function. However, the converse is not true, i.e. if $C = C_f C_t$ then G_f and G_t are not necessarily probabilistically independent. Moreover, if G_f and G_t are uncorrelated then this an insufficient condition for $G = G_f G_t$ to have a separable covariance function. The sufficient condition requires that G_f and G_t are second-order uncorrelated, that is

$$\mathbb{E}(G(f_1, t_1)G(f_2, t_2) = \mathbb{E}(G_f(f_1)G_f(f_2))\mathbb{E}(G_t(t_1)G_t(t_2)).$$

Thus, in this application the separable-covariance assumption implies that the time and frequency dimensions can be described by second-order uncorrelated processes, but are not necessarily probabilistically independent of one another. This separability assumption is invariant of Gaussianity assumptions or otherwise. Interestingly, Gaussianity of G with separable covariance does not imply that G_f and G_t are necessarily Gaussian [Rougier].

However, the main purpose of the assumption becomes apparent for the Romance data set subsequently (as described in Section 3.6.3) when use of the separable model of covariance overcomes the challenge of covariance rank deficiency. The implications of making this separability assumption are discussed in Section 3.6.3 where it is shown that for the purposes of forming a basis, the validity of the assumption is not of great concern.

Under the separable covariance assumption for two-dimensional data the CFA optimality equation equivalent to (3.4.4) is:

$$\int_{\mathcal{F}} \int_{\mathcal{T}} (B_f(f_1, f_2)B_t(t_1, t_2) - \lambda_q W_f(f_1, f_2)W_t(t_1, t_2))h_q(f_2, t_2)\, dt_2\, df_2 = 0. \tag{3.6.2}$$

It can be shown that the solutions to this equation can be obtained as the product of the solutions to two CFAs performed on the frequency and time covariances separately. Thus given any canonical function pairs $(h_{q_f}(f_2), \lambda_{q_f})$ and $(h_{q_t}(t_2), \lambda_{q_t})$ from a frequency and time CFA respectively, the products provide a solution to (3.6.2): $h_q(f_2, t_2) = h_{q_f}(f_2)h_{q_t}(t_2)$ and $\lambda_q = \lambda_{q_f}\lambda_{q_t}$. Moreover, any solution to (3.6.2) can be obtained from such products. A proof of this result

is given in Section 3.6.2 and is very useful when proceeding to obtain numerical solutions to (3.6.2).

3.6.2 Product solutions for separable-CFA

For CFA the separability assumption is:

$$B((f_1,t_1),(f_2,t_2)) = B_f(f_1,f_2)B_t(t_1,t_2)$$

$$W((f_1,t_1),(f_2,t_2)) = W_f(f_1,f_2)W_t(t_1,t_2)$$

We want to find solutions to:

$$\int_{\mathcal{F}}\int_{\mathcal{T}} (B((f_1,t_1),(f_2,t_2)) - \lambda_q W((f_1,t_1),(f_2,t_2)))h_q(f_2,t_2)\,dt_2\,df_2 = 0$$

which under the separability assumptions is:

$$\int_{\mathcal{F}}\int_{\mathcal{T}} (B_f(f_1,f_2)B_t(t_1,t_2) - \lambda_q W_f(f_1,f_2)W_t(t_1,t_2))h_q(f_2,t_2)\,dt_2\,df_2 = 0 \tag{3.6.3}$$

Considering frequency and time in turn, for each we wish to find $\lambda_1 > \cdots > \lambda_q > \cdots$ corresponding to $h_1(\cdot),\cdots,h_q(\cdot),\cdots$ which solve:

$$\int_{\mathcal{F}} (B_f(f_1,f_2) - \lambda_{q_f} W_f(f_1,f_2))h_{q_f}(f_2)\,df_2 = 0 \tag{3.6.4}$$

$$\int_{\mathcal{T}} (B_t(t_1,t_2) - \lambda_{q_t} W_t(t_1,t_2))h_{q_t}(t_2)\,dt_2 = 0 \tag{3.6.5}$$

Lemma 3.6.1. *The solutions to* (3.6.3) λ_q *and* $h_q(f_2,t_2)$ *can be constructed as follows* $\lambda_q = \lambda_{q_f}\lambda_{q_t}$ *and* $h_q(f_2,t_2) = h_{q_f}(f_2)h_{q_t}(t_2)$ *if and only if* λ_q *and* $h_q(f_2,t_2)$ *are solutions to* (3.6.4) *and* (3.6.5).

Proof. Under reasonable conditions we can rewrite (3.6.3)

$$\begin{aligned}&\int_{\mathcal{F}}\int_{\mathcal{T}}(B_f(f_1,f_2)B_t(t_1,t_2))h_q(f_2,t_2)\,dt_2\,df_2\\=&\int_{\mathcal{F}}\int_{\mathcal{T}}\lambda_q(W_f(f_1,f_2)W_t(t_1,t_2))h_q(f_2,t_2)\,dt_2\,df_2\end{aligned}$$

Now without loss of generality select any pair of solutions to (3.6.4) and (3.6.5) rewrite $\lambda_q = \lambda_{q_f}\lambda_{q_t}$ and $h_q(f_2,t_2) = h_{q_f}(f_2)h_{q_t}(t_2)$

$$\begin{aligned}\Rightarrow&\int_{\mathcal{F}}B_f(f_1,f_2)h_{q_f}(f_2)\,df_2\int_{\mathcal{T}}B_t(t_1,t_2)h_{q_t}(t_2)\,dt_2\\=&\int_{\mathcal{F}}\lambda_{q_f}W_f(f_1,f_2)h_{q_f}(f_2)\,df_2\int_{\mathcal{T}}\lambda_{q_t}W_t(t_1,t_2)h_{q_t}(t_2)\,dt_2\end{aligned}$$

Through rearrangement of (3.6.4) and (3.6.5) it is clear that $\lambda_q = \lambda_{q_f}\lambda_{q_t}$ and $h_q(f_2,t_2) = h_{q_f}(f_2)h_{q_t}(t_2)$ are solutions of (3.6.3). This proves sufficiency.

Now necessity. Suppose there is $h_q(f_2,t_2) \neq h_{q_f}(f_2)h_{q_t}(t_2)$ and corresponding λ solving (3.6.3), and we now proceed so as to reach a contradiction. Assuming that the absolute value of the statement is finite when integrated, the integrals can be exchanged to give:

$$\begin{aligned}&\int_{\mathcal{F}}\int_{\mathcal{T}}(B_f(f_1,f_2)B_t(t_1,t_2))h_q(f_2,t_2)\,dt_2\,df_2\\=&\int_{\mathcal{F}}\int_{\mathcal{T}}\lambda(W_f(f_1,f_2)W_t(t_1,t_2))h_q(f_2,t_2)\,dt_2\,df_2\end{aligned}$$

and

$$\begin{aligned}&\int_{\mathcal{T}}\int_{\mathcal{F}}(B_f(f_1,f_2)B_t(t_1,t_2))h_q(f_2,t_2)\,df_2\,dt_2\\=&\int_{\mathcal{T}}\int_{\mathcal{F}}\lambda(W_f(f_1,f_2)W_t(t_1,t_2))h_q(f_2,t_2)\,df_2\,dt_2\end{aligned}$$

Evaluating the inner integrals gives:

$$\int_{\mathcal{F}} k_1 B_f(f_1, f_2) h_{q_f}(f_2)\, df_2 = \int_{\mathcal{F}} \lambda k_2 W_f(f_1, f_2) h_{q_f}(f_2)\, df_2$$

$$\int_{\mathcal{T}} k_3 B_t(t_1, t_2) h_{q_t}(t_2)\, dt_2 = \int_{\mathcal{T}} \lambda k_4 W_t(t_1, t_2) h_{q_t}(t_2)\, dt_2$$

where $k_i \in \mathcal{R}$. However, by collecting the constants it is clear that $h_{q_f}(f_2)$ and $\frac{\lambda k_2}{k_1}$ solve (3.6.4), and that $h_{q_t}(t_2)$ and $\frac{\lambda k_4}{k_3}$ solve (3.6.5).

Thus we have $h_q(f_2, t_2) = h_{q_f}(f_2) h_{q_t}(t_2)$ and $\lambda = \frac{\lambda k_2}{k_1} \frac{\lambda k_4}{k_3}$ but this contradicts our supposition.

□

3.6.3 Separable-CVA

As demonstrated by Lemma 3.6.1, the overall solutions to a CFA optimality problem with a separable covariance structure are found as the product of solutions to CFAs of the decomposed covariance functions. We propose combining a tensor decomposable covariance structure with CVA in order to obtain numerical solutions to the decomposition of the separable-CFAs. This, when taking products, also gives solutions to the overall CFA. While separable covariance structures have been adopted elsewhere in the literature (e.g. Aston and Kirch [2012], Jones and Moriarty [2012]), this is a novel approach for both CVA and CFA. Even though the assumption behind separable covariance is strong, the accuracy of the assumption for CVA and CFA only impacts on basis efficiency not basis validity as if the complete basis is retained then it will still span the space. Thus, if the data is far from separable, then simply a higher number of dimensions will be needed to retain the same amount of information.

The main purpose of assuming a tensor-decomposable covariance structure is to overcome the obstacle of rank-deficient sample covariance matrices caused by the length of the observations exceeding the number of observations (i.e. $p > n$). This is not just a problem with the Romance speaker data set but is commonly encountered with functional data sets due to their often high-dimensionality (e.g. Long et al. [2005]). Rank deficiency obstructs using CVA to obtain numerical solutions to CFA. Theoretically in CFA an inverse function W^{-1} is neither required nor is usually bounded, whereas in CVA $\boldsymbol{W}^{-1}$ is needed for the eigenanalysis of $\boldsymbol{W}^{-1}\boldsymbol{B}$ but cannot be obtained because in this case $\boldsymbol{W}$ is singular.

In the observational matrix setting, $\boldsymbol{C}$ is separable if:

$$\boldsymbol{C}((f_1,t_1),(f_2,t_2)) = \boldsymbol{C}_f(f_1,f_2) \otimes \boldsymbol{C}_t(t_1,t_2)$$

where $\otimes$ is the standard Kronecker product, c.f. (3.6.1). Using known results of the Kronecker product (see Lancaster and Tismenetsky [1985] for example), the separability assumption in the multivariate setting implies:

$$\boldsymbol{W}^{-1}\boldsymbol{B} = (\boldsymbol{W}_t^{-1} \otimes \boldsymbol{W}_f^{-1})(\boldsymbol{B}_t \otimes \boldsymbol{B}_f) = \boldsymbol{W}_t^{-1}\boldsymbol{B}_t \otimes \boldsymbol{W}_f^{-1}\boldsymbol{B}_f \tag{3.6.6}$$

where the estimates of separate within- and between-language covariance matrices in the frequency direction are:

$$\hat{\boldsymbol{B}}_f[f_1,f_2] = \frac{1}{n_l-1}\sum_{l=1}^{n_l}\frac{m_{l\bullet}}{n_t}\sum_{t=1}^{n_t}\tilde{\boldsymbol{X}}_l[f_1,t]\tilde{\boldsymbol{X}}_l[f_2,t]$$
$$\hat{\boldsymbol{W}}_f[f_1,f_2] = \frac{1}{n-n_l}\sum_{l=1}^{n_l}\sum_{d=1}^{n_d}\sum_{m=1}^{m_{ld}}\frac{1}{n_t}\sum_{t=1}^{n_t}\tilde{\boldsymbol{X}}^d_{l,m}[f_1,t]\tilde{\boldsymbol{X}}^d_{l,m}[f_2,t]$$

where $\tilde{\boldsymbol{X}}_l[i,j] = \bar{\boldsymbol{X}}_l[i,j] - \bar{\boldsymbol{X}}[i,j]$ and $\tilde{\boldsymbol{X}}^d_{l,m}[i,j] = \boldsymbol{X}^d_{l,m}[i,j] - \bar{\boldsymbol{X}}_l[i,j]$ with equivalent estimates for the time direction. Treating each frequency and time sample as a separate observation leads to the product covariance matrices $\boldsymbol{W}$ and $\boldsymbol{B}$ having higher ranks than previously. Explicitly, for $\boldsymbol{W}^{-1} = (\boldsymbol{W}_f \otimes \boldsymbol{W}_t)^{-1}$ to be nonsingular, we need that $nn_f \geq n_t$ and $nn_t \geq n_f$. This is equivalent to requiring $n \geq \frac{\max(n_f,n_t)}{\min(n_f,n_t)}$. This contrasts to the previous condition $n \geq p = n_f n_t$. So the new requirement is usually significantly more relaxed, and CVA can often then be implemented.

An eigenanalysis of $\boldsymbol{W}_f^{-1}\boldsymbol{B}_f$ produces eigenvalues $(\lambda_{f1},\lambda_{f2},\ldots,\lambda_{fn_f})$ and corresponding eigenvectors $(\boldsymbol{c}_{f1},\ldots,\boldsymbol{c}_{fn_f})$ with equivalent output for the time covariances $\boldsymbol{W}_t^{-1}\boldsymbol{B}_t$. Sorting decreasingly, $(\lambda_{f1},\ldots,\lambda_{fn_f}) \otimes (\lambda_{t1},\ldots,\lambda_{tn_t})$ produces a vector denoted $(\lambda_1,\lambda_2,\ldots,\lambda_{n_f n_t})$ and the Kronecker product of corresponding eigenvectors results in $(\boldsymbol{c}_1,\boldsymbol{c}_2,\ldots,\boldsymbol{c}_{n_f n_t})$ of size $n_f \times n_t$, solving the overall CVA. It should be noted that while this basis defined is based on an assumption of separability, it nevertheless provides a complete basis of the space. So although when separability does not hold the basis is less efficient and is rather longer than it needs to be, the basis is still valid. For further details, see an analogous argument for separable PCA in

Aston and Kirch [2012].

If using CVA for dimension reduction, once a dimension $r < p$ has been selected, each observation from the language data can be projected into r-dimensional space: $\boldsymbol{Y} = \boldsymbol{A}\tilde{\boldsymbol{X}}^T$ where $\boldsymbol{A}$ is $r \times p$ with columns $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_r$ and where $\tilde{\boldsymbol{X}}$ is $1 \times p$ and formed by concatenating the n_f rows of length n_t of the observation $\boldsymbol{X}$.

3.7 Summary

In this chapter the main data set of the thesis has been introduced. This is linguistic data based on audio recordings of speakers from 5 languages. More specifically, the acoustic data takes the form of spectrograms. The ultimate aim of the analysis is to determine whether certain aspects of speech could have evolved in a tree-like manner, i.e. can the conditional independence relationships between particular phonetic features be adequately modelled as a GLTM. This will be achieved by testing particular sets of equations and inequalities called tree constraints. However, before this part of the analysis can be performed we need to get the data into a suitable form that can be used in conjunction with tree constraints. Given the data objects are spectrograms which are two-dimensional functional objects, our analyses of the data will be in a functional data framework. Thus, in this chapter we have introduced functional tools most notably the new construction separable-CFA and the multivariate approximation separable-CVA. The purpose of these separable techniques is two-fold: firstly, the separability overcomes the problem of high dimensional data with only a modest number of observations. Secondly, the CFA and CVA can be used to project the data to a canonical basis and greatly reduce the data dimension while maintaining important aspects of variability between the languages. CFA and CVA will therefore be crucial to the main analyses that are carried out in Chapter 7 prior to the application of tree constraints.

Chapter 4

Tree constraints for discrete distributions

The purpose of this chapter is to introduce the concept of tree constraints beginning with a review for BNs with binary random variables. This will cover some results from the papers Settimi and Smith [1999, 2000] and more recently Zwiernik and Smith [2011, 2012]. An illustration utilising some of the constraints is then given for linguistic and genetic data sets. We then present our contribution towards the development of graphical inequality diagnostics for binary random variables that were developed in Shiers and Smith [2012]. A key extension of the binary case to k-state variables will then be reported focusing on the work of Allman et al. [2014].

This chapter provides the third strand of background material required for implementing the desired analysis of the linguistic data set. Chapter 3 provided a background to the linguistic application and detailed the functional approach to be taken, and Chapter 2 provided the basics of the model types we are considering. In this chapter, we take a look at the types of concepts and tools used for assessing model-compatibility and demonstrate how they are of use in other contexts. This then sets up Chapter 5 and Chapter 6 to extend these tools and methodologies to the Gaussian setting required for the functional linguistic application in Chapter 7.

4.0.1 Introduction to tree constraints

Recall from Section 2.1 that BNs with both hidden and observed variables are known as latent variable graphical models. Moreover, if these graphical models are trees and the unobserved variables are the interior nodes then these graphical models are known formally as latent tree models. Such models are commonly referred to as phylogenetic trees with the leaf variables representing extant species and the interior nodes representing extinct ancestor species. Traditionally the species have been biological organisms (e.g. Felsenstein [1983]). However, the idea naturally extends to other fields such as linguistics (e.g. Dunn et al. [2005]). For example, the evolutionary histories of languages have often been presented in a tree form with contemporary languages being found at the leaves.

The versatility of phylogenetic trees makes the study of latent tree models of particular interest from a statistical standpoint (e.g. Dutkowski and Tiuryn [2007]). However, the estimation of BNs when interior vertices are hidden is challenging since the implicit geometry of the associated probability mass or density functions of the observed variables, and hence also the likelihood, is complicated. As a consequence there are fewer tools available to assess the fit of latent tree models to data. This thesis makes a contribution to better understanding the class of GLTMs theoretically and develops methodology for assessing the fit of such models to data sets. The approach taken focuses on a fundamental feature of phylogenetic trees called latent tree constraints (for which we shall refer to as tree constraints henceforth).

Definition 4.0.1 (Tree constraints)**.** *A tree constraint is an implicit theoretical restriction on the probability space of observed variables of a latent tree model.*

Remark 4.0.2. Tree constraints are usually expressed in terms of moments of the observed variables and take the form of algebraic or semi-algebraic statements (i.e. equations or inequalities).

The nature of the tree constraints generally depend on the assumed distributions of the random variables. However, if the given model context is clear, then the following terminology can be used freely.

Definition 4.0.3 (Universal and T-specific)**.** *If a tree constraint is applicable to any tree T then it is said to be universal. If a tree constraint only applies to a particular tree T then it is said to be T-specific.*

Definition 4.0.4 (Tree-compatibility)**.** *A data set representing observations of a latent tree model is said to be tree-compatible if it is deemed to adhere to the universal tree constraints.*

Definition 4.0.5 (T-compatibility)**.** *A data set representing observations of a latent tree model is said to be T-compatible if it is deemed to adhere to T-specific tree constraints for a given tree T.*

Our interest is in using Gaussian tree constraints to test data sets for tree-compatibility and T-compatibility. In Chapter 7, a linguistic and a biological data set are tested against the constraints, which is apt as in both settings evolutionary histories have previously been modelled as phylogenetic trees. But before we derive the complete set of Gaussian tree constraints required for these analyses, we begin by reviewing the existing tree constraints for some discrete graphical models.

4.1 Binary tree constraints

The simplest such model is the phylogenetic tree where all variables are binary. The geometry of this space of models was presented for the tripod tree in Settimi and Smith [1999] and expanded in Settimi and Smith [2000] to any star tree. It has recently attracted considerable interest (for example Allman et al. [2009] and Zwiernik and Smith [2012]). These advances have encouraged some authors to proceed to use the known geometry of these spaces to support inference and learning over the space of tree models (see Drton and Sullivant [2007] for example). These focus on the polynomial constraints that are implicit in these models. However, it is well known that not only these functional relationships but also additional inequality constraints are active. Recently Zwiernik and Smith [2011] fully characterised these inequality constraints for binary phylogenetic trees. So we are at last able, at least for this important class of graphical models, to explore the inferential use for learning of these derived inequality constraints.

We begin this section by reviewing the binary tree constraints derived in Settimi and Smith [1999] for the tripod tree — we broadly describe the derivation to give an idea of the approach used. We then present the results from Settimi and Smith [2000] for a star tree with four observed nodes and moreover a general framework for obtaining higher order moments for larger star trees.

4.1.1 Constraints on a tree with three observed nodes and one hidden node

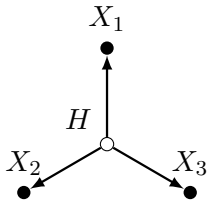


FIGURE 4.1: Tripod tree T_3.

Following Settimi and Smith [1999, Section 4] we motivate the derivations of tree constraints for the tripod tree in Figure 4.1 where the random variables are binary taking values $\{-1, +1\}$. The tripod tree has three manifest random variables X_1, X_2, X_3 and hidden variable H.

The probability space of the manifest variables can be efficiently represented in terms of moments. For any n variables denoted $X_1, \ldots, X_n$, we define n^{th} order non-central moments as

$$\mu_{1,2,\ldots,n} = \mu_{12\cdots n} = E(X_1 X_2 \ldots X_n).$$

and n^{th} order central moments as

$$\sigma_{1,2,\ldots,n} = \sigma_{12\cdots n} = E((X_1 - \mu_1)(X_2 - \mu_2)\ldots(X_n - \mu_n))$$

So, for example, the mean of variable X_i is denoted μ_i, and the covariance of X_i and X_j as σ_{ij}.

Considering the joint probability mass function for the tripod tree, for $\Pr(i, j, k) = \Pr(X_1 = i, X_2 = j, X_3 = k)$, Settimi and Smith [1999] write this in additive form:

$$\begin{aligned}\Pr(i,j,k) &= \frac{1}{8}((1 + i\mu_1)(1 + j\mu_2)(1 + k\mu_3) + ij\sigma_{12} + ik\sigma_{13} + jk\sigma_{23} + ijk\lambda_{123}) \\ &= \frac{1}{8}(1 + i\mu_1 + j\mu_2 + k\mu_3 + ij\mu_{12} + ik\mu_{13} + jk\mu_{23} + ijk\mu_{123}) \qquad (4.1.1)\end{aligned}$$

where $\lambda_{123} = \mu_{123} - \mu_1\mu_2\mu_3$. The tripod tree has the following conditional independence statements

$$X_1 \perp\!\!\!\perp X_2|H, \quad X_1 \perp\!\!\!\perp X_3|H, \quad X_2 \perp\!\!\!\perp X_3|H, \quad X_1 \perp\!\!\!\perp (X_2, X_3)|H.$$

The following constraints on the central moments hold given $X_i \perp\!\!\!\perp X_j | H$:

$$\operatorname{var}(H)\operatorname{cov}(X_i, X_j) = \operatorname{cov}(X_i, H)\operatorname{cov}(X_j, H) \tag{4.1.2}$$

for $1 \leq i < j \leq 3$. Additionally $X_1 \perp\!\!\!\perp (X_2, X_3) | H$ can be expressed by:

$$\operatorname{var}(H)\operatorname{cov}(X_2X_3, X_1) = \operatorname{cov}(X_2X_3, H)\operatorname{cov}(X_1, H) \tag{4.1.3}$$

Derivations of (4.1.2) and (4.1.3) are given in Settimi and Smith [2000]. Now let

$$\eta_i := \frac{\operatorname{cov}(X_i, H)}{\operatorname{var}(H)},\ i \in \{1, 2, 3\}$$

and

$$\mu_H := E(H).$$

Through algebraic manipulation (4.1.2) is rearranged to the form:

$$\sigma_{ij} = (1 - \mu_H^2)\eta_i\eta_j \tag{4.1.4}$$

where $1 \leq i < j \leq 3$. Rearranging (4.1.3) and substituting in equations from (4.1.2) and (4.1.4), Settimi and Smith [1999] give the final equality constraint as:

$$\sigma_{123} = -2\mu_H(1 - \mu_H^2)\eta_1\eta_2\eta_3. \tag{4.1.5}$$

Observe that $\operatorname{var}(H) = E(H^2) - E^2(H)$ and since $H \in \{-1, +1\} \Rightarrow H^2 = 1 \Rightarrow E(H^2) = 1$ and so we get $\operatorname{var}(H) = 1 - \mu_H^2$ which is seen in the above constraints. The four constraints in (4.1.4) and (4.1.3) indicate that the joint probability distribution across (X_1, X_2, X_3, H) is provided by moments of the observed variables (X_1, X_2, X_3) up to aliasing or sign changes on H. This is a well-known fact from latent variable modelling (for example, see Goodman and Mirande [1974]). By squaring (4.1.5) and substituting in (4.1.4) the following are obtained:

$$(1 - \mu_H^2)\sigma_{123}^2 = 4\mu_H^2\sigma_{12}\sigma_{13}\sigma_{23} \tag{4.1.6}$$

The first term is non-negative which implies that solutions must lie in areas where:

$$\sigma_{12}\sigma_{13}\sigma_{23} \geq 0 \tag{4.1.7}$$

Settimi and Smith [1999] go on to present inequality constraints on the probability space in terms of moments of observed random variables.

For $|\sigma_{ij}| \leq |\sigma_{ik}|, |\sigma_{jk}|$:

$$2|\sigma_{ij}| \geq |\sigma_{ik}||\sigma_{jk}| + \sqrt{\sigma_{ik}^2\sigma_{jk}^2 + \sigma_{ijk}^2} \tag{4.1.8}$$

The constraints (4.1.8) specify symmetric non-intersecting regions that the central moments can lie in. These are found in all four of the multiplicative positive octants of the second-order moment space, as specified by (4.1.7).

These inequalities in (4.1.7) and (4.1.8) are T_3-compatibility constraints. We shall refer to them as the positivity constraints and the (binary) tripod constraints. In fact, with reference to Theorem 4.3.1, from the implicit conditional independence relationship implied by a phylogenetic tree, we will see that constraints on the tripod tree appear in any phylogenetic tree and so the positivity and tripod constraints are universal tree constraints. This makes the tripod constraints very important and fundamental to any tree. When these are used with the true values of the moments, the inequalities must be satisfied for the conditional independence statements to hold and thus for T_3 to be a suitable model. Of course in practice we rarely know the true moments, and thus we can construct a three-way contingency table over (X_1, X_2, X_3), and can estimate the moments $\mu_1, \mu_2, \mu_3, \sigma_{12}, \sigma_{13}, \sigma_{23}, \sigma_{123}$. If the estimates lie outside the feasible region determined by (4.1.7) and (4.1.8) then this provides evidence that the conditional independence assumption of the tripod tree is incorrect.

Let us consider interesting points in the space of moments. The singular points are geometrically represented as the boundaries of the four regions and can be found by changing Equation (4.1.8) from an inequality to an equality. These boundaries represent zero counts in one or more cells of the contingency tables. These singular points are discussed in more detail in Section 4.1.2.

Section 4.1.4 will explore the admissible regions of the moment space graphically, both for the whole space and for an isolated positive octant.

4.1.2 Constraints on a tree with four observed nodes and one hidden node

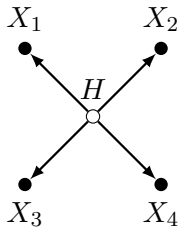


FIGURE 4.2: A directed tree with four observed nodes and one hidden node.

In Settimi and Smith [2000], the tree constraints are derived for the star tree with 4 leaves.

The tree in Figure 4.2 contains the conditional independence statements:

$$X_i \perp\!\!\!\perp X_j | H \; \forall \, i, j \in \{1, 2, 3, 4\}, i \neq j$$

As in Section 4.1.1, we will consider the additive model of the probability distribution of the observed variables:

$$\Pr(X_1 = i, X_2 = j, X_3 = k, X_4 = l) = \frac{1}{16}\big((1 + i\mu_1)(1 + j\mu_2)(1 + k\mu_3)(1 + l\mu_4)$$

$$+ij\sigma_{12}+ik\sigma_{13}+il\sigma_{14}+jk\sigma_{23}+jl\sigma_{24}+kl\sigma_{34}+ij\mu_{123}+ij\mu_{124}+ik\mu_{134}+jk\mu_{234}+ijkl\mu_{1234}\big)$$

$$i, j, k, l \in \{-1, +1\}$$

Settimi and Smith [2000] make use of Bahadur expansion for high order probability distributions in order to rewrite the non-central moments in terms of central moments (see Lancaster [1969] for generalised Bahadur expansion work and Streitberg [1990] for corrected versions for degrees of $n > 3$).

$$\mu_{ijk} = \sigma_{ijk} + \mu_i\mu_j\mu_k + \mu_i\sigma_{jk} + \mu_j\sigma_{ik} + \mu_k\sigma_{ij}$$

and

$$\mu_{1234} = \sigma_{1234} + \sum_{i \neq j < k < l} \mu_i \sigma_{jkl} + \sum_{i < j \neq k < l} \sigma_{ij}\sigma_{kl} + \sum_{i \neq j \neq k < l} \mu_i \mu_j \sigma_{kl} + \mu_i \mu_j \mu_k \mu_l$$

As before, the following equality constraints for the conditional independence model can be expressed as:

$$\sigma_{ij} = (1 - \mu_H^2)\eta_i \eta_j \; 1 \leq i < j \leq 4 \tag{4.1.9}$$

$$\sigma_{ijk} = -2\mu_H(1 - \mu_H^2)\eta_i \eta_j \eta_k \; 1 \leq i < j < k \leq 4 \tag{4.1.10}$$

Additionally for the fourth-order central moment

$$\sigma_{1234} = 2(1 - \mu_H^2)(3\mu_H^2 - 1)\eta_1 \eta_2 \eta_3 \eta_4 \tag{4.1.11}$$

Through advanced algebraic manipulation and substitution Settimi and Smith [2000] obtain the following inequality constraints:

$$0 \leq \tau_{ij} = \frac{\sqrt{4\sigma_{ij}\sigma_{ik}\sigma_{jk} + \sigma_{ijk}^2}}{2|\sigma_{ij}|} \leq 1 \tag{4.1.12}$$

$$0 \leq \tau_{ik} = \frac{\sqrt{4\sigma_{ij}\sigma_{ik}\sigma_{jk} + \sigma_{ijk}^2}}{2|\sigma_{ik}|} \leq 1 \tag{4.1.13}$$

$$0 \leq \tau_{jk} = \frac{\sqrt{4\sigma_{ij}\sigma_{ik}\sigma_{jk} + \sigma_{ijk}^2}}{2|\sigma_{jk}|} \leq 1 \tag{4.1.14}$$

for $1 \leq i < j < k \leq 4$.

These inequalities allow us to test whether the conditional independence assumptions of the tree are violated (i.e. if $\tau_{ij} > 1$ for at least one pair $i, j \; s.t. \; 1 \leq i < j \leq 4$ then this is evidence that the model may not be a good fit).

Additionally, Settimi and Smith [2000] report that the following algebraic (i.e. equality) constraints are implicit in the model:

$$\sigma_{12}\sigma_{34} = \sigma_{14}\sigma_{23} \tag{4.1.15}$$

$$\sigma_{13}\sigma_{24} = \sigma_{14}\sigma_{23} \tag{4.1.16}$$

$$\sigma_{123}\sigma_{14} = \sigma_{124}\sigma_{13} \tag{4.1.17}$$

$$\sigma_{123}\sigma_{14} = \sigma_{134}\sigma_{12} \tag{4.1.18}$$

$$\sigma_{124}\sigma_{23} = \sigma_{234}\sigma_{12} \tag{4.1.19}$$

$$\sigma_{124}^2\sigma_{134}^2 = (\sigma_{1234}\sigma_{124}\sigma_{134} + 2\sigma_{123}\sigma_{234}\sigma_{14}^2)\sigma_{14} \tag{4.1.20}$$

These equations suggest a sufficient reparametrisation of the sample space is the sample means and a subset of central moments. For instance, μ_1, μ_2, μ_3, μ_4, and σ_{ij}, σ_{kl} for i, j, k, l distinct and σ_{ijk} for $2 \leq j < k \leq 4$.

The tripod tree is identified in the star tree by selecting any three leaf nodes and so the T_3-compatibility constraints are also implicit in the 4-leaf star tree:

$$\sigma_{ij}\sigma_{ik}\sigma_{jk} > 0 \; \forall \; i, j, k \in \{1, ..., 4\}, i \neq j \neq k \neq i \tag{4.1.21}$$

$$2|\sigma_{ij}| \geq |\sigma_{ik}||\sigma_{jk}| + \sqrt{\sigma_{ik}^2\sigma_{jk}^2 + \sigma_{ijk}^2} \tag{4.1.22}$$

for $|\sigma_{ij}| \leq |\sigma_{ik}|, |\sigma_{jk}|$ where $1 \leq i < j \neq k < 4$. Finally, for the 4-leaf star tree, Settimi and Smith [2000] present a set of semi-algebraic (i.e. inequality) constraints including third and fourth-order moments:

$$\frac{4\sqrt{3}}{9} \geq |\sigma_{ijk}| \geq \frac{4(|\sigma_{ijl}\sigma_{ikl} - \sigma_{il}\sigma_{1234}|)\sigma_{jkl}^2}{(|\sigma_{ijl}\sigma_{ikl}|)^{\frac{1}{2}}(|3\sigma_{ijl}\sigma_{ikl} - 2\sigma_{il}\sigma_{1234}|)^{\frac{3}{2}}} \tag{4.1.23}$$

As mentioned in Section 4.1.1, the boundary points of these graphical regions represent zeros in the marginal tables of each manifest node, and occur where the inequality constraints are equalities. This is equivalent to having a manifest node equal to H. In the space of any three of the second-order central moments (see graphs in Section 4.1.4), the fold lines represent having zero entries in two of these tables, and each of the cusp points corresponds to zeros in all three

relevant tables. The maximum value of σ_{ijk} is (as stated earlier) $\frac{4\sqrt{3}}{3}$, and this occurs when $\sigma_{1234} = 0$ and $\sigma_{ij} = \sigma_{ik} = \sigma_{jk} = \frac{2}{3}$, which is in fact the degenerate distribution $X_i = X_j = X_k = H$. In summary, the boundary points of the geometry of the parameter spaces all relate to varying degrees of degenerate distributions.

From a geometric perspective, it is interesting to note that if the marginal data on the manifest nodes is inconsistent with the conditional independence model, then the sample estimates will correspond to points outside of these feasible regions. When considering second-order moments this can be visualised explicitly by plotting the tree-compatible regions of the second-order moment space and the second order sample moments. A visualisation of these regions is given in Section 4.1.4.

4.1.3 Beyond four observed variables

The derivation of constraints for trees involving five or more observed variables with a shared interior hidden node follows the same strategy as for lower order constraints. Thus for a tree with conditional independence

$$\perp\!\!\!\perp_{i=1}^{n} X_i | H$$

Settimi and Smith [2000] advocate the following the general procedure:

1. Write the n^{th} central moment $\sigma_{12\ldots n}$ of the observed variables in terms of in terms of central moments of a lower order than k and the non-central moment of degree n.

2. Write the non-central moment of degree n in terms of central moments of H and observed variables — this can be done as each X_i can be expressed as linear functions of H. Thus the n^{th} central moment can be expressed in terms of just central moments of the observed variables and H.

3. This can be repeated for each order k central moment where $k < n$. Thus a set of equations is formed that describes the central moment space of the model. These simultaneous equations can be solved using any standard mathematical software, and the final equations and inequality constraints are obtained through elimination of the moments of hidden variables.

Hence, Settimi and Smith [2000] provided a framework for deriving the applicable constraints for any n binary variables with a shared hidden binary variable. This introduces a significant number of extra constraints to assess for each extra observed variable. However, for a fixed number of observations, the higher the order of the moment the less confidence that may be placed in the estimate of the moment. There is a reason for this. The higher the order of the moment the more sensitive this moment is to tail behaviour or outliers (e.g. see Welling [2005]). Thus, due to reasons of reliability and practicality for larger models it may judged that only the lower order moments should be used for assessing compatibility.

4.1.4 Geometry of the tree-compatible regions

We have touched upon the geometric descriptions of the tree-compatible regions throughout this chapter but we have not yet provided a visual representation of these descriptions. Of course this is only possible for up to three moments at once and thus in general we are restricted to lower order constraints. We focus on the second-order tree constraints where we are able to plot the regions of the moment space that are consistent with the tripod model in Section 4.1.1 albeit for a fixed value of the third-order moment. By graphically representing the regions in three dimensions we get a more tangible understanding of the way that these constraints are acting on the low-order moments. For higher dimensions we do not have the privilege of being able to graphically represent the regions, though heuristically the lower order regions may aid understanding of analogous concepts in higher dimensions, for example boundaries and cusps.

Considering the second-order moments of the tripod constraints, the regions are determined by the union of inequalities seen in (4.1.8) intersected with (4.1.7). The third-order central moment can be varied through its possible range of values 0 to $\frac{4\sqrt{3}}{9}$. Plots for three different values of σ_{123} are presented in Figures 4.3–4.5 for $\sigma_{123} = 0, \frac{1}{9}, \frac{1}{3}$.

It has already been discussed that the regions only lie in the four multiplicatively non-negative octants (that is: (+,+,+), (+,-,-), (-,+,-) and (-,-,+)), that they are non-intersecting regions (except for the case where $\sigma_{123} = 0$ where the boundaries meet), and that they are symmetrical. The colours each represent the boundary of one of the three possible inequality constraints in (4.1.8) — the boundaries being defined by replacing the inequality sign with an equals sign.

From (4.1.15)-(4.1.17) it is clear that as μ_{123} is varied continuously from 0 through to $\frac{4\sqrt{3}}{9}$ the constraints become more demanding and so the size of the admissible regions decreases. Thus at $\mu_{123} = \frac{4\sqrt{3}}{9}$ there are only four admissible points in the parameter space of the second-order central moments, namely: $(\frac{2}{3}, \frac{2}{3}, \frac{2}{3})$, $(-\frac{2}{3}, -\frac{2}{3}, \frac{2}{3})$, $(-\frac{2}{3}, \frac{2}{3}, -\frac{2}{3})$, $(\frac{2}{3}, -\frac{2}{3}, -\frac{2}{3})$. At the other extreme when $\mu_{123} = 0$, the four previously separate regions are now joined at the single point $(0, 0, 0)$.

If observations are obtained and contingency tables formed for the model being discussed, the estimates of the central moments can be calculated. Once the third-order central moment has been calculated, the plot of the second-order central moment parameter space can be drawn. The co-ordinate of the second-order central moments estimates can then be marked on the plot. This may assist in signifying to what extent the observed data is within or outside the admissible region. The addition of a visual aid can provide more insight than simply checking the relevant constraint from (4.1.8). In particular, since we can see how the constraints might be violated — not just that they have been. If the marked point is only just outside (or inside) the admissible region, then this may suggest the evidence for rejecting (or not rejecting) the model is less strong.

4.1.5 Alternative parametrisations

It is useful to note that central moments are not the only way of expressing tree constraints. An alternative representation of the binary constraints is given in Zwiernik and Smith [2012] and Zwiernik and Smith [2011]. In the former paper, a novel reparametrisation is given for the binary latent tree model co-ordinate space changing the description from central moments to so-called tree cumulants, the full details of which can be found in Zwiernik and Smith [2012, Section 3.2]. In combination with the alternative binary coding $\{0, 1\}$ (as opposed to $\{-1,1\}$) this produces an elegant product-form parametrisation and is useful for obtaining the full semi-algebraic description of the model space (see the main result of the paper in Zwiernik and Smith [2011, Theorem 4.7]). However, as second-order and third-order tree cumulants are equal to the equivalent central moments, we can directly present Zwiernik and Smith [2011, Proposition 2.5]. This is equivalent to the results we report in Section 4.1.1 but under the binary random variable encoding $\{0, 1\}$.

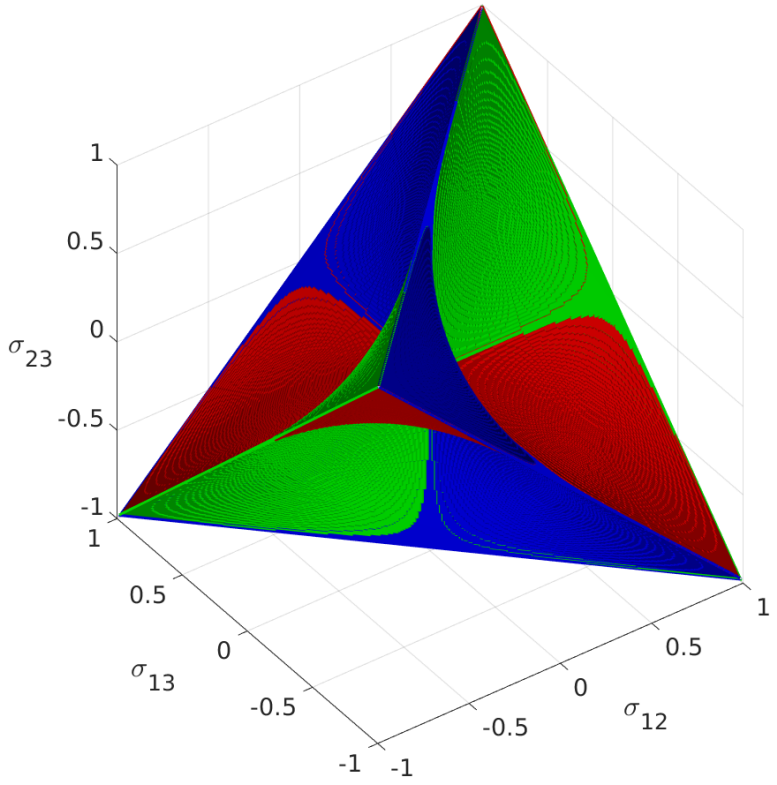


FIGURE 4.3: Covariance space of tripod tree T_3 for $\sigma_{123} = 0$.

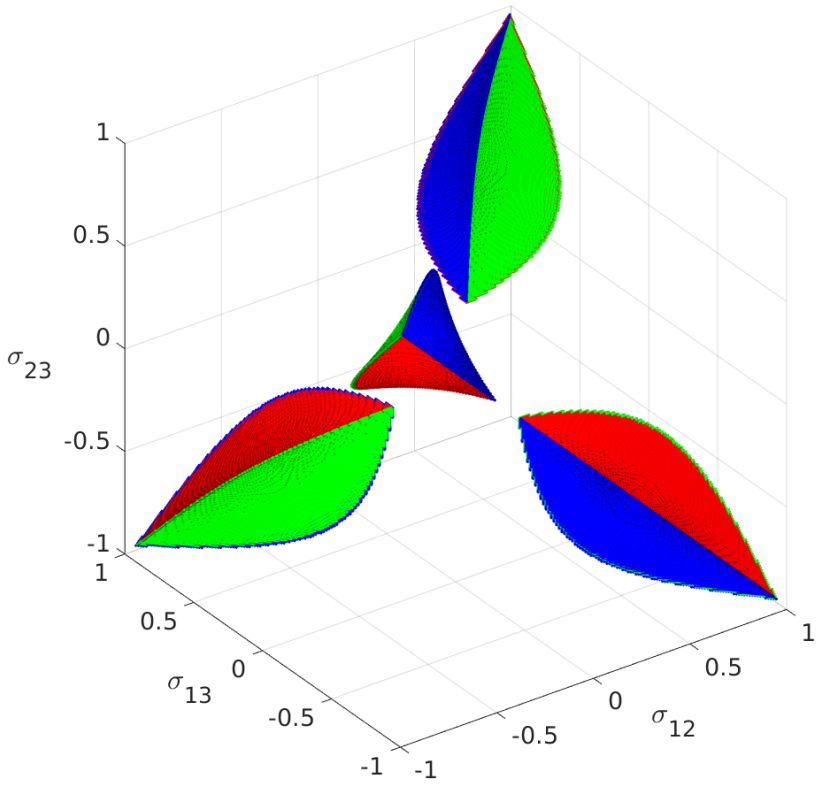


FIGURE 4.4: Covariance space of tripod tree T_3 for $\sigma_{123} = \frac{1}{9}$.

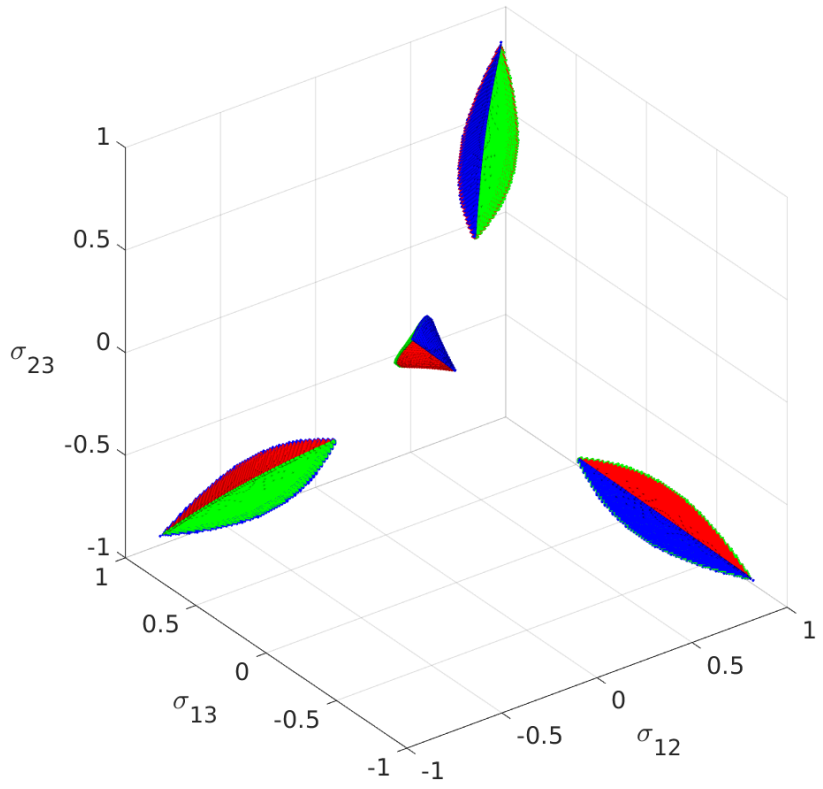


FIGURE 4.5: Covariance space of tripod tree T_3 for $\sigma_{123} = \frac{1}{3}$.

Proposition 4.1.1 (Proposition 2.5 from Zwiernik and Smith [2011])**.** *Given an observed joint probability table P ($2 \times 2 \times 2$), the data is consistent with the tripod tree structure if and only if:*

$\sigma_{123} = 0$ *and at least two of the three covariances* σ_{12}, σ_{13}, σ_{23} *vanish*

or

$$\sigma_{12}\sigma_{13}\sigma_{23} > 0 \tag{4.1.24}$$

$$|\sigma_{jk}|\sqrt{\det(P)} + \sigma_{123}\sigma_{jk} \leq (1 + 2(1 - \mu_i))\sigma_{jk}^2 \tag{4.1.25}$$

$$|\sigma_{jk}|\sqrt{\det(P)} - \sigma_{123}\sigma_{jk} \leq (1 - 2(1 - \mu_i))\sigma_{jk}^2 \tag{4.1.26}$$

where $\det(P) = \sigma_{123}^2 + 4\sigma_{12}\sigma_{13}\sigma_{23}$.

In Section 4.3 we introduce some graphical inequality diagnostics and when we apply these tools to some genetic data we will use the formulation of the tripod constraints given in Proposition 4.1.1.

4.2 Examples of tree constraint testing

We now present two examples that demonstrate how tree constraints can be used to assess tree-compatibility. The two applications are in linguistics and phylogenetics, both of which are topics known for using latent tree models for describing evolutionary relationships between species. The examples shown here should only be considered as illustrative. Although both applications explore questions typical to the subject areas and proceed in a sensible manner, without consultation with domain experts we cannot give much weight to the conclusions of the analyses. With this proviso noted, we proceed with the applications with particular note to the methodology that can carry across to other data sets.

4.2.1 Tree-compatibility of Indo-European languages using binary random variables

The aim of this analysis is to determine whether four selected Indo-European languages French, Italian, Spanish and Brazilian Portuguese can be adequately described using a binary latent tree model. This is achieved by checking whether the inequality constraints derived in Section 4.1.2 are respected by the sample estimates of the data. Of course, it is important that a suitable set of binary random variables are selected to begin with. The data set is a subset of that used in Nicholls and Gray [2008] (and is denoted Dyen et al. in Section 7.1 of that paper). The data set is based upon Dyen et al. [1997] which itself makes use of the famous Swadesh list of 200 word meanings [Swadesh, 1952]. The Swadesh list comprises word meanings that are known to have a low level of borrowing between languages — borrowing can be considered a linguistic equivalent of horizontal gene transfer in the genetic context. Thus the Swadesh list provides information about the historical relationships between languages largely focused on gradual evolutionary development. The data set was formed by taking one of 87 Indo-European languages and one of 200 word meanings and then identifying all words within that language with that particular meaning. This was then repeated for each language and meaning pair. Words with a shared meaning are said to be homologous if they are believed to share a common ancestor or origin. For each word meaning, words that are homologous (as judged by linguistic experts) are said to belong to the same cognate class. For example:

TABLE 4.1: Example of words with given meaning for each of four languages.

	'all'	**'to sit'**	**'to burn'**
Brazilian Portuguese	todo(to)	—	queimar
French	tout	asseoir	bruler
Italian	tutto	sedere	ardere, bruciare
Spanish	todo	sentasse	arder

TABLE 4.2: The corresponding cognate classes for the words in Table 4.1.

	'all'	**'to sit'**	**'to burn'**
Brazilian Portuguese	c=1	—	c=3
French	c=1	c=2	c=4
Italian	c=1	c=2	c=4, 5
Spanish	c=1	c=2	c=5

For the word meaning 'all' in English, an equivalent word was given in the data set for each of the four languages: 'tout', 'tutto', 'todo' and 'todo(to)'. It can be read from the data set that these four words are deemed homologous and so they share the cognate class denoted $c = 1$ (say). Now considering the verb 'to sit' we have the rare occurrence that the data set in this case does not provide a word for one of the languages (or in circumstances such a word does not exist). This means that there is no class code for Brazilian Portuguese for the word meaning 'to sit'. This occurs rarely in the data set and is thus unlikely to materially affect the analysis. The final example word meaning is 'to burn' where we have the occurrence that two words are provided for Italian: 'ardere' and 'bruciare'. There are three cognate classes containing {bruler, bruciare}, {ardere, arder} and {queimar}. Notice that there are two cognate classes associated with Italian and 'to burn'. The data set is presented as an 2665×87 binary data matrix $\boldsymbol{Z}$ where each row represents a cognate class $c = 1, \ldots, 2665$ and each column relates to one of the 87 languages. If a language j has a word in cognate class i then $\boldsymbol{Z}[i, j] = 1$ otherwise absence is indicated by a -1 or 0 depending on your binary coding choice. Hence, the data matrix relating

to the example in Tables 4.1 and 4.2 is

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 \end{bmatrix}$$

where the columns are ordered Brazilian Portuguese, French, Italian and Spanish.

We denote the submatrix of $\boldsymbol{Z}$ that contains the four languages of interest as $\boldsymbol{Z}^*$. We treat each of the languages as an observed variable with each of the cognate classes being considered as an observational unit. We can then assess whether $\boldsymbol{Z}^*$ is compatible with a binary latent tree model using the constraints described in Section 4.1.1 and Section 4.1.2.

The required central moment and mean estimates for checking for tree constraint violations are provided below (where the four languages are coded $1 =$ Brazilian Portuguese, $2 =$ French, $3 =$ Italian and $4 =$ Spanish).

$$\hat{\mu}_1 = -0.8507 \quad \hat{\mu}_2 = -0.8544 \quad \hat{\mu}_3 = -0.8454 \quad \hat{\mu}_4 = -0.8522$$

$$\hat{\sigma}_{12} = 0.1839 \quad \hat{\sigma}_{13} = 0.1991 \quad \hat{\sigma}_{14} = 0.2256 \quad \hat{\sigma}_{23} = 0.2101 \quad \hat{\sigma}_{24} = 0.1916 \quad \hat{\sigma}_{34} = 0.2097$$

$$\hat{\sigma}_{123} = 0.2951 \quad \hat{\sigma}_{124} = 0.2953 \quad \hat{\sigma}_{134} = 0.3152 \quad \hat{\sigma}_{234} = 0.3017$$

$$\hat{\sigma}_{1234} = 0.3470$$

The full set of inequality constraints were evaluated, namely (4.1.7), (4.1.8), (4.1.12)–(4.1.14) and (4.1.23). All of the constraints were satisfied with the exception of (4.1.23) that was violated. This suggests that the four languages are not tree-compatible, but that any three of the languages are indeed T_3-compatible as all of the tripod constraints are satisfied. We discuss some of the difficulties in making inferences about these types of violations in Section 4.2.3.

4.2.2 Assessing evolutionary history using the COI gene

Phylogenetic trees can be constructed using gene sequences where the vectors of data are obtained from DNA sequences coded into binary. Each base in a sequence has one of four chemicals either T or C (pyrimidines) or A or G (purines). Transversions (which is when a sequence jumps from a pyrimidine to a purine or vice versa) occur at a lower rate than transitions (jumps within pyrimidines and purines) and so transversions can be considered of more interest [Yang, 2007]. We can thus encode T and C as 1, and A and G as -1 [Vij and Biswas, 2005, p.8]. When fitting phylogenetic trees to data there has been some use of constraints implied by conditional independence (e.g. Casanellas and Fernández-Sánchez [2007]), however without the inequality constraints given in Settimi and Smith [2000], Zwiernik and Smith [2011] this can lead to erroneously fitting a tree to data. The sequences used for genetic analyses usually have hundreds of entries. For example, Barcode of Life Data Systems (BOLD Systems) requires sequences with a minimum of five hundred base pairs (BPs) [Ratnasingham and Hebert, 2007a].

The genetic data obtained from BOLD Systems [Ratnasingham and Hebert, 2007b] is from a particular region of the mitochondrial gene, cytochrome c oxidase I (COI). Hebert et al. [2004] published the first practical paper using this gene region suggesting the gene region "as a DNA barcode for the identification of animal species" and since then COI has had increasingly widespread use in animal species classification.

To illustrate the technique we consider the unresolved problem of how to model the evolution of placental mammals. For example, Teeling and Hedges [2013] recently surveyed the competing theories as to the ancestral root of placental mammals and found that despite advances in phylogenetic techniques and data sizes that there remain three serious possibilities. The disagreement is about the ordering of three groups of species called clades. A clade is a group of all species that are descendants of a common ancestor (and does not exclude any descendants). The three clades of interest are *boreoeutheria*, *xenarthra* and *afrotheria*.

In Teeling and Hedges [2013, Figure 1] the three proposed orderings of the clades are presented and Figure 4.6 here generalises the graph, the question then being the location of each clade in the positions A, B and C. However, this precludes the possibility that a latent tree model is not the appropriate model class. Reading the tree as a rooted BN, a necessary condition for

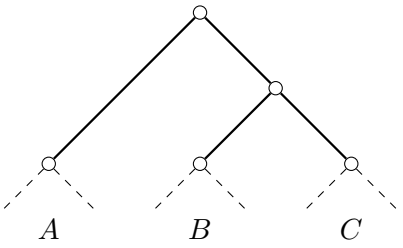


FIGURE 4.6: Outline of possible placental mammal phylogenetic tree where A, B and C are also trees and represent each of the three clades.

tree-compatibility is that tree constraints hold for any set of extant species within these groups. The minimum analysis involves selecting a species from each of these three clades and using the binary encoding of each COI gene to test them against tripod inequality constraints. It is this approach that we use in our example. The proof that the tripod constraints apply is provided by Theorem 4.3.1. This minimal analysis is undoubtedly a simplification but the principle is correct. A more detailed analysis might involve a larger number of species, more constraints, and more extensive sections of genetic data.

In our example we select *Pongo pygmaeus* (Bornean orangutan) from *boreoeutheria*, *Dasypus novemcinctus* (nine-banded armadillo) from *xenarthra* and *Loxodonta africana* (African bush elephant) from *afrotheria*. In this instance the choice of particular species was arbitrary though motivated by the images used in Teeling and Hedges [2013]. Clearly a large number of such choices could be tested in a full-blown analysis.

The species are coded $1 =$ orangutan, $2 =$ armadillo and $3 =$ elephant.

$$\hat{\mu}_1 = 0.1170 \quad \hat{\mu}_2 = 0.2226 \quad \hat{\mu}_3 = 0.3132$$

$$\hat{\sigma}_{12} = 0.4004 \quad \hat{\sigma}_{13} = 0.4049 \quad \hat{\sigma}_{23} = 0.2208$$

$$\hat{\sigma}_{123} = 0.1656$$

We find that the tripod constraints (4.1.7), (4.1.8) are satisfied and so based on this analysis the data set is T_3-compatible. Therefore, this offers no evidence against a tree adequately describing the evolutionary history of the species and thus one of the three rootings of the tree surveyed in Teeling and Hedges [2013] could be valid.

4.2.3 Discussions of examples

In these two examples we have motivated two contexts where a phylogenetic tree structure might be assessed and we demonstrated how to implement straightforward tests of tree-compatibility using up to fourth-order moments. However, it is worth considering the limitations of these analyses. Firstly, there has been no expert linguistic or phylogenetic guidance during the analysis as already mentioned. Secondly, the data sets are not definitive. Although the Swadesh list is a sensible choice it is just one of a number of lists, many of which have been constructed with more contemporary linguistic knowledge in mind. Likewise, the COI gene is only one of several genes that have been suggested for identifying species. With advances in technology it is possible for much larger numbers of BPs or even entire genomes to form the basis of genetic studies. Finally, we are only using point estimates for the moments and so ignoring any estimate error. It is interesting to note that for the linguistic example the only violations occur in the constraints that make use of fourth-order moments, and it is the higher order moment estimates that tend to be least accurate as argued earlier. To get a sense of reliability of these results there are several approaches that could be taken. For example, a non-parametric method would be to bootstrap the data and record the proportion of samples that adhere to each constraint. Alternatively, a prior distribution could be assigned to each estimate and a Bayesian hierarchical model could be constructed and through simulation a posterior probability of tree-compatibility can be estimated. These probabilistic methods could potentially be used to additionally incorporate the algebraic constraints though that is beyond the scope of these short examples. A search of current research suggest that this is the first example of this type of diagnostic to appear in the literature.

4.3 A first step into graphical inequality diagnostics

This section is based upon the text of the paper Shiers and Smith [2012]. Here we demonstrate how some tree constraints for binary random variables can be used for inference by providing the foundation for various diagnostics. These are primarily designed for the early stages of a phylogenetic analysis. The first point we note is that all trees must satisfy certain cubic inequalities associated with all the triples of its observed variables. A simple diagnostic is produced

which therefore simply checks whether data appears to fit with any tree structure or whether some of these inequalities appear to be violated. For if these constraints are significantly broken then we know that no tree will fit the data well and a search for a more elaborate graphical model for the data is needed. We also indicate how functions of the associated statistics can be used as a guide to selecting candidate tree models that are likely to score highly in model selection. In this way, these preprocessing tools can be used to help speed up and stabilise numerical scoring methods over the class.

In the literature, most efforts addressing inference and identifiability on tree models have focused on the algebraic geometry (see Drton and Sullivant [2007] and Allman et al. [2009]). However, we have found that there is great practical benefit in matching sample estimates against inequality constraints on the observed variables implied by the hidden tree structure (as illustrated in Section 4.2). In this paper, following suggestions of the mentioned authors, we focus on diagnostics based on the set of lower order constraints as they apply to all binary phylogenetic trees.

4.3.1 Two useful graphical results

The following theorem which has appeared previously, albeit as a corollary of more general theory in different forms and different contexts, is under-exploited but is particularly useful. The constraints we can use for a tree diagnostic can be derived simply in the following graphical way, which demonstrates that the tripod constraints are universal constraints.

Theorem 4.3.1. *Given any strictly trivalent phylogenetic tree T, for all triples X_i, X_j, X_k there exists a unique hidden variable H_{ijk} such that H_{ijk} separates X_i, X_j and X_k) in T. i.e.* $\perp\!\!\!\perp (X_i, X_j, X_k)|H_{ijk}$

Proof. Let X_i, X_j, X_k be any three manifest variables (leaves) on a phylogenetic tree T. Recall that $\overline{ab}$ denotes the path between X_a and X_b and furthermore that $\overline{abc}$ denotes a path between X_a and X_c with the path containing (at least) X_b.

By properties of a tree, there is exactly one path with no repeated edges $\overline{ij}$ and similarly one such path $\overline{ik}$ and $\overline{jk}$. Note that the intersection of the paths $\overline{ij}$ and $\overline{ik}$ has at least one hidden

vertex as it contains the hidden node adjacent to X_i. Denote the vertex in this intersection furthest from X_i as X_h, thus the intersection of $\overline{ih}$ and $\overline{kh}$ is the vertex X_h. Now consider the repeating path $\overline{jhihk}$. By removing the repeating nodes and their edges, a non-repeating subpath $\overline{jhk}$ is formed, with X_h being the only remaining node from the intersection of $\overline{ij}$ and $\overline{ik}$. Thus $X_h = H_{ijk}$ and furthermore H_{ijk} is unique as no other vertex appears on all three (non-repeating) paths $\overline{ij}$, $\overline{ik}$ and $\overline{jk}$. □

This result allows us to construct a diagnostic test to check whether any tree could be consistent with a sample data set (see Section 4.3.3). This method is not the only means of assessing tree structures (e.g. the retention index Farris [1989]) but has the advantage of being very simple to implement and is also well-grounded in theory. When the diagnostic does not reject the tree class, there is a second way in which the distributions of triples can be used to guide the search for promising candidate models.

First note that, by its definition, associated with every hidden variable $H \in \mathcal{H}$ of a strictly trivalent tree $\mathcal{T}$ is a partition $\Lambda(H, \mathcal{T})$ of the manifest variables into 3 subsets, each subset being the leaves of a subtree rooted at H. Interestingly, these partitions uniquely define a tree $\mathcal{T}$. Thus we have:

Theorem 4.3.2. *Each strictly trivalent tree T is uniquely identified by its set of partitions and*

$$\mathcal{X}(T) := \{\Lambda(H, T) : H \in \mathcal{H}\}$$

acts as an identifier, under the assumption of faithfulness (see Spirtes et al. [2001]).

The faithfulness property (see Definition 2.1.17 makes the assumption that the conditional independences in T map to those in the probability distribution. Also note that Theorem 4.3.2 is essentially a graph-topological theorem and thus we are interested in the underlying structure (i.e. the skeleton) and not concerned about the location of the root.

Proof. Let H' be a vertex in $\mathcal{T}$ which is a leaf of the subtree $\mathcal{T}^{\mathcal{H}}$ consisting of all hidden vertices and their connecting edges in $\mathcal{T}$. Then since $\mathcal{T}$ is strictly trivalent and H' is an interior vertex in $\mathcal{T}$, H' must be connected to two manifest vertices of $\mathcal{T}$ which we label X_{m-1} and X_m. Thus $\{X_{m-1}\}$ and $\{X_m\}$ are singletons in $\Lambda(H', \mathcal{T})$.

Suppose there exists m, the lowest number of leaves a strictly trivalent tree can have such that there exists nonisomorphic $\mathcal{T}_1$ and $\mathcal{T}_2$ with $\mathcal{X}(\mathcal{T}_1) = \mathcal{X}(\mathcal{T}_2)$. For $m = 3$ there is only one strictly trivalent tree, so $m \geq 4$. Now select $H' \in \mathcal{H}$ such that $\Lambda(H', \mathcal{T}_1)$ $(= \Lambda(H', \mathcal{T}_2))$ contains observed vertices X_{m-1}, X_m as singletons.

Note then that for all other $H \in \mathcal{H}$ by the definition of $\mathcal{X}(\mathcal{T}_1)$ (and $\mathcal{X}(\mathcal{T}_2)$) the pair $\{X_{m-1}, X_m\}$ are contained in the same subset of both partitions $\Lambda(H, \mathcal{T}_1)$ and $\Lambda(H, \mathcal{T}_2)$. So

$$\mathcal{X}(\mathcal{T}') = \{\Lambda(H, \mathcal{T}_i) : H \in \mathcal{H} \setminus \{H'\}, i = 1, 2\}$$

is isomorphic to the partition set $\{X_1, \ldots, X'_{m-1}\}$ where X'_{m-1} is identified with $\{X_{m-1}, X_m\}$.

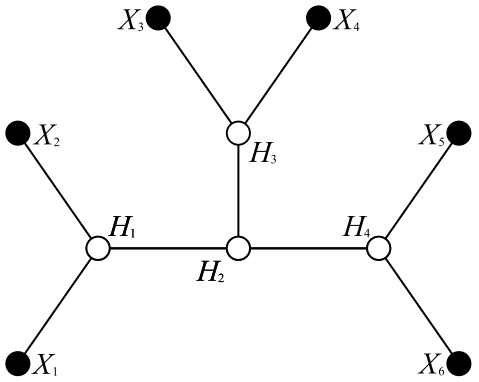
Now define trees $\mathcal{T}'_i$ from $\mathcal{T}_i$, $i = 1, 2$ each having $m - 1$ observed vertices: In $\mathcal{T}'_i$ replace $H' \in \mathcal{H}$ by a manifest variable X'_{m-1} then delete vertices X_{m-1}, X_m and their connecting edges.

By construction $\mathcal{X}(\mathcal{T}'_1) = \mathcal{X}(\mathcal{T}'_2)$, yet $\mathcal{T}'_1$ and $\mathcal{T}'_2$ have $m - 1$ manifest variables so by the definition of m they must be isomorphic. But then by construction $\mathcal{T}_1$ and $\mathcal{T}_2$ are also isomorphic. Thus no such m exists and we obtain our required contradiction. □

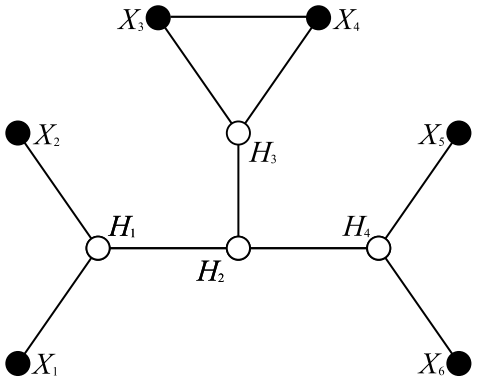
Thus if there exists a unique phylogenetic tree, then the estimated moments of the triple will identify it. We will show that this simple result allows us to preselect good candidate trees for model selection. See Section 4.3.4 for more details.

4.3.2 An illustration of the constraints

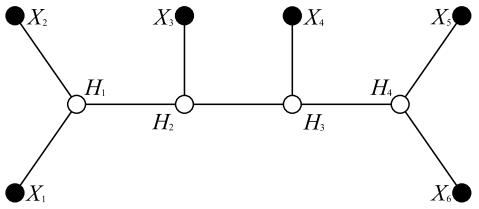
In this section we simulate binary data from the two non-isomorphic strictly trivalent trees with 6 leaves. Then from two other graphical models chosen to mirror typical variations which, for scientific reasons, we might expect of the tree. We adopted the binary coding $\{0, 1\}$ (in contrast to Section 4.2) which has no practical consequence other than the form of the tree constraints used. Unsurprisingly the empirical moments derived from the tree-generated data satisfy all the constraints whilst the empirical moments calculated from the non-tree data violates some of the constraints. The four graphs are shown in Figure 4.7.



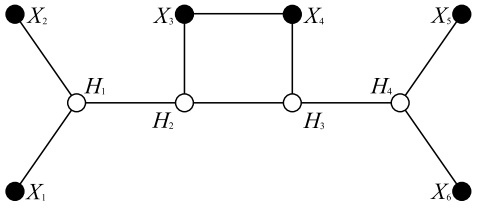
(a) 6-leaf Tree I.



(b) Non-tree I.



(c) 6-leaf Tree II.



(d) Non-tree II.

FIGURE 4.7: 6-leafed trees and non-trees.

The variations of the two trees were chosen to be similar in order to highlight any other differences. The probability distributions of the graphs were simulated so expected means for each X_i were the same across graphs. Because of this, the differences in the graphs may be expected to be exhibited through the differences in the higher order moments.

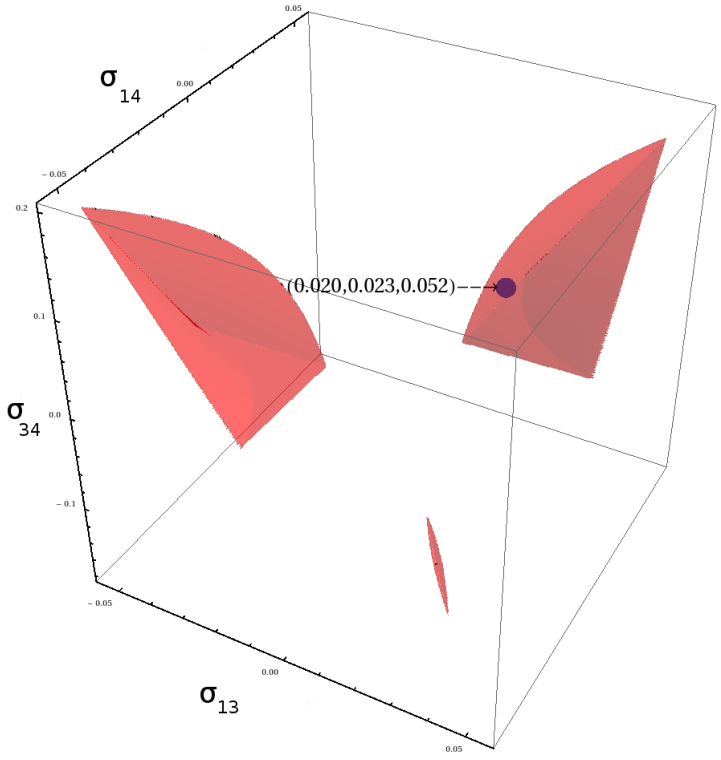
TABLE 4.3: Estimates of sample covariances for Figure 4.7a and Figure 4.7c.

Moment	**Tree I**	**Tree II**
$\hat{\sigma}_{12}$	0.054	0.054
$\hat{\sigma}_{13}$	0.020	0.025
$\hat{\sigma}_{14}$	0.023	0.023
$\hat{\sigma}_{15}$	-0.017	-0.013
$\hat{\sigma}_{16}$	-0.017	-0.013
$\hat{\sigma}_{23}$	0.037	0.048
$\hat{\sigma}_{24}$	0.044	0.044
$\hat{\sigma}_{25}$	-0.032	-0.025
$\hat{\sigma}_{26}$	-0.032	-0.025
$\hat{\sigma}_{34}$	0.053	0.041
$\hat{\sigma}_{35}$	-0.023	-0.023
$\hat{\sigma}_{36}$	-0.023	-0.023
$\hat{\sigma}_{45}$	-0.027	-0.035
$\hat{\sigma}_{46}$	-0.027	-0.035
$\hat{\sigma}_{56}$	0.150	0.150

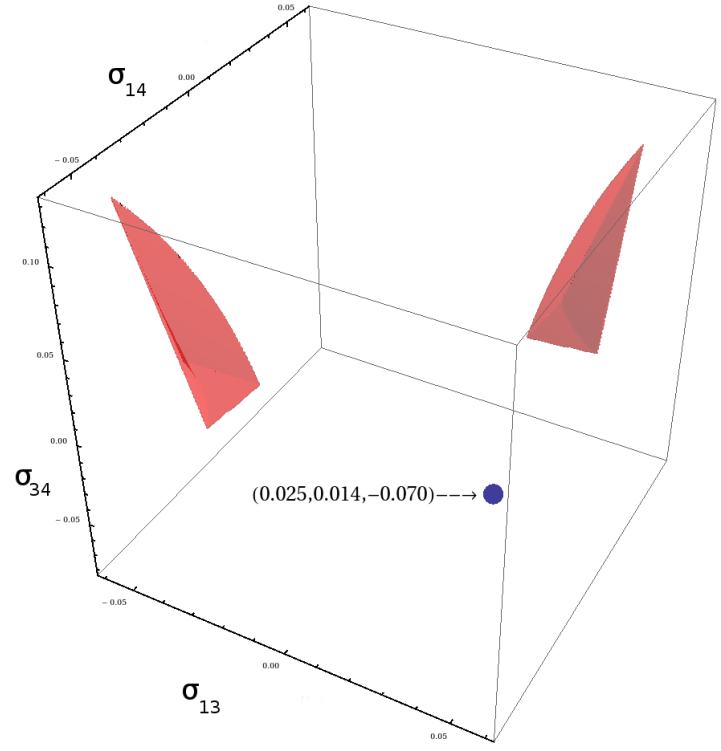
For large enough samples, the sample moments of the trees always satisfied the constraints demanded of their population analogues whilst the non-tree modifications did not. Both non-trees experienced the same violations: (X_1, X_3, X_4) and (X_2, X_3, X_4) for (4.1.25), (X_3, X_4, X_5) and (X_3, X_4, X_6) for (4.1.26), and all versions of (4.1.24) which involve X_{34} (for example $X_{13}X_{14}X_{34} < 0$). Note that all violated constraints include both indices 3 and 4 (where the additional edge is found). Moreover, every inequality involving indices 3 and 4 is broken. This suggests that these diagnostics could provide more insight than a binary acceptance or rejection of the tree structure; the violated constraints in these cases hint that the random variables X_3 and X_4 may be responsible for the exceptions.

The geometries of the admissible regions of the covariances $\sigma_{13}, \sigma_{14}, \sigma_{34}$ are similar between the trees, and similar between the non-trees. The graphs of the admissible covariance regions are shown for Figure 4.7a and Figure 4.7b in Figure 4.8a and Figure 4.8b respectively, along with the estimates of the observed covariances plotted. The sample covariances are revisited in Section 4.3.5.

The sample covariances of the two 6-leafed trees are compared in Table 4.3. When considering the absolute values of these moments we notice that for Tree I, $\hat{\sigma}_{56}$, $\hat{\sigma}_{12}$ and $\hat{\sigma}_{34}$ have the greatest magnitudes. These relate to the pairs of observed nodes topologically closest to each other. The



(a) Triple 134 - Tree I. The sample moments lie within the tree-compatible region which provides evidence towards tree-compatiblity of the data.



(b) Triple 134 - Non-tree I. The sample moments lie well outside the tree-compatible region which is consitent with the data being non-tree generated.

FIGURE 4.8: Plots of covariance point estimates.

same can be said of Tree II for $\hat{\sigma}_{56}$ and $\hat{\sigma}_{12}$, but now $\hat{\sigma}_{23}$ and $\hat{\sigma}_{24}$ rank above $\hat{\sigma}_{34}$ in magnitude. This appears to reflect the structure of Tree II where X_3 and X_4 are not directly joined to the same vertex. This alludes to a potential method of narrowing a set of tree models based on magnitudes of sample covariances. However, further model simulation and a greater variety of joint probabilities would need to be considered before any secure inference could be made about this. In Chapter 6, we instead consider Gaussian random variables and are able to perform inference on specific trees using probabilistic methods. It may be possible for equivalent methods to be constructed for binary random variables which would then be complementary to the graphical diagnostics.

4.3.3 Application of diagnostics

To illustrate the use of diagnostics we again use mitochondrial genetic data from the COI gene as we did in Section 4.2.2. Each sequence is of length 883 BPs and obtained from BOLD Systems 3 [Ratnasingham and Hebert, 2007b] for six species from the class Mammalia:

X_1 — *Ailurus fulgens* (red panda)

X_2 — *Procyon lotor* (raccoon)

X_3 — *Ailuropoda melanoleuca* (giant panda)

X_4 — *Ursus maritimus* (polar bear)

X_5 — *Tremarctos ornatus* (spectacled bear)

X_6 — *Ursus malayanus* (sun bear)

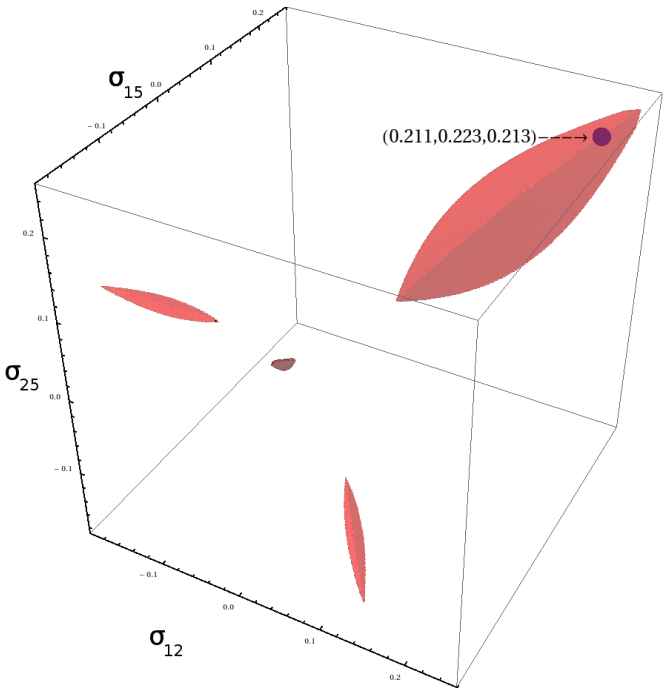
The BPs are encoded into binary similarly to Section 4.2.2 but with A and G as 0 and T and C as 1 to reflect the choice of $\{0, 1\}$ over $\{-1, 1\}$. Note that this mean we must use the results from Proposition 4.1.1 to match this parametrisation.

The genetic data did violate some constraints for five of the triples: (X_4,X_6,X_k) for $k = \{1, 2, 3, 5\}$, and (X_2,X_4,X_5). Figure 4.9a shows a covariance triple point within the admissible region, and contrastingly Figure 4.9b shows a different sample covariance triple outside

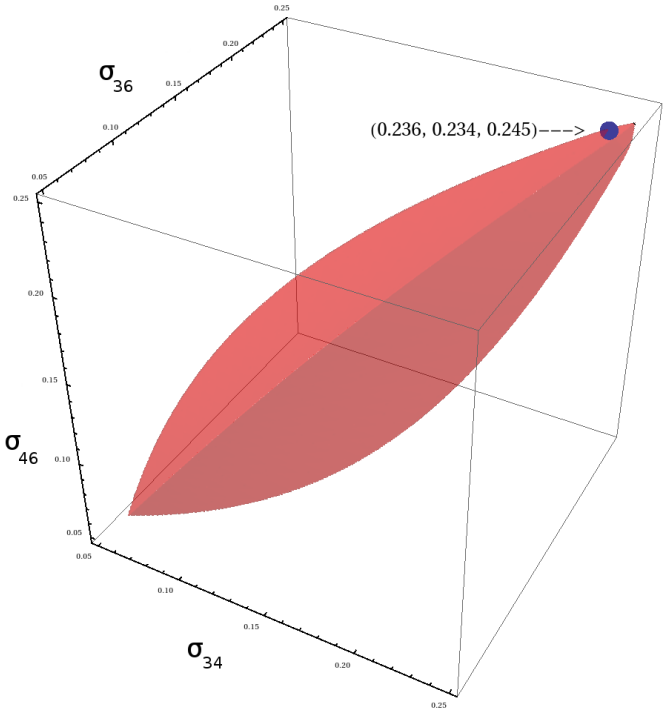
the region associated with a tree. Superficially, this suggests that a tree is not a suitable model for the data. However, because the sample size is not large these violations could be attributed to random error, particularly given the closeness of the point in Figure 4.9b to the boundary of the admissible region. This is explored in Section 4.3.4.

However, one observation specific to the above data is that unlike the simulated data violations, $\hat{\sigma}_{ij}\hat{\sigma}_{ik}\hat{\sigma}_{jk} > 0$ is not violated. Because this constraint only involves lower order moments, this might suggest that the evidence for violations existing might be considered weaker.

At this stage of course, it is important to remember that the types of analyses conducted here are simply a demonstration of how the constraints could be used. The diagnostics are so far simply indicative rather than definitive.



(a) Triple 125. The sample covariances relating to the red panda, raccoon and spectacled bear, lie within the tree-compatible region suggesting a tree model may be appropriate given the observed data.



(b) Triple 346. This plot of the three sample covariances between the giant panda, polar bear and sun bear lies just outside the tree-compatible region. Given the closeness to the boundary, the evidence is less clear as to whether the tree inequality constraint has actually been violated.

FIGURE 4.9: Point estimates of covariances.

4.3.4 The effect of sample size

We now examine the likely power of these diagnostics for studies like those given below using data generated from the graphs in Figure 4.7 but now using smaller samples. Since it is known whether the graphs are trees or not, for those that are trees we know that any violations of the constraints must be due to estimation error of moments. The inequalities can be tested for each of the four graphs using different sample sizes ($n = 500$, 883, 1500, 5000), and repeated 10^4 times. For the 20 combinations of triples on each graph the number of triples which violate at least one constraint is recorded. This is relevant as even for a tree, some violations (depending on the sample size) will be expected due to noise. The frequencies of these violations are displayed in Tables B.1–B.4 in Appendix B. A comparison of these results is shown in Figure 4.10 for Tree I and Non-tree I from Figure 4.7.

It is noticeable at a sample size of 883 (as in Section 4.3.3) there is a clear shift between the

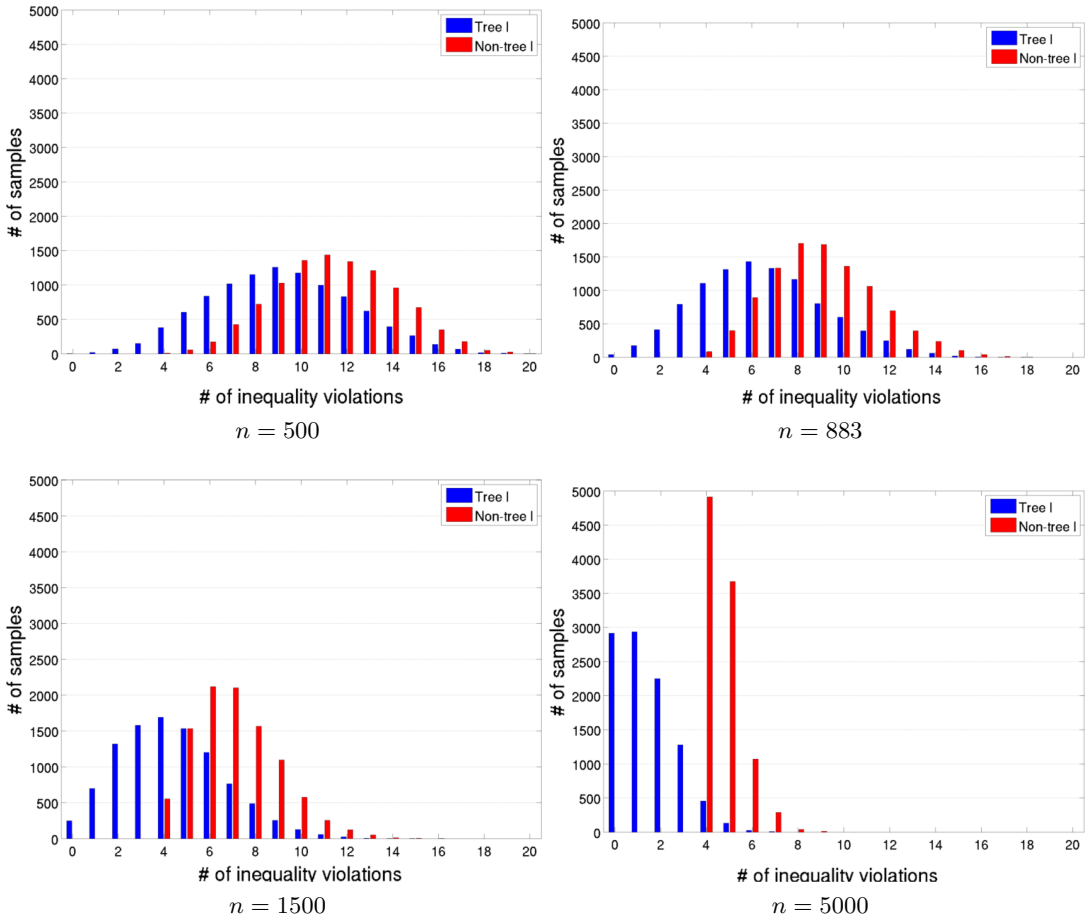


FIGURE 4.10: Frequency of violations on a tree (blue) and non-tree (red) for 4 sample sizes.

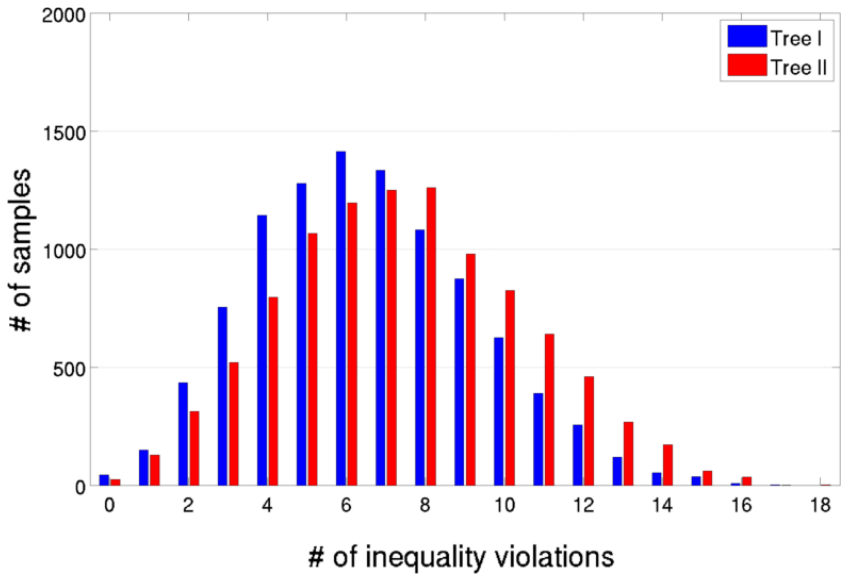


FIGURE 4.11: Frequency of violations ($n = 883$).

modal violations in the non-tree model relative to the tree model. If 5 or fewer violations are observed Tree I seems the more likely model, and if 11 or more then Non-tree I appears the favourable hypothesis. Note that the frequencies for the non-tree are likely to be conservative (shifted to the left) compared to more complex non-trees, and so it might be expected that results are often even clearer in practice.

We compared the two six-leafed tree models and noted that they performed similarly in terms of number of violations. This can be seen in particular for sample size 883 in Figure 4.11. More generally though, the sample size required for these diagnostics to be useful is likely to vary depending on the topology of the models, as well as the underlying true joint probability table. However, in the examples we have looked at, a large number of violations would suggest a non-tree model even for a moderate sample size.

Returning to the application in Section 4.3.3 where 5 violations were observed, the frequency of violation plots can be studied. Figure 4.12 shows the plot for Tree I and Non-tree I, where the arrow indicates the frequencies for 5 violations. If the plot is thought to be typical of other distributions on trees of this size, the level of violations would suggest that the gene sequence data is more likely to conform to a tree than a non-tree. The compared non-tree is from the simplest subset of non-trees (only one additional edge). A more complex tree might be expected to produce more violations and so additional weight may be given to the hypothesised tree.

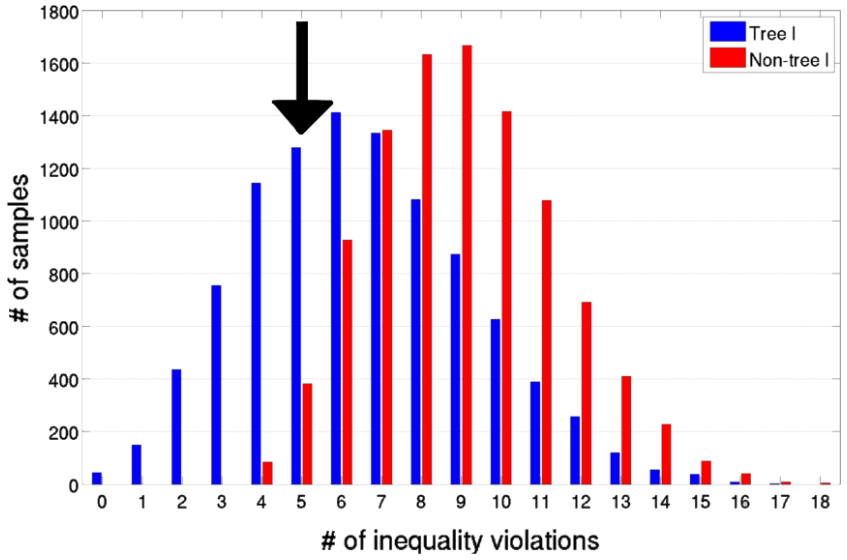


FIGURE 4.12: Frequency of violations ($n = 883$).

4.3.5 Inferring trees from moments

It can be shown that the first 3 moments provide us with consistent but inefficient estimates of a tree. In Settimi and Smith [1998] it was shown that (provided none of the terms are zero) for any triple X_i, X_j, X_k with $(X_i \perp\!\!\!\perp X_j \perp\!\!\!\perp X_k)|H_{ijk}$

$$S_{ijk} = \ln(|\mu_{ij}|) + \ln(|\mu_{ik}|) + \ln(|\mu_{jk}|) - 2\ln(|\mu_{ijk}|)$$

where S_{ijk} (the signature) depends only on the marginal distributions of the triple. From Theorem 4.3.2, for large enough data sets the signatures of the corresponding sample quantities will indicate candidate tree partitions $\mathcal{X}(\mathcal{T})$ and hence, from the theorem, candidates $\mathcal{T}$. Note that triples (i, j, k) and (i', j', k') share the same separator H in $\mathcal{T}$ if the pairs (i, i'), (j, j') and (k, k') all lie in different subsets in $\Lambda(H, \mathcal{T})$. So these $n - 2$ partitions can be calculated by first clustering the signatures by magnitude into up to $n - 2$ clusters and from this deducing $\Lambda(H, \mathcal{T})$. For small trees this method, simply using the statistics already calculated for our first diagnostics, allows us to identify some promising trees.

Figure 4.14 shows the standardised signatures S^*_{ijk} for both trees in Figure 4.7 with the signatures of interest labelled. The clustering of the signatures demonstrates the power of the diagnostic - there is remarkably clear separation of all S^*_{12k} and S^*_{i56} which supports the topologies of the trees. The signatures involving 3 and 4 are less distinct, but this may be in part

unavoidable overlapping of true clusters. However, trees like Tree I have an interior node and it can be shown that this leads to signatures with a higher variance. The wide spread in Tree I of the middle cluster (overlapping) with the X_3 X_4 cluster is indicative of this.

When applied to the genetic example data set we get strong clustering of signatures involving the raccoon and giant panda, and some clustering of polar bear and sun bear - the signatures for the remaining species (red panda and spectacled bear) are dispersed. This diagnostic would thus suggest Tree II as a starting point for an analysis fitting the data, with the latter species being the singletons.

Finally, the sample covariances $\hat{\mu}_{ij}$ themselves can provide some indication of the topology of the tree if correlations between each manifest variable and its hidden parent are of about the same magnitude. For then $\hat{\mu}_{ij}$ tends to be higher when the number of edges between these vertices is fewer (as utilised in Harmeling and Williams [2011]). For example, considering Section 4.3.2, for Tree I $\hat{\mu}_{56}$, $\hat{\mu}_{12}$ and $\hat{\mu}_{34}$ have the greatest absolute values and for Tree II $\hat{\mu}_{56}$ and $\hat{\mu}_{12}$ have greatest magnitude, but now $\hat{\mu}_{23}$ and $\hat{\mu}_{24}$ rank above $\hat{\mu}_{34}$ in magnitude. This appears to reflect the structure of Tree II where X_3 and X_4 do not share a common parent.

It follows that non-metric MDS can be used on a function of sample covariances, allowing the relationships to be displayed graphically. Here we use the function $\delta_{ij} = -2\ln(|\hat{\mu}_{ij}|)$.

The resulting plots (Figure 4.13) for a 2D scaling relate to the trees in Figure 4.7. They were generated using a sample sizes of 883, which gives a relevant comparison with the gene data (see Figure 4.15). The size of the plotted points relates to the number of times a variable is involved in a violation; the more violations, the smaller the marker. More precisely, inequality violation is expressed through the size of the plotted points. The relative sizes are determined via $(1 + V_i)^{-1/2}$ where V_i is the number of violations for variable X_i.

For Tree I, the MDS indicates clear pairings (X_1, X_2), and (X_5, X_6), plus X_3 and X_4 are reasonably close which suggests these three pairs are a distance of 2 edges apart giving us the topology here. However in Tree II X_3 and X_4 are further apart. So these might be conjectured as singletons (giving us Tree II). Note that this ambiguity may be caused in part by some $|\mu_{ij}|$ being close to zero.

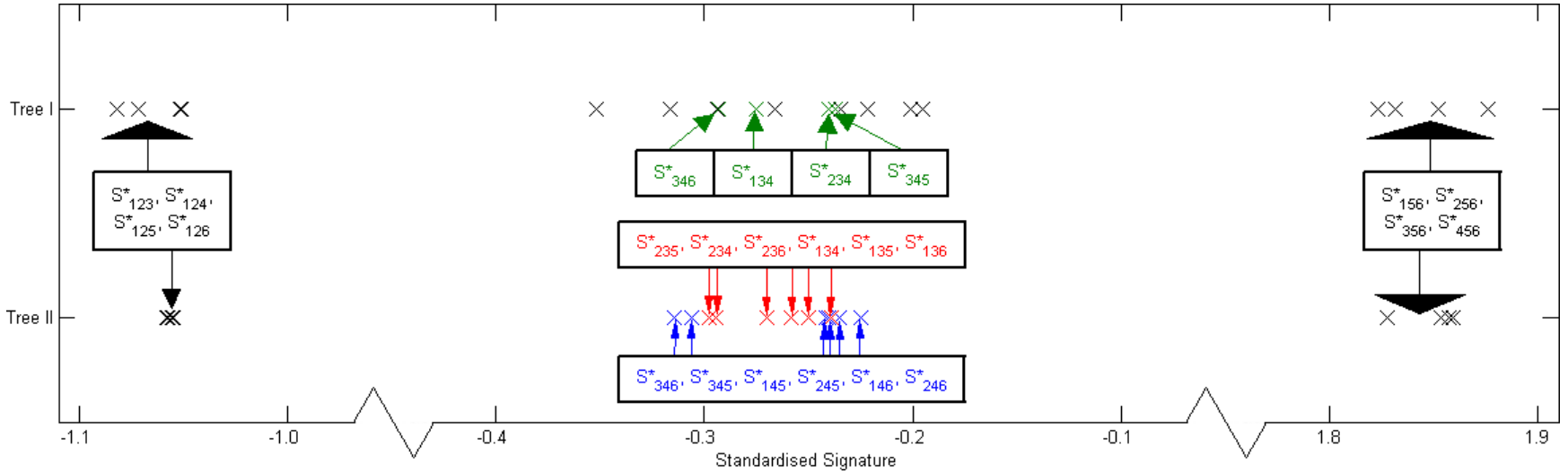
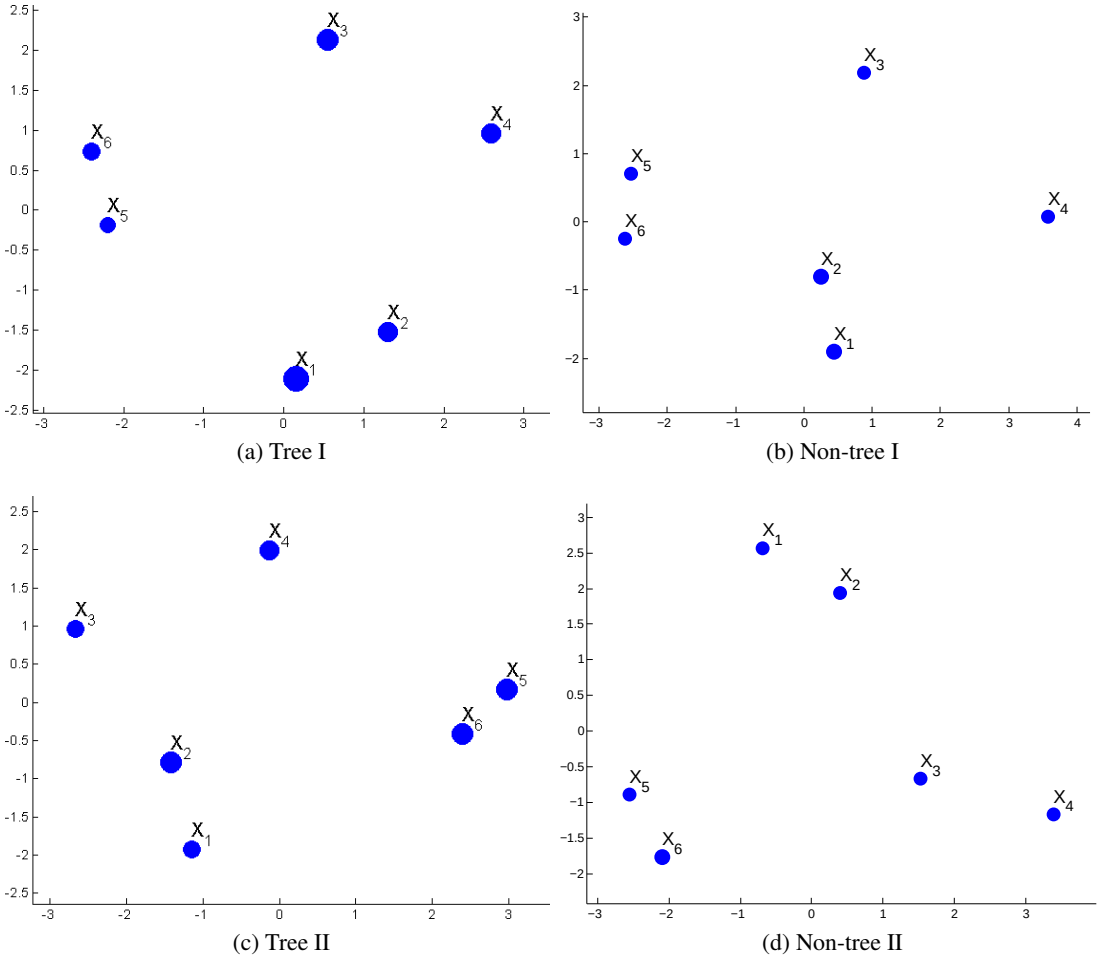


FIGURE 4.14: Plot of standardised signatures for Tree I and Tree II.

Figure 4.15 is the result of performing the same MDS for the genetic data with the hope that the plot indicates one of the two six-leafed trees. Unlike for the previous plots, the points are separated into two groups of three. This matches the form of Tree II, and although admittedly some of the distances are similar it could be hypothesised that X_2 and X_4, and X_5 and X_6 are pairs (i.e. they are both connected to a common hidden vertex), with X_1 and X_3 as singletons (i.e. they are connected to a hidden vertex and no other manifest variable is).



(a) Tree I (b) Non-tree I

(c) Tree II (d) Non-tree II

FIGURE 4.13: MDS for trees in Figure 4.7.

It is interesting to note that the ambiguity is reflected in a more detailed analysis of this data set. We are currently applying these simple methods described to preselect good trees with the hope of achieving a large time saving at little cost to accuracy. We can then use the subset of trees as a starting point for MCMC likelihood-based model selection algorithms, which otherwise often get stuck within local maxima (e.g. see Chor et al. [2000]). However, a discussion of these applications is beyond the scope of the analysis presented in this thesis.

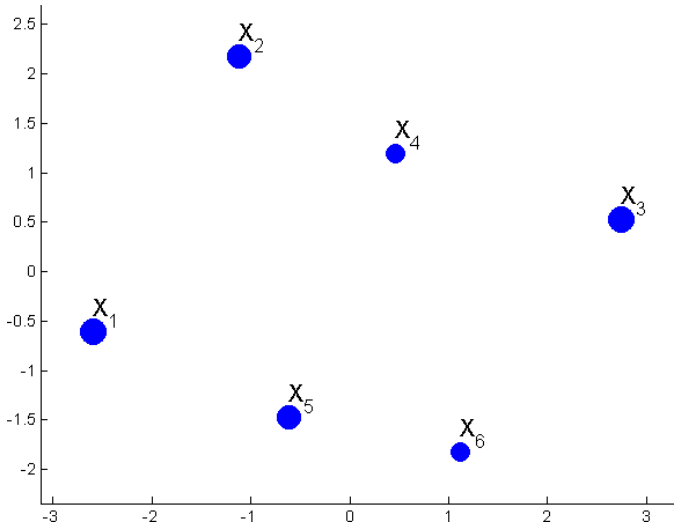


FIGURE 4.15: MDS for gene data.

4.3.6 Discussion of graphical inequality diagnostics

In this section we have illustrated how some simple graphical properties of trees allow us to construct useful diagnostics based on inequality violations and certain functions of sample moments up to order 3. The diagnostics are trivial to calculate and for the sizes of data sets used in phylogenetic trees (now typically 1500 BPs if using BOLD Systems) provide a relatively powerful method. In particular the inequality diagnostics are complementary to the algebraic methods developed by Drton and Sullivant [2007] which are based on different functional constraints. We are currently developing analogous inequality diagnostics using the same graphical properties but for differently distributed variables.

One appealing and unusual feature of our methods based on low order moments is that as we add more species to the putative tree we simply need to check the new triples introduced by the additional manifest variables. So in this sense it scales up. Furthermore if violations of the tree structure are discovered our methods also sometimes allow us to identify a subset of manifest variables on which a tree is valid. So for example in the two non-trees of Figure 4.7 we can still deduce that a tree might be valid on X_1, X_2, X_5, X_6 since no violations of the inequality constraints are apparent.

Of course, rather than simple singularities it is important to develop more general theory for tree diagnostics so that they can be exploited routinely. But even in this naive form, our methods

appear surprisingly effective.

4.4 Extensions beyond binary variable trees

In a discrete setting, the natural extension beyond binary variables is k-state variables. For many genetic applications this can be motivated by the desire to code BPs into $k = 4$ states instead of $k = 2$ binary. But the benefit is clear in any setting that requires more than two states. Under mild conditions such as non-singularity of parameters and strictly positive parameter values, Allman et al. [2014] derive a semi-algebraic description of the k-state general Markov model for trees with n leaves. In the paper, the joint distribution of the model leaves is denoted P and is considered as an n-dimensional $k \times \cdots \times k$ tensor. This ultimately allows the main result Allman et al. [2014, Theorem 5.7] to be presented in terms of conditions on functions of P such as principal minors and ranks. A search of the literature has not identified a use of this result in any applications yet and so there is an opportunity to investigate the challenges of implementation and develop suitable probabilistic diagnostics to assess tree-compatibility.

An obvious expansion of the existing tree-constraint literature is from discrete variables to continuous, and in Chapter 5 we shall derive a set of tree constraints for GLTMs. It turns out that there are some parallels with the presentations of the descriptions for the k-state n-tree and the Gaussian n-tree (Allman et al. [2014, Theorem 5.7] and (in this thesis) Theorem 5.3.4 respectively). In the former, two-thirds of the theorem is expressed in terms of tripod tree constraints and quartet tree constraints; in the latter the conditions are solely in terms of tripod and quartet trees results. This indicates the universal nature of the tripod tree constraints as well T-specific quartet trees for trees with at least $n = 4$ leaves — both are fundamental components of graphical tree models which heuristically explains their appearance in both theorems.

Chapter 5

Gaussian tree constraints

In this chapter we derive a complete description of the space of GLTMs in terms of correlations between observed variables. This is achieved by taking advantage of the link between tree metrics and the space of phylogenetic oranges. The purpose of obtaining the algebraic and semi-algebraic description of the space is to be able to construct tools for assessing whether a particular data set is compatible with the class of GLTMs. We will then apply this in Chapter 7 to applications that are typically implicitly modelled by GLTMs. This chapter is based upon the first half of the paper Shiers et al. [2016].

5.1 A constraint on the covariance of Gaussian latent tree models

To introduce the idea of Gaussian tree constraints we demonstrate a direct and explicit method of deriving the most fundamental of the constraints. Although the most easily attainable of the constraints it has not been widely exploited for the purpose of investigating tree-compatibility.

Consider the precision matrix Σ^{-1} related to the tripod tree in Figure 5.1 with one latent and three manifest Gaussian random variables.

It is well known in the Gaussian setting that X_i is independent of X_j conditional on all other observed variables if and only if the corresponding entry $\Sigma^{-1}_{ij} = 0$ (see Lauritzen [1996, Chapter 5] for example). By Gaussian setting, we refer to the GLTM as given in Definition 5.2.2 and

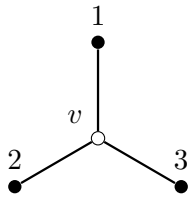


FIGURE 5.1: Tripod tree.

again in Definition 2.1.27. Thus the precision matrix for the univariate Gaussian tripod tree has the form:

$$\boldsymbol{\Sigma}^{-1} = \left(\begin{array}{ccc|c} \tau_1 & 0 & 0 & \tau_{1H} \\ 0 & \tau_2 & 0 & \tau_{2H} \\ 0 & 0 & \tau_3 & \tau_{3H} \\ \hline \tau_{H1} & \tau_{H2} & \tau_{H3} & \tau_H \end{array}\right)$$

where the final row/column relates to the interior hidden variable, $\tau_{Hi} = \tau_{iH}$ for $i \in \{1,2,3\}$ and $\tau_{ii} = \tau_i$ for $i \in \{1,2,3,H\}$. The covariance matrix resulting from taking the inverse of the precision matrix can be expressed algebraically in terms of entries of Σ^{-1}:

$$\boldsymbol{\Sigma} = \frac{1}{\tau_1\tau_2\tau_3}\left(\begin{array}{ccc|c} \frac{\tau_{1H}^2+\tau_2\tau_3\tau_1^2}{\tau_1^2} & \frac{\tau_{1H}\tau_{2H}}{\tau_1\tau_2} & \frac{\tau_{1H}\tau_{3H}}{\tau_1\tau_3} & -\frac{\tau_{1H}}{\tau_1} \\ \frac{\tau_{1H}\tau_{2H}}{\tau_1\tau_2} & \frac{\tau_{2H}^2+\tau_1\tau_3\tau_2^2}{\tau_2^2} & \frac{\tau_{2H}\tau_{3H}}{\tau_2\tau_3} & -\frac{\tau_{2H}}{\tau_2} \\ \frac{\tau_{1H}\tau_{3H}}{\tau_1\tau_3} & \frac{\tau_{2H}\tau_{3H}}{\tau_2\tau_3} & \frac{\tau_{3H}^2+\tau_1\tau_2\tau_3^2}{\tau_3^2} & -\frac{\tau_{3H}}{\tau_3} \\ \hline -\frac{\tau_{1H}}{\tau_1} & -\frac{\tau_{2H}}{\tau_2} & -\frac{\tau_{3H}}{\tau_3} & \frac{1}{\tau_H-\tau_{1H}^2\tau_{2H}^2\tau_{3H}^2} \end{array}\right)$$

Noting that the diagonal entries $\tau_1, \tau_2, \tau_3, \tau_H$ in $\boldsymbol{\Sigma}^{-1}$ represent the reciprocals of the conditional variances and so are non-negative (and non-zero for non-degenerate model), the following constraint can be derived:

$$\sigma_{12}\sigma_{13}\sigma_{23} = \frac{1}{(\tau_1\tau_2\tau_3)^3}\frac{\tau_{1H}\tau_{2H}}{\tau_1\tau_2}\frac{\tau_{1H}\tau_{3H}}{\tau_1\tau_3}\frac{\tau_{2H}\tau_{3H}}{\tau_2\tau_3} = \frac{\tau_{1H}^2\tau_{2H}^2\tau_{3H}^2}{(\tau_1\tau_2\tau_3)^5} \geq 0$$

Since correlations are simply covariances scaled by standard deviations we can immediately deduce that $\rho_{12}\rho_{13}\rho_{23} \geq 0$. Therefore, we have derived a semi-algebraic constraint of the Gaussian latent tripod model in terms of observed correlations. Moreover, considering the importance of the tripod tree as a fundamental part of any binary tree (recall Theorem 4.3.1), for any binary tree with n_l observed leaf nodes, there are $\binom{n_l}{3}$ tripod trees that must be valid — one for each

triple — and thus $\binom{n_l}{3}$ such positivity constraints. Therefore:

$$\rho_{ij}\rho_{ik}\rho_{jk} \geq 0 \,\forall\, 1 \leq i < j < k \leq n_l. \tag{5.1.1}$$

We denote this equation the positivity constraint. The algebraic manipulation required is quite involved to obtain a suitable factorisation of $\mathbf{\Sigma}$. Thus to derive further constraints another more subtle approach is considered.

5.2 A formal description of the model

To give a thorough derivation of tree Gaussian latent tree constraints we provide a more formal description of the set-up. Let $Z = (Z_u)_{u\in U}$ be a random vector whose components are indexed by the vertices of an undirected tree $T = (U, E)$ with edge set $E \subset U \times U$. The tree T induces a Gaussian tree model $N(T)$ for Z that is a Gaussian graphical model on T [Lauritzen, 1996, Section 5.2]. For two nodes $u, v \in U$, let $\overline{uv}$ denote the (unique) path between u and v. Then the model $N(T)$ is the collection of all multivariate normal distributions on $\mathbb{R}^{|U|}$ under which Z_u and Z_v are conditionally independent given a subvector Z_C whenever the set $C \subset U \setminus \{u, v\}$ contains a node on $\overline{uv}$. It follows that a normal distribution with correlation matrix $R = (\rho_{uv})$ belongs to $N(T)$ if and only if

$$\rho_{uv} \quad = \quad \prod_{e\in\overline{uv}} \rho_e \qquad \text{for all } u, v \in U,$$

where $\rho_e := \rho_{uv}$ when e is the edge (u, v). To obtain this equation, note that for three nodes $u, v, w \in U$ the conditional independence of Z_v and Z_w given Z_u is equivalent to $\rho_{vw} = \rho_{uv}\rho_{uw}$ [Wright, 1921].

As motivated, we are concerned with GLTMs in which only the tree's leaves correspond to observed random variables. In this case the set of leaves is denoted by V. For example, considering the Romance languages, a possible evolutionary tree is displayed in Figure 5.2 with extant languages as leaves.

Definition 5.2.1. *Let V be a finite set. We say that $T = (T, \phi)$ is a semi-labelled tree with the underlying tree $T = (U, E)$ and labelling map ϕ if $\phi : V \to U$ is such that every vertex of T*

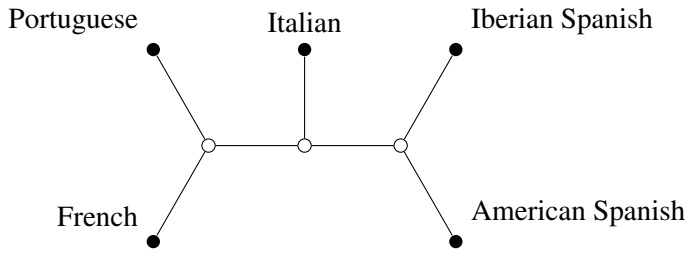


FIGURE 5.2: Quintet tree relating five Romance languages.

with degree ≤ 2 is necessarily contained in the image $\phi(V)$. In other words ϕ is any labelling of nodes of T with the labelling set V such that degree ≤ 2 vertices are labelled. We say that T is a phylogenetic tree if ϕ is a bijection between V and the leaves of T.

For examples, see Figure 5.3.

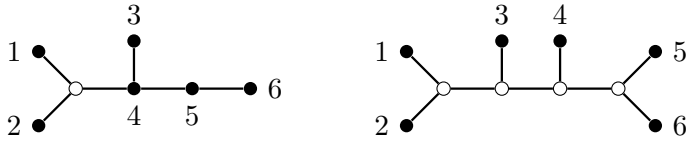


FIGURE 5.3: On the left - a semi-labelled tree with the labelling set $\{1, 2, 3, 4, 5, 6\}$. On the right - a binary phylogenetic tree with six leaves.

Definition 5.2.2. *The GLTM $M(T)$ for the subvector $X := (Z_v)_{v \in V}$ is the set of all V-marginal distributions of the distributions in $N(T)$, where the V-marginal distributions are those associated with leaf variables.*

The parametrization of $M(T)$ is induced from the parametrisation of $N(T)$ and given by

$$\rho_{ij} \quad = \quad \prod_{e \in \overline{ij}} \rho_e \qquad \text{for all } i, j \in V. \tag{5.2.1}$$

As the variances σ_{uu} for $u \in U \setminus V$ never appear in the parametrization, without loss of generality, we may assume that they are equal to 1. Because there are no constraints on the variances σ_{vv} for $v \in V$, we may consider the standardised version of X. Thus, from now on, Σ denotes a correlation matrix and furthermore we consider $M(T)$ to have the above parametrization restrictions.

We proceed by relating $M(T)$ to the space of phylogenetic oranges [Engström et al., 2012, Gill et al., 2008, Kim, 2000, Moulton and Steel, 2004]. This gives the complete description of the

semi-algebraic structure of the model $M(T)$. Such a complete description was known for a simple tree with only four leaves (see Pearl and Xu [1987, Theorem 2]) and for a star tree (see Bekker and de Leeuw [1987]). For a general tree, only the defining equations were known (see Sullivant [2008, Corollary 6.5]).

5.3 From tree metrics to phylogenetic oranges

In this section we briefly recall basic results for metrics defined on trees and the corresponding space of phylogenetic oranges. This space is of interest as it equivalent to the space of correlation matrices with strictly positive entries and can be used to build up a description of the complete space as is shown in Section 5.4. A special role is played by binary trees which, recall, are trees whose inner nodes all have degree three. Here we follow the definition of phylogenetic oranges as given by Kim [2000] and the concept of tree metrics as introduced by Buneman [1974] and discussed by Moulton and Steel [2004] for example.

Let $T = (U, E)$ be a tree with leaf set $V \subseteq U$. Associate to each edge a positive number d_e, which we interpret as length of this edge. Then for any two leaves $i, j \in V$ we can compute the distance between them

$$d_{ij} = \sum_{e \in \overline{ij}} d_e. \tag{5.3.1}$$

It is straightforward to check that a collection of such edge lengths for all pairs $i, j \in V$ forms a metric:

(i) $d_{ij} > 0$ trivially for $i \neq j$.

(ii) $d_{ij} = 0$ trivially for $i = j$.

(iii) It is clear that $d_{ij} = d_{ji}$ as summation is commutative.

(iv) Recall that there is a unique interior node h that separates any three leaves i, j, k. Thus $d_{ij} = d_{ih} + d_{hj}$ and $d_{ik} + d_{jk} = d_{ih} + d_{hk} + d_{jh} + d_{hk}$. By (iii) we then see that $d_{ij} < d_{ik} + d_{jk}$.

The set of all metrics that arise in this way for all T with leaves labelled by V is called the space of tree metrics. We recall the following result.

Theorem 5.3.1 (Buneman [1974]). *A collection of positive numbers d_{ij} for $i, j \in V$ forms a tree metric if and only if for all (not necessarily distinct) $i, j, k, l \in V$ we have*

$$\max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\} \geq d_{ij} + d_{kl}.$$

Equivalently, for any three sums $d_{ik} + d_{jl}$, $d_{il} + d_{jk}$, $d_{ij} + d_{kl}$ two are equal and not less than the third. Moreover, if it happens then generically T is uniquely identified.

In the above theorem by generically we mean that the statement holds outside of a measure zero set corresponding to vanishing of some edge lengths d_e. A more precise statement is also possible if we allow semi-labelled trees, see Semple and Steel [2003, Section 7]. A more careful analysis shows that this generic tree is always a binary tree, that is, trees with all inner nodes having degree three. The usual triangle inequality follows from setting i, j, k distinct and $k = l$ in Theorem 5.3.1, which implies that every tree metric is a metric on V.

We say that $A|B$ is a split of the set $A \cup B \subseteq V$ if $A \cap B = \emptyset$. A split $A|B$ is compatible with T if there exists an edge of T such that removing this edge leaves A and B in two disjoint components.

Corollary 5.3.2. *The space of all tree metrics on a fixed tree T is a metric satisfying: for any four distinct leaves i, j, k, l such that $ij|kl$ is compatible with T we have*

$$d_{ik} + d_{jl} \quad = \quad d_{il} + d_{jk} \quad \geq \quad d_{ij} + d_{kl}.$$

An important closely related space defined over a tree is the space of phylogenetic oranges (see e.g. Kim [2000], Moulton and Steel [2004]). For a given T, the set of all points in $\mathbb{R}^{m(m-1)/2}$ in this space is denoted $\mathrm{PO}(T)$ and given by:

$$\rho_{ij} \quad = \quad \prod_{e \in \overline{ij}} \rho_e \qquad \text{for all } i, j \in V, \rho_{ij} \geq 0. \tag{5.3.2}$$

Note that this is a similar parametrisation to $\mathrm{M}(T)$ the GLTM (5.2.1) but the edge correlations ρ_e are non-negative. We always assume that T is a binary tree with the set of leaves V. The union of all $\mathrm{PO}(T)$ for all such binary trees is denoted by $\mathrm{PO}(V)$. Furthermore, by $\mathrm{PO}_+(T)$

and $\mathrm{PO}_+(V)$ we denote the subset of $\mathrm{PO}(T)$ and $\mathrm{PO}(V)$ respectively, where all coordinates are assumed to be strictly positive, which implies in particular that the corresponding edge correlations ρ_e must be strictly positive.

It is clear that $\mathrm{PO}(T) \subset M(T)$. The link between the model $M(T)$ and the space of tree metrics comes from the following result.

Lemma 5.3.3. *The space of tree metrics on a fixed tree T is isomorphic to $\mathrm{PO}_+(T)$. The isomorphism is given by $d_{ij} = -\log(\rho_{ij})$.*

Proof. First note that in $\mathrm{PO}_+(T)$ all edge correlations ρ_e in the parametrization (5.3.2) must be strictly positive. Taking $-\log(\cdot)$, (5.3.2) yields

$$-\log(\rho_{ij}) = \sum_{e \in \overline{ij}} -\log(\rho_e).$$

Because $\rho_e \in (0, 1]$ we have that $-\log(\rho_e) > 0$. Changing $d_{ij} = -\log(\rho_{ij})$ and $d_e = -\log(\rho_e)$ and comparing with (5.3.1) gives the result. □

This lemma lets us obtain the semi-algebraic description of $\mathrm{PO}(V)$ and $\mathrm{PO}(T)$ for any fixed tree.

Theorem 5.3.4. *Let $\Sigma = [\rho_{ij}]_{i,j \in V}$ and suppose that $\rho_{ij} \geq 0$ for all $i, j \in V$. The following two statements hold.*

(1) $\Sigma \in \mathrm{PO}(V)$ if and only if for every four (not necessarily distinct) elements i, j, k, l in V at least two out of three products

$$\rho_{ik}\rho_{jl} \qquad \rho_{il}\rho_{jk} \qquad \rho_{ij}\rho_{kl}$$

are equal and less or equal to the third. Moreover, if this holds then T such that $\Sigma \in \mathrm{PO}(T)$ is generically identified uniquely.

(2) If T is a fixed tree then the space $\mathrm{PO}(T)$ has dimension $|E|$ and is described by the following set of constraints. For any four distinct elements i, j, k, l of V such that the split $ij|kl$ is

compatible with T we have

$$\rho_{ik}\rho_{jl} \quad = \quad \rho_{il}\rho_{jk} \quad \leq \quad \rho_{ij}\rho_{kl}. \tag{5.3.3}$$

Moreover, for any three distinct leaves i, j, k

$$\rho_{ij}\rho_{ik} \quad \leq \quad \rho_{jk}. \tag{5.3.4}$$

Before proceeding with the proof, we introduce the definition of a toric cube, a concept that we use in the proof of Theorem 5.3.4.

Definition 5.3.5 (Engström et al. [2012])**.** *A toric precube in n dimensions is a subset C of the standard cube $[0,1]^n$ that is defined by binomial inequalities*

$$x_1^{r_1} \cdot \ldots \cdot x_n^{r_n} \leq x_1^{w_1} \cdot \ldots \cdot x_n^{w_n}, \quad \textit{where} \quad r_i, w_j \in \mathbb{N} \cup \{0\}, x \in [0,1].$$

A toric cube is a toric precube that also satisfies $C = C \cap (0,1]^n$.

Proof. Assume first that all correlations ρ_{ij} are strictly positive, that is $\Sigma \in \mathrm{PO}_+(V)$ or $\Sigma \in \mathrm{PO}_+(T)$. In this case Lemma 5.3.3 gives an isomorphism with the space of tree metrics, whose constraints are given in Theorem 5.3.1 and Corollary 5.3.2. Translating these constraints via $d_{ij} = -\log(\rho_{ij})$ gives exactly the constraints in the proposed theorem. These constraints describe a closed set, which is the smallest closed set containing $\mathrm{PO}_+(V)$ so it is enough to show that the closure (in Euclidean space) of $\mathrm{PO}_+(T)$ is equal to $\mathrm{PO}(T)$. First note that $\mathrm{PO}(T)$ is a toric cube, that is, it is given as the image of a hypercube $[0,1]^{|E|}$ under a monomial map. By Engström et al. [2012, Theorem 1] every toric cube is equal to the closure of its interior, which completes our argument. □

The relationship between $M(T)$ and $\mathrm{PO}(T)$ may be further refined, which will give us an explicit description of $M(T)$. Define a multiplicative group $G = \{-1, 1\}^{|V|}$ that acts on the sample space $\mathbb{R}^{|V|}$ by reflections across axes. We identify this group with the group of diagonal $|V| \times |V|$-matrices with ± 1 on the diagonal. With this identification for every $\epsilon = (\mathrm{diag}(\epsilon_v)) \in G$ the action on an element $x \in \mathbb{R}^{|V|}$ is just given by regular matrix multiplication, $x \mapsto \epsilon \cdot x =$

$(\epsilon_v x_v)$. The action of G on the sample space induces the action on the model space

$$\Sigma \mapsto \epsilon \cdot \Sigma = \Sigma' = [\rho'_{ij}], \qquad \text{where } \rho'_{ij} = \epsilon_i \epsilon_j \rho_{ij}.$$

This also induces the action on the parameter space including the edge correlations $\rho = (\rho_e)$ so that if $\epsilon_v = -1$ then the sign of the edge correlation for the unique edge of v changes. It is an easy check that $[\epsilon \cdot \rho] = \epsilon \cdot \Sigma$. We therefore have the following result.

Proposition 5.3.6. *For every T the model $M(T)$ is invariant under G, that is $G \cdot M(T) = M(T)$.*

Denote by $\mathcal{S}_+(V)$ the space of all symmetric positive definite $|V| \times |V|$-matrices.

Theorem 5.3.7. *Let T be a tree and let $\Sigma = [\rho_{ij}] \in \mathcal{S}_+(V)$ be a correlation matrix. We have $\Sigma \in M(T)$ if and only if*

$$\Sigma' := [|\rho_{ij}|] \in \mathrm{PO}(T) \text{ and } \rho_{ij}\rho_{ik}\rho_{jk} \geq 0 \text{ for any three distinct } i, j, k \in V.$$

Proof. For sufficiency, if $\Sigma \in M(T)$ then each ρ_{ij} has representation (5.2.1). Thus $|\rho_{ij}| = \prod_{e \in \overline{ij}} |\rho_e|$ and hence Σ' also lies in $\mathrm{PO}(T)$. To show that $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ note that the tree spanned on three leaves i, j, k necessarily has a unique vertex v that lies on the intersection of paths $\overline{ij}$, $\overline{ik}$ and $\overline{jk}$. Moreover, by (5.2.1),

$$\rho_{ij}\rho_{ik}\rho_{jk} = \prod_{e \in \overline{ij}} \rho_e \prod_{e \in \overline{ik}} \rho_e \prod_{e \in \overline{jk}} \rho_e = \prod_{e \in \overline{iv}} \rho_e^2 \prod_{e \in \overline{jv}} \rho_e^2 \prod_{e \in \overline{kv}} \rho_e^2 \geq 0.$$

For necessity, we use the action of G. Distinguish one node in V and label it by $\{1\}$. Let $\epsilon \in G$ be such that for $i \in V$, $\epsilon_i = -1$ if $\rho_{1i} < 0$ and $\epsilon_i = 1$ otherwise. Then $\Sigma = \epsilon \cdot \Sigma'$ because: $\epsilon_1 \epsilon_i |\rho_{1i}| = \rho_{1i}$ for all $i \in V \setminus \{1\}$ and $\epsilon_i \epsilon_j |\rho_{ij}| = \rho_{ij}$ for $i, j \in V \setminus \{1\}$. This last equality follows from our assumption that $\rho_{1i}\rho_{1j}\rho_{ij} \geq 0$ and thus the sign of $\rho_{1i}\rho_{1j}$ is equal to the sign of ρ_{ij}. Now, since $\Sigma' \in \mathrm{PO}(T) \subset M(T)$ and $\Sigma = \epsilon \cdot \Sigma'$, Proposition 5.3.6 implies that $\Sigma \in M(T)$. □

Corollary 5.3.8. *We have $G \cdot \mathrm{PO}(T) = M(T)$ or more precisely the model $M(T)$ is isomorphic to the space of orbits of the action of G on the space of phylogenetic oranges $\mathrm{PO}(T)$.*

The link between the space of phylogenetic oranges $\mathrm{PO}(T)$ and the space of tree metrics $M(T)$ will help us to study the space of GLTMs.

5.4 The complete set of Gaussian tree constraints

Now that the link from the space of phylogenetic oranges to the space of GLTMs has been established, we can derive the complete set of associated tree constraints. We begin by considering the smallest binary tree, the tripod tree as shown in Figure 5.1.

Example 5.4.1. Let T be the tripod tree. By Theorem 5.3.7 and Theorem 5.3.4, the space of correlation matrices in $M(T)$ is described by

$$\rho_{12}\rho_{13}\rho_{23} \geq 0, \quad |\rho_{12}\rho_{13}| \leq |\rho_{23}|, \quad |\rho_{12}\rho_{23}| \leq |\rho_{13}|, \quad |\rho_{13}\rho_{23}| \leq |\rho_{12}|.$$

If $\rho_{12}, \rho_{13}, \rho_{23} \geq 0$ then by Theorem 5.3.4 (2) the space described by the above inequalities corresponds to $\mathrm{PO}(T)$. There are three other sign patterns for $\rho_{12}, \rho_{13}, \rho_{23}$ that assure that $\rho_{12}\rho_{13}\rho_{23} \geq 0$. For every such pattern we obtain a copy of $\mathrm{PO}(T)$ and hence $M(T)$ is given by four copies of $\mathrm{PO}(T)$ as depicted in Figure 5.4.

The description of this model contains no implicit equations. Basic calculus shows that this model fills $\frac{2}{\pi^2} \approx 0.2$ part of the volume of the space of all 3×3 correlation matrices. This shows the importance of including inequality constraints in our description.

We now provide the complete set of constraints to give an explicit description of $M(T)$ via Theorem 5.3.4 and the action of group G.

Proposition 5.4.2. *If T is a fixed tree then the space $M(T)$ has dimension $|E|$ (where recall E is the number of edges) and is described by the following set of constraints. For any four distinct elements i, j, k, l of V such that the split $ij|kl$ is compatible with T we have*

$$\frac{\rho_{ik}\rho_{jl}}{\rho_{ij}\rho_{kl}} = \frac{\rho_{il}\rho_{jk}}{\rho_{ij}\rho_{kl}} \leq 1. \tag{5.4.1}$$

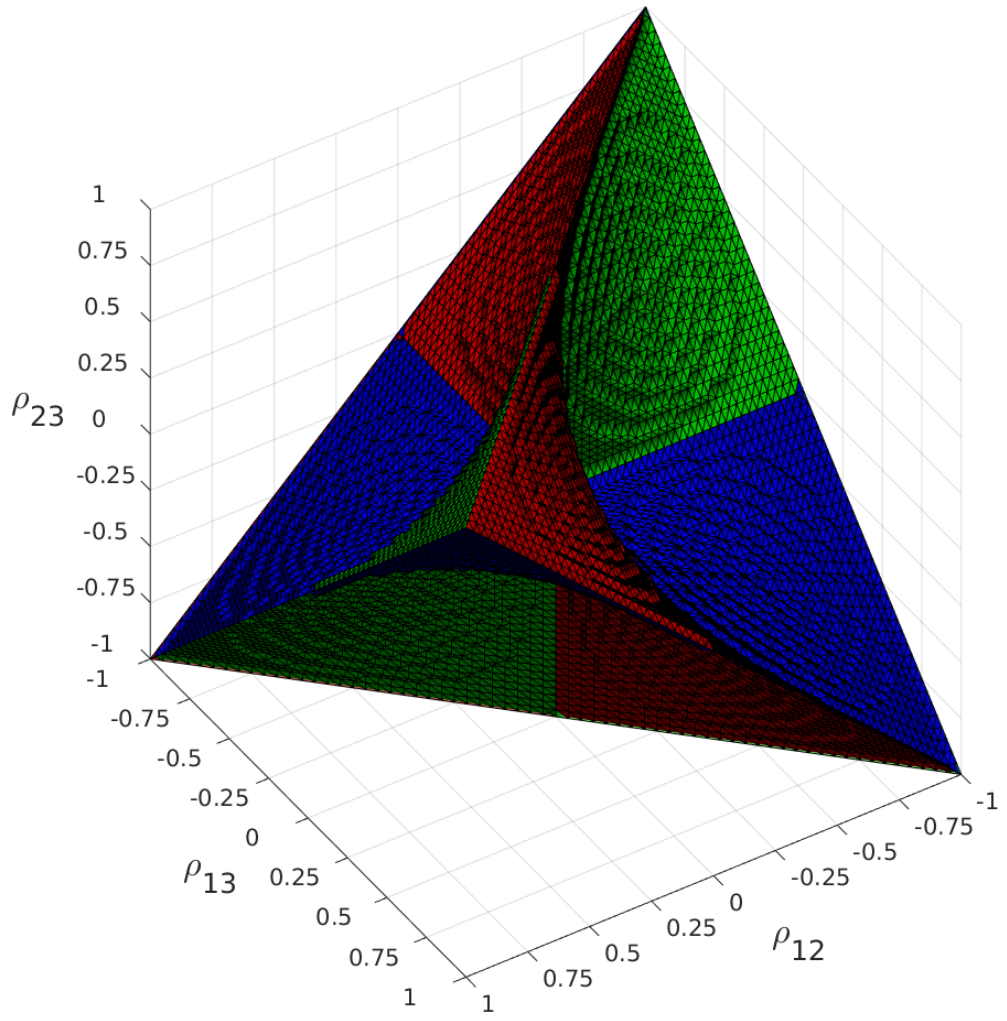


FIGURE 5.4: Region in correlation space consistent with the tripod tree model.

Moreover, for any three distinct leaves i, j, k

$$(\rho_{ij} - \rho_{ik}\rho_{jk})(\rho_{ik} - \rho_{ij}\rho_{jk})(\rho_{jk} - \rho_{ij}\rho_{ik}) \geq 0. \tag{5.4.2}$$

Remark 5.4.3. We refer to (5.4.2) as the tripod constraints and to (5.4.1) as the tetrad constraints. Furthermore, simple bounding shows that (5.4.2) implies the positivity constraint (5.1.1) $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$:

$$\begin{aligned}
0 &\leq (\rho_{ij} - \rho_{ik}\rho_{jk})(\rho_{ik} - \rho_{ij}\rho_{jk})(\rho_{jk} - \rho_{ij}\rho_{ik}) \\
&= \rho_{ij}\rho_{ik}\rho_{jk} + \rho_{ij}^3\rho_{ik}\rho_{jk} + \rho_{ij}\rho_{ik}^3\rho_{jk} + \rho_{ij}\rho_{ik}\rho_{jk}^3 - \rho_{ij}^2\rho_{ik}^2 - \rho_{ij}^2\rho_{jk}^2 - \rho_{ik}^2\rho_{jk}^2 - \rho_{ij}^2\rho_{ik}^2\rho_{jk}^2 \\
&\leq \rho_{ij}\rho_{ik}\rho_{jk} + \rho_{ij}^3\rho_{ik}\rho_{jk} + \rho_{ij}\rho_{ik}^3\rho_{jk} + \rho_{ij}\rho_{ik}\rho_{jk}^3 \\
&= (\rho_{ij}\rho_{ik}\rho_{jk})(1 + \rho_{ij}^2 + \rho_{ik}^2 + \rho_{jk}^2)
\end{aligned}$$

But given that $1 + \rho_{ij}^2 + \rho_{ik}^2 + \rho_{jk}^2 \geq 0$ means $\rho_{ij}\rho_{ik}\rho_{jk} \geq 0$ as required.

Proof of Proposition 5.4.2. First note that, by Theorem 5.3.4 (2), $\mathrm{PO}(T)$ satisfies all the above constraints. Moreover, all these constraints are invariant with respect to the action of G and hence $G \cdot \mathrm{PO}(T)$ also satisfies these constraints. By Corollary 5.3.8, $G \cdot \mathrm{PO}(T)$ is equal to $M(T)$. Hence it is now enough to show that every point Σ satisfying the above constraints also lies in $M(T)$. Equivalently we can show that there always exists $\epsilon \in G$ such that $\Sigma' := \epsilon \cdot \Sigma$ lies in $\mathrm{PO}(T)$. As in the proof of Theorem 5.3.7 take $\epsilon_i = \mathrm{sgn}(\rho_{1i})$. Then $\rho'_{ij} = \epsilon_i\epsilon_j\rho_{ij} = |\rho_{ij}| \geq 0$. Since constraints in (5.4.1) and (5.4.2) are G-invariant, they also hold for Σ'. In particular Σ' satisfies (5.3.3). Moreover,

$$(\rho'_{ij} - \rho'_{ik}\rho'_{jk})(\rho'_{ik} - \rho'_{ij}\rho'_{jk})(\rho'_{jk} - \rho'_{ij}\rho'_{ik}) \geq 0,$$

which means that for each such a triple either all terms are nonnegative or exactly two are negative. If all are nonnegative then $\Sigma' \in \mathrm{PO}(T)$ by Theorem 5.3.4(2). Suppose that the first two are negative. But $\rho'_{ij} < \rho'_{ik}\rho'_{jk}$ and $\rho'_{ik} < \rho'_{ij}\rho'_{jk}$ implies that

$$\rho'_{ij}\rho'_{ik} < \rho'_{ij}\rho'_{ik}\rho'^2_{jk},$$

which is of course impossible because $\rho'_{jk} \in [0, 1]$. □

This result can be viewed as a generalisation of the main results in Bekker and de Leeuw [1987], Pearl and Xu [1987] from star trees to general trees. Equations defining $M(T)$ were already given in Sullivant [2008, Corollary 6.5]. We have now derived the complete set of Gaussian tree constraints in a similar fashion to the binary tree constraints in Chapter 4. In order to be able to utilise these constraints with the Romance language data set it is necessary to develop some methodology, which is the focus of Chapter 6.

Chapter 6

A probabilistic approach to Gaussian tree constraints

This chapter presents a selection of methods that together make use of the full set of Gaussian latent tree constraints. This suite of tools not only incorporates the algebraic constraints, but also moves beyond simplistic binary outcomes of tree-compatibility to introduce more nuanced probabilistic approaches. In the case of inequality constraints we can elicit posterior probabilities that a given data set is consistent with the semi-algebraic tree space. For the algebraic constraints we make the link between the equalities and the tetrad analyses described in Bollen and Ting [1993]. Using the work of Drton et al. [2008] the relevant moments for the equality constraints are derived for use with exploratory and confirmatory tetrad analyses.

Once we have developed the tools in this chapter, we can cover two main scenarios. Firstly, given a data set, we can assess the suitability of the GLTM class as a whole. For example, in Section 7.1.5 we will go on to test the suitability the space of quintet trees for a linguistic data set. Secondly, if presented with a candidate tree for a particular data set (as the output of a search algorithm say), then we are able to examine in absolute terms whether this is a good fit for the data. We subsequently demonstrate such a scenario in Section 7.2.1 where we test yeast species against a phylogenetic tree listed in the literature. Knowing whether a tree is the correct model is pertinent in phylogenetic settings where the effect of horizontal gene transfer can vary

greatly (e.g. Hao and Golding [2008]) and consequently the relevant search space is not clear. This chapter covers some of the latter half of the paper Shiers et al. [2014].

6.1 Utilising semi-algebraic tree constraints

In this section it is important to distinguish between covariances and correlations explicitly and so we denote covariances as Σ_C and the corresponding correlations as Σ_R, as we are sometimes using distributions that require the covariance representation instead of the correlation. The true correlation matrix Σ_R is not usually known in practice so it is not usually possible to directly test for tree-compatibility using the inequalities and equalities in Proposition 5.4.2. Instead we must work with the sample correlations $\hat{\Sigma}_R$ to test for violations in the constraints. Of course, any sample analogues of the parametric constraints we have derived above will only hold approximately and in particular, using point estimates of correlations with the semi-algebraic constraints will only give a crude binary assessment of each inequality. In this section we therefore describe how the semi-algebraic constraints can be used more formally to give an indication of tree-compatibility. The algebraic constraints are dealt with in Sections 6.2 and 6.3.

Before proceeding, it is worth noting that if the assumption is made that Σ_R will only have positive entries then the constraints that describe $\mathrm{PO}_+(T)$ can be used as they are but in combination with the trivial constraint $\rho_{ij} \geq 0$ for all $i, j \in V$. However, in some circumstances it is preferable to allow for the possibility of negative correlations. For example, in some evolutionary processes some species traits may diverge from a common ancestor. This could be caused by zero sum conditions on limited resources leading to divergence of growth curves for example and so negative correlation. Alternatively the construction of the analysis may induce negative correlations, e.g. through standardisation of a data set.

A straightforward but effective assessment of T-compatibility constraints can be obtained from the posterior probabilities that each of the tree inequalities is satisfied by applying an inverse-Wishart prior on the sample covariance. More precisely, if $\hat{\Sigma}_C$ is a sample matrix based on a sample X comprising n samples from $N_m(0, \Sigma_C)$, then the estimated scatter matrix is calculated as $S = n\hat{\Sigma}_C = XX^T$ and it is well known that the scatter matrix is Wishart distributed $S \sim$

$\mathcal{W}_m(n, \Sigma_C)$ [Wishart, 1928a] with probability density function

$$p(S) = \frac{\det(S)^{\frac{n-m-1}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\Sigma_C^{-1}S))}{2^{\frac{nm}{2}} \det(\Sigma_C)^{\frac{n}{2}} \Gamma_m(\frac{n}{2})}$$

where $\Gamma_m(\cdot)$ is the multivariate gamma function (see Section 2.2). As previously discussed, a common prior distribution for unknown covariance Σ_C is the inverse-Wishart $\mathcal{W}_m^{-1}(n_0, \Sigma_0)$, resulting in the posterior density $p(\Sigma_C|X)$ which is inverse-Wishart distributed with parameters as indicated: $\mathcal{W}_m^{-1}(n_0 + n, \Sigma_0 + S)$. By letting $n_0 = m$, $\Sigma_C|X$ can be sampled with each draw being translated to a correlation and then tested against the constraints. Then after N draws from the posterior distribution an estimate of the posterior probability that Σ_C satisfies the positivity constraint can be obtained.

Considering Example 5.4.1, an estimate of the probability of Σ_C satisfying the semi-algebraic structure of $M(T)$ can be constructed using indicator functions. For each draw l from the relevant inverse-Wishart posterior distribution for $\hat{\Sigma}_C$, the following identity is evaluated:

$$r_{123}^l(\hat{\Sigma}_C) = \mathbb{1}_{\{(\tilde{\rho}_{12}-\tilde{\rho}_{13}\tilde{\rho}_{23})(\tilde{\rho}_{13}-\tilde{\rho}_{12}\tilde{\rho}_{23})(\tilde{\rho}_{23}-\tilde{\rho}_{12}\tilde{\rho}_{13})\geq 0\}} \tag{6.1.1}$$

where $\tilde{\rho}_{ij}, i, j = 1, 2, 3$ are the correlations corresponding to covariance draw l of the posterior, the index l being dropped to keep the notation clean. The posterior probability of tree-compatibility is thus estimated using:

$$R_{123}(\hat{\Sigma}_C) := \frac{1}{N}\sum_{l=1}^{N} r_{123}^l(\hat{\Sigma}_C) \tag{6.1.2}$$

For a tree with four variables such that $\overline{12}$ and $\overline{34}$ do not intersect, the final test of inequality constraints is:

$$R_{12|34}(\hat{\Sigma}_C) = \frac{1}{N}\sum_{l=1}^{N} \mathbb{1}_{\{\frac{\tilde{\rho}_{14}\tilde{\rho}_{23}}{\tilde{\rho}_{12}\tilde{\rho}_{34}}\leq 1\}} \mathbb{1}_{\{\frac{\tilde{\rho}_{13}\tilde{\rho}_{24}}{\tilde{\rho}_{12}\tilde{\rho}_{34}}\leq 1\}} \prod_{\substack{1\leq i<j\\ <k\leq 4}} r_{ijk}^l(\hat{\Sigma}_C) \tag{6.1.3}$$

If only considering $\mathrm{PO}_+(T)$ then for both (6.1.1) and (6.1.3), a further condition is required that all three $\tilde{\rho}_{ij} > 0$ for each draw.

Proposition 6.1.1. *If $\Sigma \in M(T) \setminus PO(T)$ then the inequality sign in (5.1.1) for the tripod tree is reversed for two of the three constraints.*

Proof. Given $\Sigma \in M(T) \setminus PO(T)$ we know that at exactly two of $\rho_{12}, \rho_{13}, \rho_{23} \leq 0$. Without loss of generality, assume $\rho_{12}, \rho_{13} \leq 0$ and $\rho_{23} \geq 0$. Considering the tripod tree as the simplest representation of these correlations, we recall that there is a single interior node h that lies on the path between all three leaf nodes $1, 2, 3$. We can deduce that $\rho_{1h} \leq 0$ and $\rho_{2h}, \rho_{3h} \geq 0$. If we replace ρ_{1h} with $\epsilon\rho_{1h}$ where $\epsilon = -1$ then $\epsilon\rho_{12}, \epsilon\rho_{13}, \rho_{23} \geq 0$ and the three tripod constraints hold as presented in (5.1.1):

$$\epsilon\rho_{12} \geq \epsilon\rho_{13}\rho_{23} \quad \epsilon\rho_{13} \geq \epsilon\rho_{12}\rho_{23} \quad \rho_{23} \geq \epsilon\rho_{12}\epsilon\rho_{13}.$$

Thus removing ϵ, the three constraints become:

$$\rho_{12} \leq \rho_{13}\rho_{23} \quad \rho_{13} \leq \rho_{12}\rho_{23} \quad \rho_{23} \geq \rho_{12}\rho_{13}.$$

□

Corollary 6.1.2. *The estimator* (6.1.1) *accounts for the reversal of the inequalities as described in Proposition 6.1.1.*

Proof. Without loss of generality, assume $\rho_{12}, \rho_{13} \leq 0$ and $\rho_{23} \geq 0$. Thus substituting the relevant constraints into the indicator of (6.1.1) gives $(\rho_{13}\rho_{23} - \rho_{12})(\rho_{12}\rho_{23} - \rho_{13})(\rho_{23} - \rho_{12}\rho_{13}) = (\rho_{12} - \rho_{13}\rho_{23})(\rho_{13} - \rho_{12}\rho_{23})(\rho_{23} - \rho_{12}\rho_{13})$ trivially. □

Proposition 6.1.3. *If $(\rho_{12} - \rho_{13}\rho_{23})(\rho_{13} - \rho_{12}\rho_{23})(\rho_{23} - \rho_{12}\rho_{13}) \geq 0$ then*

(i) if $\rho_{12}, \rho_{13}, \rho_{23} \in (0, 1)$ then $\rho_{12} - \rho_{13}\rho_{23} \geq 0, \rho_{13} - \rho_{12}\rho_{23} \geq 0, \rho_{23} - \rho_{12}\rho_{13} \geq 0$

(ii) if $\rho_{12} \in (0, 1)$ and $\rho_{13}, \rho_{23} \in (-1, 0)$ then $\rho_{12} - \rho_{13}\rho_{23} \geq 0, \rho_{13} - \rho_{12}\rho_{23} \leq 0, \rho_{23} - \rho_{12}\rho_{13} \leq 0$.

We exclude the cases of measure 0 where $\rho_{12}, \rho_{13}, \rho_{23} \in \{-1, 0, 1\}$. These correspond to boundaries or joints of the tree-compatible regions and can produce exceptions to Proposition 6.1.3. These do not occur in practice during simulation and theoretically boundary cases will never be sampled.

Proof. We will prove this by assuming that two violations occur in the relevant semi-algebraic constraints and then proceed to reach a contradiction. Thus, we will have demonstrated that if $(\rho_{12} - \rho_{13}\rho_{23})(\rho_{13} - \rho_{12}\rho_{23})(\rho_{23} - \rho_{12}\rho_{13}) \geq 0$ then all the constituent products are also non-negative.

(i) Suppose that $\rho_{12}, \rho_{13}, \rho_{23} \in (0, 1)$ and without loss of generality suppose that $\rho_{12} \leq \rho_{13}\rho_{23}$ (a violation). Now suppose that also $\rho_{13} \leq \rho_{12}\rho_{23}$. Then

$$\rho_{13}\rho_{23} \leq \rho_{12}\rho_{23}^2\rho_{12} \leq \rho_{13}\rho_{23} \leq \rho_{12}\rho_{23}^2 \implies \rho_{12} \leq \rho_{12}\rho_{23}^2.$$

But given $\rho_{12}, \rho_{23} \in (0, 1)$ we have a contradiction.

(ii) Suppose that $\rho_{12} \in (0, 1)$ and $\rho_{13}, \rho_{23} \in (-1, 0)$. Now we consider two cases:

Case 1: Suppose that $\rho_{12} \leq \rho_{13}\rho_{23}$ and suppose that also $\rho_{13} \geq \rho_{12}\rho_{23}$. Then

$$\rho_{13} \geq \rho_{12}\rho_{23} \geq \rho_{13}\rho_{23}^2 \implies \rho_{13} \geq \rho_{13}\rho_{23}^2 \implies 1 \leq \rho_{23}^2 \text{ as } \rho_{13} \in (-1, 0).$$

But $\rho_{23}^2 \in (0, 1)$ so we have a contradiction.

Case 2: Suppose that $\rho_{13} \geq \rho_{12}\rho_{23}$ and suppose that also $\rho_{23} \geq \rho_{12}\rho_{13}$. Similarly to before,

$$\rho_{12}\rho_{23} \geq \rho_{12}^2\rho_{13} \implies \rho_{13} \geq \rho_{12}\rho_{23} \geq \rho_{12}^2\rho_{13} \implies 1 \leq \rho_{12}^2.$$

But $\rho_{12}^2 \in (0, 1)$ so we have a contradiction.

This covers all possibilities (up to permutation) since we know that either 0 or 2 correlations are negative for tree-compatibility to hold due to the positivity constraint (5.1.1). □

Remark 6.1.4. Proposition 6.1.1, Corollary 6.1.2 and Proposition 6.1.3 together provide an alternative proof for the latter part of the proof of Proposition 5.4.2.

Note that the tetrad inequality constraints in 5.4.1 can be instead written such that the sign determines whether there is a violation. i.e.

$$\frac{\rho_{13}\rho_{24}}{\rho_{12}\rho_{34}} = \frac{\rho_{14}\rho_{23}}{\rho_{12}\rho_{34}} \leq 1 \quad \Rightarrow \quad \rho_{12}\rho_{34} - \rho_{13}\rho_{24} \geq 0, \quad \rho_{12}\rho_{34} - \rho_{14}\rho_{23} \geq 0. \tag{6.1.4}$$

With this latter representation, the estimator equivalent to 6.1.3 is given by:

$$R_{12|34}(\hat{\Sigma}) := \frac{1}{N}\sum_{l=1}^{N}(\alpha^{l}_{12|34}\prod_{\substack{1\leq i<j\\ <k\leq 4}} r^{l}_{ijk}(\hat{\Sigma})) \tag{6.1.5}$$

where

$$\alpha^{l}_{12|34} := \mathbb{1}_{\{F^{l}(\hat{\Sigma})(\tilde{\rho}_{12}\tilde{\rho}_{34}-\tilde{\rho}_{13}\tilde{\rho}_{24})\geq 0\}}\mathbb{1}_{\{F^{l}(\hat{\Sigma})(\tilde{\rho}_{12}\tilde{\rho}_{34}-\tilde{\rho}_{14}\tilde{\rho}_{23})\geq 0\}} \tag{6.1.6}$$

and $F^{l}(\hat{\Sigma}) = \operatorname{sgn}(\tilde{\rho}_{12}\tilde{\rho}_{13}\tilde{\rho}_{14})$ is the signum function. The role of $F(\cdot)$ is to adapt the inequalities to account for tree-compatible regions which do not lie in purely positive correlation space. The description of $\mathrm{PO}_{+}(T)$ allows us to easily extend to $M(T)$ as it is obtained as the union of rotations of $\mathrm{PO}_{+}(T)$. This corresponds to the tetrad inequalities (5.3.3) reversing in sign if $F^{l}(\hat{\Sigma})$ is negative.

Proposition 6.1.5. *The function* $\operatorname{sgn}(\tilde{\rho}_{12}\tilde{\rho}_{13}\tilde{\rho}_{14})$ *found in (6.1.6) reverses the tetrad inequalities when necessary.*

Proof. Consider the quartet tree as show in Figure 6.1. We can write the tetrad inequalities in (5.4.1) in terms of edge correlations:

$$\rho_{14}\rho_{23} = \rho_{13}\rho_{24} = \rho_{1g}\rho_{2g}\rho_{3h}\rho_{4h}\rho^{2}_{gh} \leq \rho_{1g}\rho_{2g}\rho_{3h}\rho_{4h} = \rho_{12}\rho_{34}$$

Observe that the inequality above only reverses if an odd number of edge correlations (excluding ρ_{gh}) are negative. Also note that $\operatorname{sgn}(\rho_{1g}\rho_{2g}\rho_{3h}\rho_{4h}\rho^{2}_{gh}) = \operatorname{sgn}(\rho_{1g}\rho_{2g}\rho_{3h}\rho_{4h})$. Finally, observe that $\operatorname{sgn}(\rho_{12}\rho_{13}\rho_{14}) = \operatorname{sgn}(\rho_{1g}\rho_{2g}\rho_{3h}\rho_{4h})$ which completes the proof. □

Corollary 6.1.6. *The tetrad inequalities need only be reversed if 3 of the correlations* ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, ρ_{34} *are negative and specifically those correlations are* $\rho_{ij}, \rho_{ik}, \rho_{il}$ *for* i, j, k, l *distinct.*

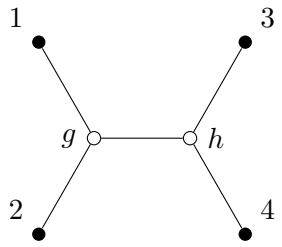


FIGURE 6.1: Quartet tree

Proof. This follows from Proposition 6.1.5 where if the $\operatorname{sgn}(\rho_{12}\rho_{13}\rho_{14}) = -1$ then the inequality is reversed. Thus either a single ρ_{1l} is negative in which case also $\rho_{jl}, \rho_{kl} < 0$, or otherwise $\rho_{12}, \rho_{13}, \rho_{14} < 0$. To complete the proof consider exactly three negative correlations that do not fit the stated pattern, without loss of generality $\rho_{12}, \rho_{13}, \rho_{24}$ or $\rho_{12}, \rho_{13}, \rho_{23}$. In either case the positivity constraint fails e.g. $\rho_{23}\rho_{24}\rho_{34} < 0$ and $\rho_{12}\rho_{13}\rho_{23} < 0$ respectively. This is accounted for in (6.1.3) and so the sign of the tetrad inequalities is irrelevant for the simulation. □

In general, the former estimator (6.1.3) shall be used as it is cleaner and simpler to implement. The alternative estimator (6.1.5) can be used if there is a concern about numerical accuracy caused by small values of ρ_{ij} in the denominator — though such cases are rare.

6.2 The sample distribution of algebraic constraints

The sampling approach described in Section 6.1 does not extend to the algebraic constraints as the set of draws from the posterior satisfying an equality constraint will have zero probability. Thus an alternative approach is taken using sample distributions of minors of a covariance matrix.

6.2.1 Means and variances

Considering Theorem 5.3.4, it is clear that information on whether a Gaussian distribution lies in some $M(T)$ for these particular constraints is recorded in the sign of tetrad constraints $\sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk}$ and other quadratic binomials of the form $\sigma_{ii}\sigma_{jk} - \sigma_{ij}\sigma_{ik}$. Here we use the covariance matrix instead of correlation matrix as the distributional results associated with correlations are notably more difficult to deal with. Both type of constraints can be realised as minors of the

covariance matrix Σ

$$\det(\Sigma_{ij,kl}) \qquad \text{and} \qquad \det(\Sigma_{ij,ik}). \tag{6.2.1}$$

However, we are particularly interested in the former as the algebraic constraints cannot be treated in the same way as the semi-algebraic constraints in Section 6.1.

Definition 6.2.1. *A matrix A vanishes if the* $\det(A) = 0$.

Let $A|B$ be a split of the set of leaves that is induced by T by removing an edge and considering two resulting components. Consider the following block matrix

$$\begin{bmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{bmatrix}.$$

Proposition 6.2.2. *The equality constraints in (5.3.3) imply that all 2×2 minors of $\Sigma_{A,B}$ must vanish if $A \cup B = V$ for $T = (V, E) \subset M(T)$.*

Proof. Without loss of generality, consider $i, j \in A$ and $k, l \in B$, i, j, k, l distinct. Then $\{i, j\}|\{k, l\}$ and so $\sigma_{ik}\sigma_{jl} = \sigma_{il}\sigma_{jk}$ as a consequence of Theorem (5.3.4). Thus, $\det(\Sigma_{ij,kl}) = \sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk} = 0$, i.e. $\Sigma_{ij,kl}$ vanishes. □

Corollary 6.2.3. *The rank of all 2×2 minors of $\Sigma_{A,B}$ must equal 1 if $A \cup B = V$ for $T = (V, E) \subset M(T)$.*

Proof. Without loss of generality, consider $i, j \in A$ and $k, l \in B$, i, j, k, l distinct. Then

$$\Sigma_{ij,kl} = \begin{bmatrix} \sigma_{ik} & \sigma_{il} \\ \sigma_{jk} & \sigma_{jl} \end{bmatrix}.$$

Multiplying the second row of $\Sigma_{ij,kl}$ by $\frac{\sigma_{ik}}{\sigma_{jk}}$ produces:

$$\begin{bmatrix} \sigma_{ik} & \sigma_{il} \\ \sigma_{ik} & \frac{\sigma_{ik}\sigma_{jl}}{\sigma_{jk}} \end{bmatrix}.$$

Since $\sigma_{ik}\sigma_{jl} - \sigma_{il}\sigma_{jk} = 0 \implies \frac{\sigma_{ik}\sigma_{jl}}{\sigma_{jk}} = \sigma_{il}$, rows 1 and 2 of $\Sigma_{ij,kl}$ are linearly dependent. Hence, $\operatorname{rank}(\Sigma_{ij,kl}) = 1$. □

We wish to understand the distributional properties of the algebraic tetrad constraints for the Gaussian setting. Suppose that a sample $X \in \mathbb{R}^{n\times m}$ was observed and let $S = X^T X$ so that S/n is the sample covariance matrix. If C is the covariance matrix of the data generating distribution then S has Wishart distribution $\mathcal{W}_m(n, C)$. Denote by $\binom{m}{2}$ the set of all subsets of $\{1, \ldots, m\}$ of cardinality two. We are going to use the following estimator of the value of $\det(C_{I,J})$ for $I, J \in \binom{m}{2}$

$$Q_{I,J} \quad := \quad \frac{1}{n(n-1)} \det(S_{I,J}). \tag{6.2.2}$$

In this section we provide the first two moments for the random $\binom{m}{2} \times \binom{m}{2}$-matrix with entries given by $Q_{I,J}$. The following result, that shows that $Q_{I,J}$ is an unbiased estimator, follows from Drton et al. [2008, Corollary 4.2].

Proposition 6.2.4. *If* $I, J \in \binom{m}{2}$ *then* $\mathbb{E}[Q_{I,J}] = \det(C_{I,J})$.

It is convenient to introduce the following notation. For an $m \times m$ matrix A let $A^{(2)}$ denote the matrix with rows and columns indexed by elements $\binom{m}{2}$ whose (I, J)-th element is the corresponding minor $\det(A_{I,J})$. The rows and columns are ordered in the natural order of $\binom{m}{2}$ given by

$$\{1, 2\} \prec \{1, 3\} \prec \cdots \prec \{1, m\} \prec \{2, 3\} \prec \cdots \prec \{m-1, m\}.$$

With this notation, the matrix, whose elements are estimators $Q_{I,J}$ is given by $S^{(2)}/(n(n-1))$. Proposition 6.2.4 now reads: $\mathbb{E}[S^{(2)}] = n(n-1)C^{(2)}$.

Remark 6.2.5. This given ordering of $\binom{m}{2}$ is induced from the natural ordering on $\{1, \ldots, m\}$ given by $1 < 2 < \cdots < m$. A different ordering of $\{1, \ldots, m\}$ will lead to a different ordering of $\binom{m}{2}$.

We next provide the variance for $Q_{I,J}$. The formulae depend on the cardinality of $I \cap J$.

Proposition 6.2.6. *If* $I, J \in \binom{m}{2}$ *are disjoint, then*

$$\operatorname{var}(Q_{I,J}) = \frac{1}{n(n-1)} \left[(n+2) \det(C_{I,I}) \det(C_{J,J}) - n \det(C_{I\cup J, I\cup J}) + 3n \det(C_{I,J})^2 \right]$$

If $I = J \in \binom{m}{2}$ *then*

$$\operatorname{var}(Q_{I,I}) = \frac{4n+2}{n(n-1)} \det(C_{I,I})^2.$$

Moreover, if $I, J \in \binom{m}{2}$ *such that* $I = \{i, j\}$, $J = \{i, k\}$ *and* $j \neq k$, *then*

$$\operatorname{var}(Q_{I,J}) = \frac{4n+2}{n(n-1)} \det(C_{I,J})^2 + \frac{n+2}{n(n-1)} C_{i,i}^2 \det(C_{jk,jk} - C_{jk,i} C_{i,i}^{-1} C_{i,jk}).$$

Remark 6.2.7. The result for $I = J \in \binom{m}{2}$ cannot (and is not intended to) be derived directly from result for the $I, J \in \binom{m}{2}$ disjoint.

Proof. If I, J are disjoint, this is a classical result of Wishart [1928b], see also Drton et al. [2008, Corollary 5.6]. The second part of the result, covering the case $|I \cap J| \geq 1$, follows from the main theorem of Drton et al. [2008] (see Drton and Goia [2012] for the corrected version). □

6.2.2 Covariances between minors

There is no simple explicit formula for covariances of various 2-minors but they can be computed if the true distribution of C is known.

By Drton et al. [2008, Proposition 3.3]

$$\operatorname{cov}[S^{(2)}] = [(C^{1/2})^{(2)} \otimes (C^{1/2})^{(2)}] \cdot (\operatorname{cov}(W^{(2)})) \cdot [(C^{1/2})^{(2)} \otimes (C^{1/2})^{(2)}], \tag{6.2.3}$$

where W has standard Wishart distribution $\mathcal{W}_m(n, I)$ and $\otimes$ is the Kronecker product. This follows from

$$S = C^{1/2} W C^{1/2} \sim \mathcal{W}_m(n, C^{1/2} I_m C^{1/2}) = \mathcal{W}_m(n, C)$$

and the Cauchy–Binet formula [Olkin and Marshall, 2014]

$$(AB)^{(k)} = (A)^{(k)} (B)^{(k)}.$$

In the rest of this section we provide a complete description of the covariance matrix $\operatorname{cov}(W^{(2)})$. This matrix has many symmetries that we want to exploit both in the exposition below and in the computations. First note that $\det W_{I,J} = \det W_{J,I}$ for all $I, J \in \binom{m}{2}$ and hence

$$\operatorname{cov}(\det W_{I,J}, \det W_{K,L}) = \operatorname{cov}(\det W_{J,I}, \det W_{K,L}) = \operatorname{cov}(\det W_{K,L}, \det W_{I,J}).$$

This enables us to take the following convention. Using the natural order of $\binom{m}{2}$ we always assume that

$$I \preceq J, \quad K \preceq L \quad \text{and} \quad I \preceq K. \tag{6.2.4}$$

If we can find $\operatorname{cov}(\det W_{I,J}, \det W_{K,L})$ for all I, J, K, L satisfying (6.2.4) we can then fill out the rest of the matrix $\operatorname{cov}(W^{(2)})$ using basic symmetries.

Let $A \Delta B := (A \setminus B) \cup (B \setminus A)$ be the symmetric difference of two sets.

Proposition 6.2.8. *Suppose that $I, J, K, L \in \binom{m}{2}$ satisfy (6.2.4) and $(I, J) \neq (K, L)$. If $I \Delta J \neq K \Delta L$, then* $\operatorname{cov}(\det W_{I,J}, \det W_{K,L}) = 0$.

Proof. Note that necessarily either $I \neq J$ or $K \neq L$ (otherwise $I \Delta J = \emptyset = K \Delta L$) and hence either $\det C_{I,J} = 0$ or $\det C_{K,L} = 0$, where C is the identity matrix. By Proposition 6.2.4

$$\mathbb{E}(\det W_{I,J})\mathbb{E}(\det W_{K,L}) = n^2(n-1)^2 \det C_{I,J} \det C_{K,L} = 0,$$

By Drton et al. [2008, Proposition 3.4] also $\mathbb{E}(\det W_{I,J} \det W_{K,L}) = 0$. □

In the rest of this section we assume that $I \Delta J = K \Delta L$. If $(I, J) = (K, L)$ then trivially $\operatorname{cov}(\det W_{I,J}, \det W_{K,L}) = \operatorname{var}(\det W_{I,J})$ and by Proposition 6.2.6

$$\frac{1}{n^2(n-1)^2} \operatorname{var}(\det W_{I,J}) = \begin{cases} \frac{4n+2}{n(n-1)} & \text{if } I = J \\ \frac{2}{n(n-1)} & \text{if } I \cap J = \emptyset \\ \frac{n+2}{n(n-1)} & \text{otherwise.} \end{cases} \tag{6.2.5}$$

For the remaining cases we are going to use the following result, which deals with general minors of size r.

Theorem 6.2.9 (Drton et al. [2008], Theorem 4.5)**.** *Suppose $I, J, K, L \in \binom{m}{r}$ such that $I \Delta J = K \Delta L$. Fix an ordering of $\{1, \ldots, m\}$ that induces an ordering on $\binom{m}{r}$ (c.f. Remark 6.2.5) that*

satisfies

$$\begin{aligned}(I\cap J)\setminus(K\cap L) \quad &\preceq \quad I\setminus J \quad \preceq \quad J\setminus I \quad \preceq \quad (K\cap L)\setminus(I\cap J)\\ (I\setminus J)\cap(K\setminus L) \quad &\preceq \quad (I\setminus J)\cap(L\setminus K),\\ (J\setminus I)\cap(K\setminus L) \quad &\preceq \quad (J\setminus I)\cap(L\setminus K).\end{aligned} \tag{6.2.6}$$

If any terms result in an order comparison with the empty set then by convention this is said to hold. If the ordering here holds and if $W\sim\mathcal{W}_m(n,I)$, then

$$\begin{aligned}\mathbb{E}[\det(W_{I,J})\det(W_{K,L})] = \frac{n!}{(n-r)!}\cdot\frac{(n+2)!}{(n+2-|I\cap J\cap K\cap L|)!}\cdot\\ \cdot\frac{(n-r+|(I\cap J)\setminus(K\cap L)|)!}{(n-r)!}\cdot|(I\setminus J)\cap(K\setminus L)|!\cdot|(I\setminus J)\cap(L\setminus K)|!\end{aligned}$$

In the rest of this section we analyse the special case of Theorem 6.2.9, when $r=2$. Our motivation is to obtain a more concrete version, with a more algorithmic approach to the ordering constraint (6.2.6). If $(I,J)\neq(K,L)$ and $I\Delta J = K\Delta L$ then we have three cases

(i) $I=J$ and $K=L$ and either $|I\cap K|=0$ or $|I\cap K|=1$

(ii) $I=\{i,j\}$, $J=\{i,k\}$, $K=\{j,l\}$ $L=\{k,l\}$ for some distinct i,j,k,l

(iii) $I=\{i,j\}$, $J=\{k,l\}$, $K=\{i,k\}$, $L=\{j,l\}$ for some distinct i,j,k,l

Proposition 6.2.10. *If $I,J,K,L\in\binom{m}{2}$ satisfy (6.2.4) and $I\Delta J = K\Delta L$ then given $I=J$, $K=L$ and $I\neq K$, we have*

$$\frac{1}{n^2(n-1)^2}\operatorname{cov}(\det W_{I,J},\det W_{K,L}) = \begin{cases}0 & \text{if } |I\cap K|=0\\ \frac{2}{n} & \text{if } |I\cap K|=1.\end{cases}$$

Proof. If $I=J$ and $K=L$ then, by Proposition 6.2.4, $\mathbb{E}(\det W_{I,J})=\mathbb{E}(\det W_{K,L})=1$. By Theorem 6.2.9 $\mathbb{E}(\det W_{I,J}\det W_{K,L})$ is equal, up to sign, to 1 if $|I\cap K|=0$ and to $\frac{n+2}{n}$ if $|I\cap K|=1$. To show that the sign is always positive we use the fact that $I\prec K$ by (6.2.4). If $|I\cap K|=0$, the ordering constraint (6.2.6) is trivially satisfied. If $|I\cap K|=1$ then it is satisfied because $I\setminus K\prec K\setminus I$. □

Obtaining remaining covariances is more subtle because they can take different signs depending on whether (6.2.6) holds or not. In our special case these technical conditions can be translated into checking parity of certain permutations.

Definition 6.2.11. *We say that a sequence* $i_1, \ldots, i_k$ defines *a partition* σ *of the set* $\{i_1, \ldots, i_k\}$ *if* $\sigma(o_j) = i_j$ *for* $j = 1, \ldots, k$*, where* $o_1, \ldots, o_k$ *is the sequence of* i_j*'s given in an increasing order. This is also sometimes called the one-line notation for permutations.*

For example if $i_1 = 2, i_2 = 1, i_3 = 4, i_4 = 3$ then $\{i_1, i_2, i_3, i_4\} = \{1, 2, 3, 4\}$ and the sequence $2, 1, 4, 3$ defines permutation such that $\sigma(1) = 2$, $\sigma(2) = 1$, $\sigma(3) = 4$ and $\sigma(4) = 3$.

Proposition 6.2.12. *If* $I, J, K, L \in \binom{[m]}{2}$ *satisfy (6.2.4) and* $I \Delta J = K \Delta L$ *then for some distinct* i, j, k, l*, where* $I = \{i, j\}, J = \{i, k\}, K = \{j, l\}, L = \{k, l\}$ *then*

$$\frac{1}{n^2(n-1)^2} \operatorname{cov}(\det W_{I,J}, \det W_{K,L}) = (-1)^{\operatorname{sgn}(\sigma)} \frac{1}{n}$$

where σ *is a permutation of the set* $\{i, j, k, l\}$ *defined by the sequence* i, j, k, l*.*

Proof. Since $I \neq J$ then $\mathbb{E}(\det W_{I,J}) = 0$ and hence to compute this covariance it suffices to compute $\mathbb{E}(\det W_{I,J} \det W_{K,L})$. By Theorem 6.2.9 this second order moment is equal (up to sign) to $1/n$ and the sign is positive if $i < j < k < l$. Because I, J, K, L satisfy (6.2.4), we necessarily have $j < k$ and $i < l$ so there are six possible orderings

$$\begin{aligned} &i < j < k < l, && j < k < i < l, && j < i < l < k \\ &i < l < j < k, && j < i < k < l, && i < j < l < k. \end{aligned}$$

Because $i < j < k < l$ gives a positive sign and the sign of $\det W_{ij,kl}$ changes if you swap rows or columns, we check directly that only the last two situations lead to negative values of $\mathbb{E}(\det W_{I,J} \det W_{K,L})$. Again, a direct check shows that only the last two sequences define permutations with negative parity. □

Proposition 6.2.13. *If* $I, J, K, L \in \binom{[m]}{2}$ *satisfy (6.2.4) and* $I \Delta J = K \Delta L$ *then for some distinct* i, j, k, l*, where* $I = \{i, j\}, J = \{k, l\}, K = \{i, k\}, L = \{j, l\}$ *then*

$$\frac{1}{n^2(n-1)^2} \operatorname{cov}(\det W_{I,J}, \det W_{K,L}) = (-1)^{\operatorname{sgn}(\sigma)} \frac{1}{n(n-1)}$$

where σ is a permutation of the set $\{i, j, k, l\}$ defined by the sequence i, j, k, l.

Proof. We proceed in a similar way as in the proof of Proposition 6.2.12. To determine the sign note that, because I, J, K, L satisfy (6.2.4), necessarily $\min\{i, j\} < \min\{k, l\}$, $\min\{i, k\} < \min\{j, l\}$ and $j < k$. There are three possible orderings that satisfy these constraints

$$i < j < k < l \qquad i < l < j < k \qquad i < j < l < k.$$

To see this note that $j < k$ and that there are three possible ways to position l: $j < k < l$, $l < j < k$ and $j < l < k$. But now for each of these orderings the position of i is already determined. Since the sign of $\mathbb{E}(\det W_{ij,kl} \det W_{ik,jl})$ is positive if $i < j < k < l$, then it is also positive if $i < l < j < k$ and it is negative for $i < j < l < k$. Again, a direct check confirms that the sign coincides with the parity of the corresponding permutation. □

6.3 Quartets and applications of tetrad analyses

6.3.1 The method of quartets

For any four distinct leaves $i, j, k, l \in V$ we say that $q_{ij,kl} = ij|kl$ forms a quartet of T if the paths $\overline{ij}$ and $\overline{kl}$ are disjoint. A binary tree T displays the set of quartets $\mathcal{Q}$ if each quartet $q \in \mathcal{Q}$ is a quartet of T. A set of quartets $\mathcal{Q}$ is said to determine T if T displays $\mathcal{Q}$ and T is the unique tree displayed by $\mathcal{Q}$ [Semple and Steel, 2003]. For T we denote such a set as $\mathcal{Q}_T$. Thus, quartets can be considered as fundamental components of binary trees; see also Dress et al. [2012]. A set $\mathcal{Q}_T$ is said to be minimal if there exists no element $q \in \mathcal{Q}_T$ such that $\mathcal{Q}_T \setminus \{q\}$ determines T. Grünewald et al. [2008, Theorem 2.4] provides the minimum size of any $\mathcal{Q}_T$ (i.e. the size of the smallest minimal defining quartet set). Furthermore, Semple and Steel [2003, Theorem 6.8.8] provides a quick method for constructing minimal defining sets of quartets that define binary phylogenetic trees.

Quartets are related to the $Q_{I,J}$ defined in (6.2.2) when $I, J \in \binom{m}{2}$. Let $V \subset U$ be such that $V = \{i, j, k, l\}$ with distinct elements. Consider three random variables $Q_{ik,jl}$, $Q_{il,jk}$ and $Q_{ij,kl}$. By Theorem 5.3.4 we expect that one of them will be approximately zero and the other

two will be equal. Using this fact, these $Q_{I,J}$ can be used to test the algebraic constraints in Proposition 5.4.2.

If an off-diagonal minor $\det(C_{I,J})$ is zero then $Q_{I,J}$ is zero on average but since the variance of $Q_{I,J}$ may be very high, the observed value of $Q_{I,J}$ may also deviate from zero substantially in certain scenarios. This means that we should standardise the data. Since we do not know C, the standardisation should depend on the sample covariance matrix $\hat{\Sigma}$. Testing hypotheses of the form $Q_{ij,kl} = 0$ is referred to as testing for vanishing tetrads, that is testing whether the quartet $q_{ij,kl}$ is displayed in T given the data.

To test a particular binary tree T, a set $\mathcal{Q}_T$ is required, i.e. a set of quartets $\mathcal{Q}$ that determines T. Ideally $\mathcal{Q}_T$ should be minimal to remove superfluous quartets. For each $q_{ij,kl} \in \mathcal{Q}_T$ consider the corresponding $Q_{ij,kl}$ as in (6.2.2) and define the set of these random variables as Q_T. For an arbitrary but fixed ordering, denote the column of sample means of these $Q_{ij,kl} \in Q_T$ by $\hat{Q}_T$ as provided by (6.2.2). This estimator is known to be consistent [Drton et al., 2007], and as a consequence as the sample size n tends to infinity any tree T is uniquely identified by the i, j, k, l such that $\hat{Q}_{ij,kl} = 0$. The corresponding sample covariance matrix $\hat{\Sigma}_{Q_T}$ has dimension $p = |Q_T|$. The sample covariance between minors $\hat{\Sigma}_{Q_T}$ can be obtained by calculating $\mathrm{cov}(W^{(2)})$ per the results for covariances of minors in Section 6.2.2 and by substituting C for the sample covariance of original variables $\hat{\Sigma}$. An appropriate simultaneous test statistic (6.3.1) is provided in Bollen and Ting [1993], which can be calculated and compared with the relevant value of the chi-square distribution.

$$\mathcal{T} = \hat{Q}^T \hat{\Sigma}_Q^{-1} \hat{Q} \sim \chi_p^2 \tag{6.3.1}$$

The subscript T is omitted for clarity and to avoid confusion with the matrix transpose superscript T. Compare (6.3.1) with Bollen and Ting [1993, (20)] where their Σ_{tt} is the covariance of $\sqrt{n}\hat{Q}$. Here the sample size n is incorporated implicitly through $\hat{\Sigma}_Q^{-1}$. Therefore, (6.3.1) provides a significance test for the equality constraints in (5.4.1), and furthermore, the required moments of $Q_{I,J}$ are given in Section 6.2. This provides a quick method for assessing whether a Gaussian data set appears consistent with the algebraic constraints associated with tree models.

A consideration with tetrad analyses is that there can be multiple $\mathcal{Q}_T$ and there may not always be an obvious reason for selecting one minimal defining quartet set over another. In such cases

it is recommended that a number of these sets are randomly selected to assess the robustness of the results. This issue is discussed further in Bollen and Ting [1993].

Hypothesis testing for vanishing tetrads can be used for both CTA and for ETA. There are many algorithms for obtaining candidate trees. However, often there is no way to assess the suitability of the optimal outputted tree. CTA takes a candidate tree and provides an absolute rather than relative value as to how well the data supports the purported tree. In contrast to a CTA, an ETA is used primarily when there are only a few variables because it is necessarily computationally intensive. Here the underlying tree is not known but there is an indication or prior belief that the data may be compatible with $M(T)$ for some phylogenetic tree T. Then an exhaustive search can then be performed across all trees, assessing each set of hypotheses using the observed data. If the set of plausible trees is empty then the assumption that the data is consistent with $M(T)$ may be incorrect.

6.4 Simulation results

We now investigate the reliability of the techniques through a variety of methods. The purpose is to give some comfort that the proposed methodology can be expected to work in practice.

6.4.1 Visualising reliability of semi-algebraic constraints

Here we explore the space of 3×3 positive definite correlation matrices with off-diagonal entries on a lattice of spacing 0.025. Each of the $314,087$ valid correlation matrices is treated as an observed estimate, and so $\hat{S}$ is obtained by multiplying by $n-1$ (where n is a chosen effective sample size). Using the inverse-Wishart prior as mentioned previously, the posterior density of $\Sigma|X$ is $\mathcal{W}_3^{-1}(3+n, I_3+\hat{S})$. This posterior density can then be used to simulate N realisations of Σ each of which can be tested against the tree constraints from which a posterior probability of tree-compatibility can be determined. As n increases, the effect of the prior diminishes. As N increases, the estimate of the posterior probability of tree-compatibility improves (given the particular prior). The following results are for $N = 100$ and for $n = 50, 200, 800$.

Table 6.1 displays measures of reliability of the tree constraints for different effective sample sizes n — this excludes cases where Σ lies exactly on the boundary of the tree being admissible (e.g. $\min\{|\rho_{12}|, |\rho_{13}|, |\rho_{23}|\} = 0$ for the positivity constraint) and so leaves 312,976 correlation matrices. These were excluded as no matter how large n is, the tree-compatibility remains sensitive to any perturbation, and exact boundary cases will not be observed in practice as long as the measurement precision is suitably high. The first column indicates n the effective sample size. The second column refers to the summary statistic being measured — in this case the proportion of posterior probabilities below certain thresholds and the mean posterior probabilities. The next two columns relate to those posteriors based on covariance Σ that is T_3-compatible. The final two columns relate to those posteriors that are not T_3-compatible. These in turn are split for posterior probabilities of satisfying the positivity (5.1.1) and tripod constraints (5.4.2) (denoted 'pos' and 'tri') respectively. Recall from Remark 5.4.3 that the tripod triple product implies the positivity constraint so the columns 'tri' are usually of more interest. The results are also displayed visually in Figures 6.2–6.5 where the effect of sample size can be observed and it is apparent that proximity to boundaries of the regions can be seen to have an effect on the posterior probability of tree-compatibility.

FIGURE 6.2: Space of correlations relating to 3×3 positive definite correlation matrices that satisfy the positivity constraint. The colour indicates the posterior probability of T_3-compatibility for respective sample sizes $n = \{50, 200, 800\}$.

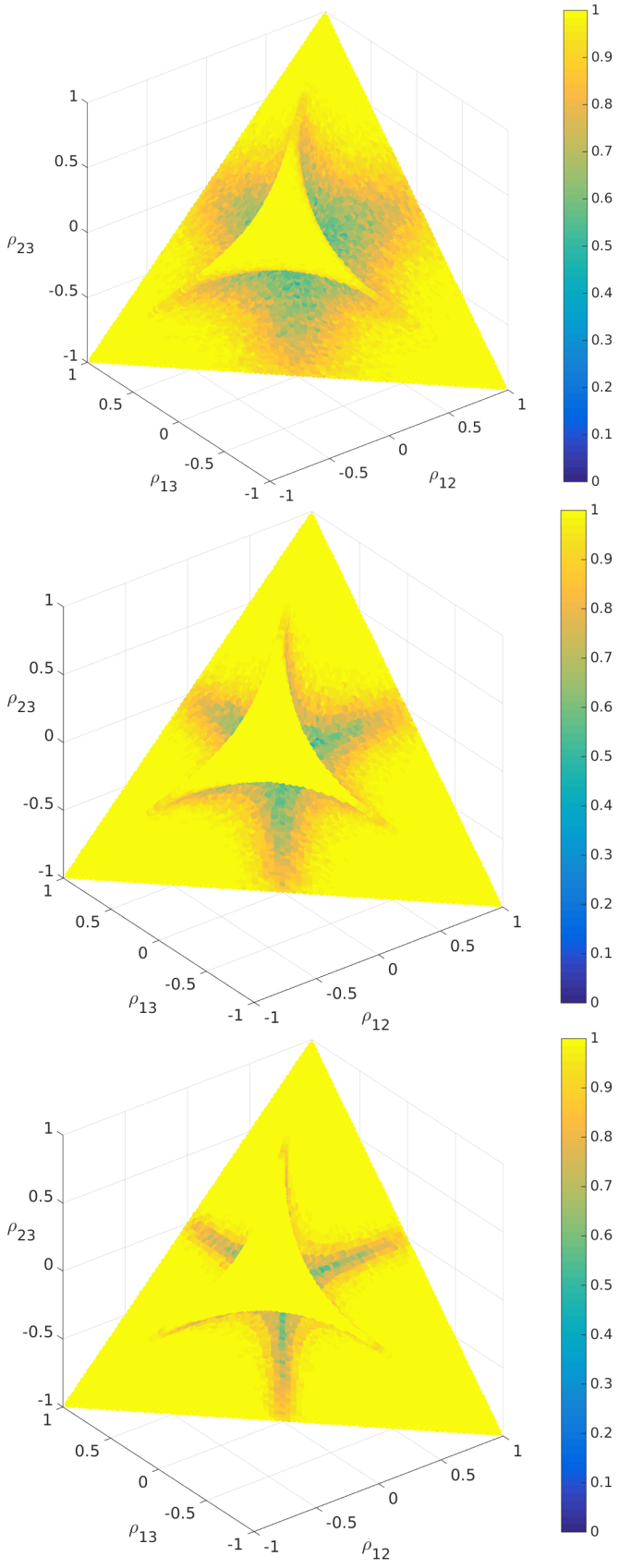


FIGURE 6.3: Space of correlations relating to 3×3 positive definite correlation matrices that do not satisfy the positivity constraint. The colour indicates the posterior probability of T_3-compatibility for respective sample sizes $n = \{50, 200, 800\}$.

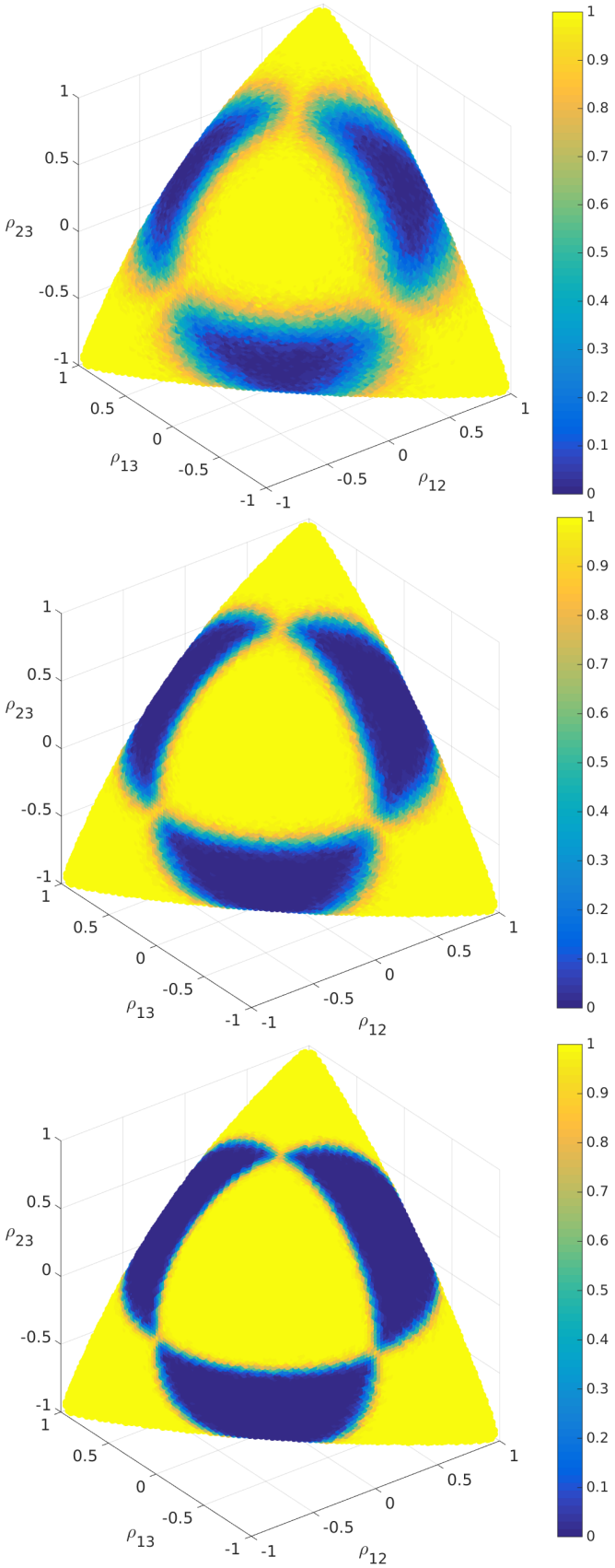


FIGURE 6.4: Space of correlations relating to 3×3 positive definite correlation matrices that satisfy the tripod constraints. The colour indicates the posterior probability of T_3-compatibility for respective sample sizes $n = \{50, 200, 800\}$.

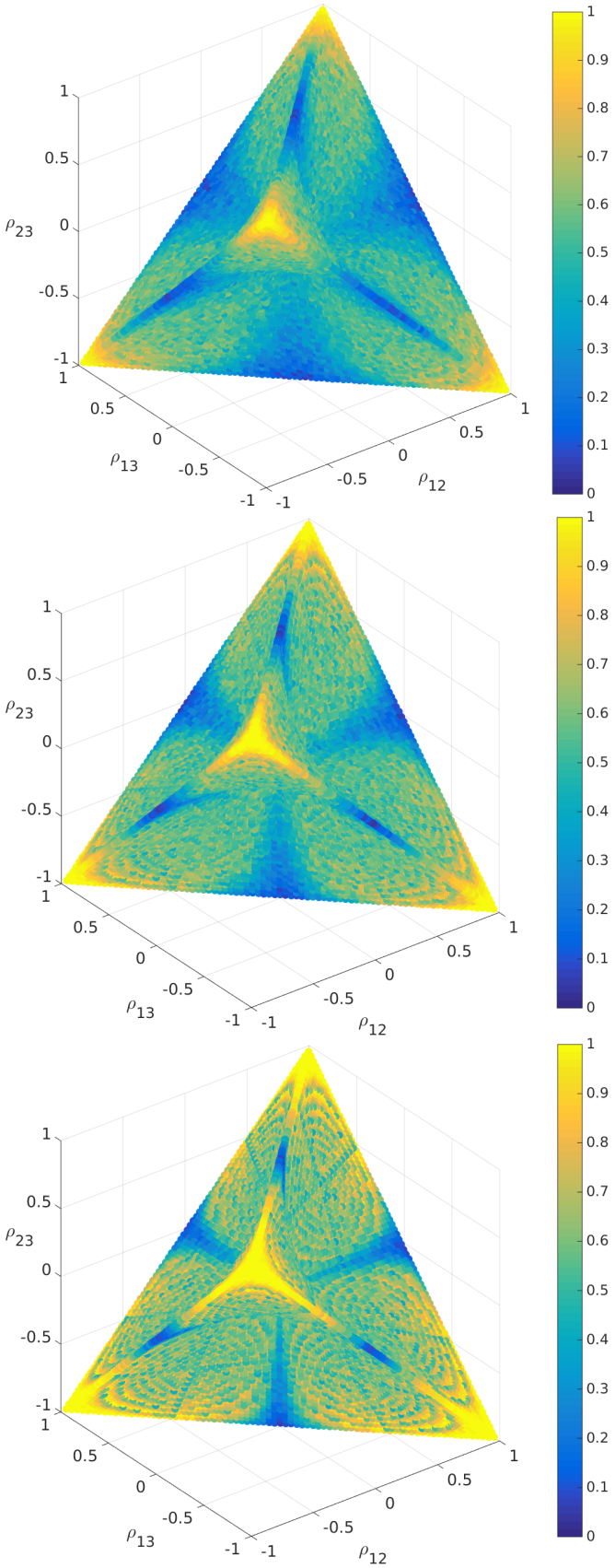
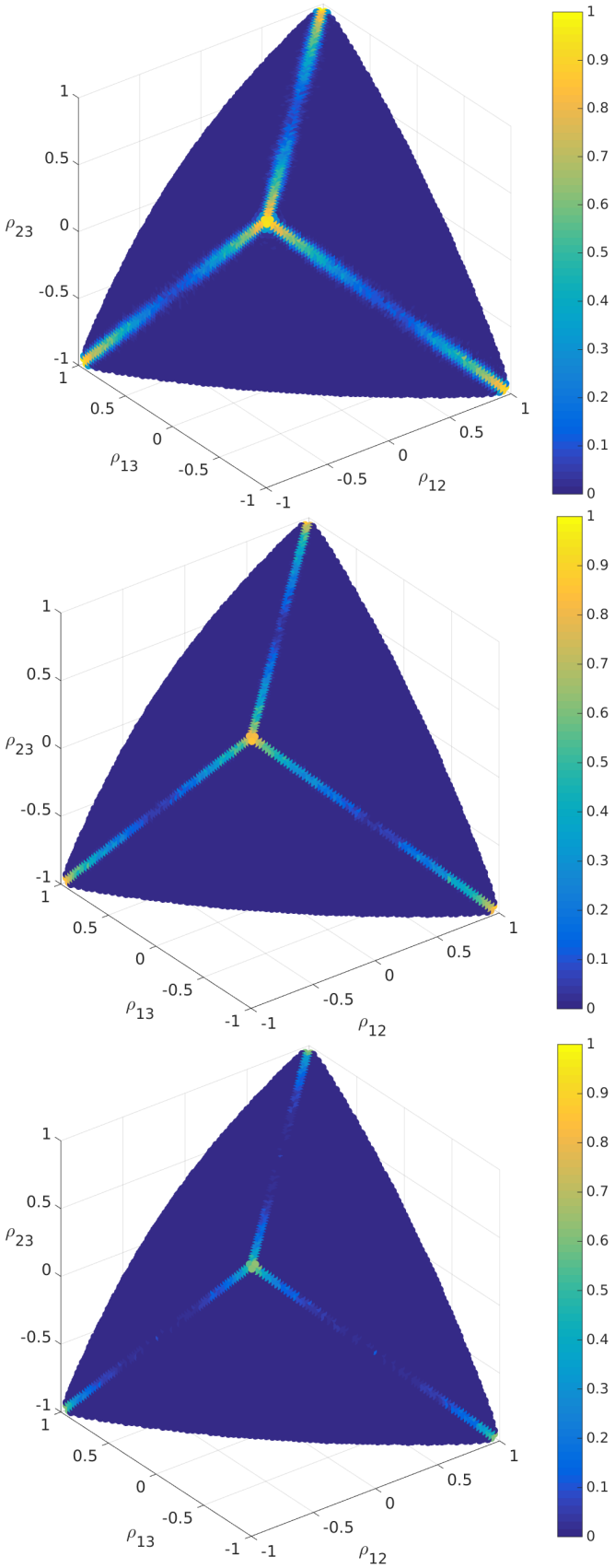


FIGURE 6.5: Space of correlations relating to 3×3 positive definite correlation matrices that do not satisfy the tripod constraints. The colour indicates the posterior probability of T_3-compatibility for respective sample sizes $n = \{50, 200, 800\}$.



The results in Table 6.1 show that as sample size increases the performance improves across the board. For instance there are fewer cases of false rejections of T_3-compatibility and fewer cases of failing to reject T_3-compatibility. Another remark is that the positivity constraint alone does not perform as well as the tripod product constraint (5.4.2), but that is not a particular problem as the full tripod constraints are advocated in Section 6.1.

TABLE 6.1: This table displays the proportion of the $312,976$ valid posterior probabilities that are below the thresholds 0.01, 0.05 and 0.10 along with the mean posterior probabilities. This is split between Σ that are T_3-compatible and those that are not.

		Given Σ is T_3-compatible		**Given Σ is not T_3-compatible**	
n	measure	pos	tri	pos	tri
50	<0.01	0	0	0.0171	0.4700
	<0.05	0	<0.0001	0.0659	0.6055
	<0.10	0	0.0005	0.1095	0.6824
	mean	0.9200	0.6868	0.5908	0.0945
200	<0.01	0	0	0.1328	0.7060
	<0.05	0	<0.0001	0.2059	0.7923
	<0.10	0	0.0004	0.2482	0.8354
	mean	0.9724	0.8211	0.5680	0.0494
800	<0.01	0	0	0.2592	0.8442
	<0.05	0	0	0.3100	0.8933
	<0.10	0	0.0002	0.3429	0.9171
	mean	0.9913	0.9058	0.5603	0.0244

More generally we can observe that the posterior probability is conservative with rejection in so much as it is very unlikely to reject tree-compatibility when the source is truly tree-compatible but does not reject tree-compatibility enough when the underlying Σ is not tree-compatible. Another way of summarising this is that the technique has very high specificity but lower sensitivity. The sensitivity improves when considering a higher threshold or larger sample size. For example, when using the threshold 0.01 with $n = 50$ sensitivity is 0.47. Increasing to $n = 800$ gives sensitivity 0.84 and using the threshold 0.1 increases it to about 0.92. In all combinations studied, specificity is approximately 1 even with higher thresholds. In practical terms, this suggests that if a tree is rejected then this appears to be strong evidence that the data is truly not compatible. In contrast, we are cautious to reject tree-compatibility so failure to reject is less conclusive.

6.4.2 Moment estimators for dimension four

We now show how the estimators for the mean and variance appear to converge to the true values for simulated data. Consider a random sample of size n from the four dimensional Gaussian distribution with mean zero and covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} 4 & 2.1 & 2.24 & 3.15 \\ \cdot & 9 & 4.704 & 6.615 \\ \cdot & \cdot & 16 & 14.4 \\ \cdot & \cdot & \cdot & 25 \end{bmatrix}, \qquad R = \begin{bmatrix} 1 & 0.35 & 0.28 & 0.315 \\ \cdot & 1 & 0.392 & 0.441 \\ \cdot & \cdot & 1 & 0.72 \\ \cdot & \cdot & \cdot & 1 \end{bmatrix}.$$

We perform the following simulations. We first fix $n = 100$. To evaluate the mean and the variance of the $Q_{I,J}$'s, we replicate our computations 1000 times. Each time we generate a sample of size 100, compute the corresponding statistics S and the minors. This first matrix contain the means of the estimators:.

$$\begin{bmatrix} 31.6 & 14.1 & 19.8 & -10.3 & -14.5 & \mathbf{0.0} \\ \cdot & 58.9 & 50.5 & 23.0 & \mathbf{15.4} & -18.1 \\ \cdot & \cdot & 90.0 & \mathbf{15.4} & 31.5 & 10.6 \\ \cdot & \cdot & \cdot & 121.6 & 98.3 & -38.1 \\ \cdot & \cdot & \cdot & \cdot & 180.9 & 22.3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 192.3 \end{bmatrix}$$

This second matrix contains the theoretical actual values of the minors.

$$\begin{bmatrix} 31.6 & 14.1 & 19.9 & -10.3 & -14.5 & \mathbf{0.0} \\ \cdot & 59.0 & 50.5 & 23.1 & \mathbf{15.4} & -18.1 \\ \cdot & \cdot & 90.1 & \mathbf{15.4} & 31.7 & 10.6 \\ \cdot & \cdot & \cdot & 121.9 & 98.5 & -38.1 \\ \cdot & \cdot & \cdot & \cdot & 181.2 & 22.3 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 192.6 \end{bmatrix}$$

The rows and columns of this matrix correspond to the six size-two subsets of $\{1,2,3,4\}$ ordered in a natural way: $\{1,2\}$, $\{1,3\}$, $\{1,4\}$, $\{2,3\}$, $\{2,4\}$, $\{3,4\}$. So for example the element $(2,5)$ of this matrix is the sample mean of $Q_{13,24}$. The three boldfaced elements correspond to the off-diagonal minors: $Q_{12,34}$, $Q_{13,24}$ and $Q_{14,23}$.

The estimated variances of $Q_{I,J}$ are given in this first matrix

$$\begin{bmatrix} 41.6 & 25.2 & 41.6 & 42.1 & 64.7 & \mathbf{15.3} \\ \cdot & 142.0 & 132.8 & 91.9 & \mathbf{70.9} & 127.2 \\ \cdot & \cdot & 336.9 & \mathbf{74.3} & 202.6 & 180.9 \\ \cdot & \cdot & \cdot & 605.8 & 510.0 & 287.5 \\ \cdot & \cdot & \cdot & \cdot & 1335.6 & 373.8 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1526.1 \end{bmatrix}$$

This second matrix provides the true theoretical variances computed from Proposition 6.2.6.

$$\begin{bmatrix} 40.5 & 25.2 & 41.3 & 42.9 & 65.3 & \mathbf{15.8} \\ \cdot & 141.3 & 132.2 & 90.2 & \mathbf{70.5} & 127.0 \\ \cdot & \cdot & 329.5 & \mathbf{73.4} & 198.6 & 182.2 \\ \cdot & \cdot & \cdot & 603.1 & 521.5 & 285.9 \\ \cdot & \cdot & \cdot & \cdot & 1333.9 & 374.9 \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1506.9 \end{bmatrix}$$

Repeating the simulation for 1000 random positive definite covariances matrices, the average percentage difference between estimates and true values are 2.5% and 3.9% for the means matrix and variance matrix respectively. Even though we used only 1000 replications, the estimation is very good — especially given the fact that in order to estimate the variance of $Q_{I,J}$ we need to estimate fourth order moments of S. However, the next subsection gives a more thorough way to investigate the estimators.

6.4.3 Assessment of test statistic

Another approach to assessing the estimators is to investigate how well the test statistic matches the intended distribution. That is, for particular sample sizes how well does the distribution of $\mathcal{T}$ match the chi-squared distribution. This assessment also acts as a check that the coding is working as expected. If it is then the distribution of T will be closer to the chi-squared distribution as the number of samples n increases.

We consider the quintet tree T_5 as in Figure 5.2. Our aim is to show that under the null hypothesis of T_5-compatibility, as the sample size n increases, the distance between the empirical CDF of $\mathcal{T}$ and the CDF of the χ^2 distribution decreases. Specifically, the χ^2 distribution we consider has 2 degrees of freedom as two embedded quartets are simultaneously tested for T_5.

For a single run of the simulation, the 7 edge correlations of T_5 are sampled from the Uniform(-1,1) and then the correlations between leaf variables are calculated to produce a 5×5 correlation matrix M. A diagonal matrix D is produced with the diagonal entries sampled from Uniform(1,5). Then define $\Sigma = DMD$ which is a covariance matrix that is T_5 compatible. Then n samples are simulated from $N_5(0, \Sigma)$. This is repeated 1000 times for the same Σ. For each of these 1000 sets of n samples the test statistic $\mathcal{T}$ is calculated. The empirical density function of the test statistic can then be estimated from these 1000 test statistics and compared to the χ^2_2 density function as is done visually in Figure 6.6.

The Kolmogorov–Smirnov (KS) statistic [Pollard, 1979, Chapter 12] is also calculated between the empirical CDF and χ^2_2 CDF. The KS statistic is selected as it accounts for differences in both shape and location. The one sample KS statistic is calculated as

$$K_n = \sup_x |F_n(x) - F(x)|$$

where n is the sample size, sup is the supremum, $F_n(x)$ is the empirical CDF, $F(x)$ is the CDF of the known distribution.

For a fixed Σ we repeat the process for three sample sizes: $n = 50, 200, 800$. Thus the output for each of the three n is a set of 1000 Kolmogorov–Smirnov statistics — the smaller the value

the closer the two CDFs. We then opt to repeat the whole process 250 times for 250 different generated $\Sigma = DMD$. Therefore, at the end we have a $250 \times 3 \times 1000$ array of KS statistics.

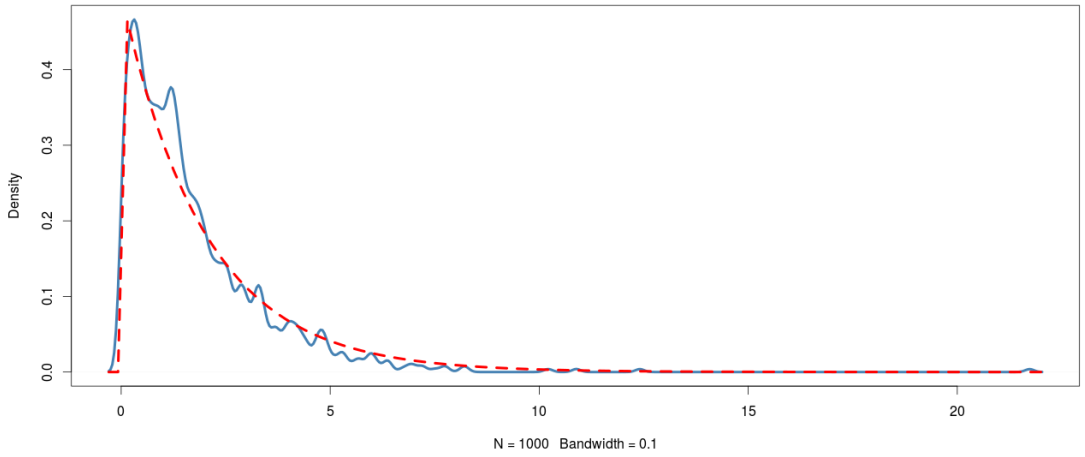


FIGURE 6.6: An example of an empirical density (solid blue line) plotted with a chi-squared degree 2 density (red dashed line). This is for sample size $n = 800$ and $N = 1000$ replications as described in the main text.

Figure 6.7 plots the values of the 250×3 KS statistics. The KS statistics related to the three sample sizes $n = 50, 200, 800$ are represented by 'x', 'y' and 'z' respectively in the colours black, red and blue respectively. To make the plot visually clearer, the 250 replicates are ordered ascendingly according to the value of the KS statistic for the $n = 50$ case along the horizontal axis. It is apparent that the larger the sample size, the smaller the KS statistic and hence the smaller the differences in shape and location of the empirical CDF which is a desirable property. Together with the previous simulations, this gives confidence that the proposed techniques are suitable for use with the data sets to be considered in Chapter 7.

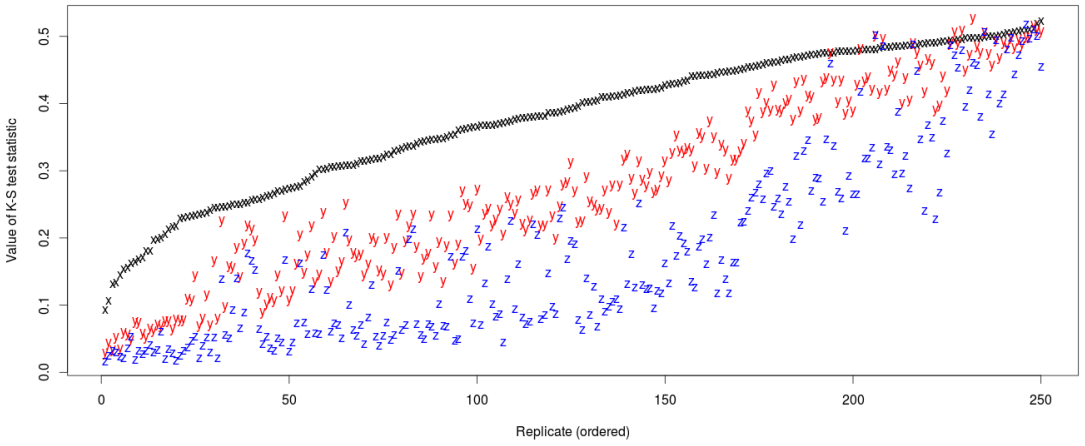


FIGURE 6.7: Comparison of KS statistics for each of the 250 simulated parameter sets — the lower the value, the closer the CDFs of the empirical density and the chi-squared density. The black 'x' represents $n = 50$, red 'y' represents $n = 200$ and blue 'z' represents $n = 800$. For visual clarity, the 250 replicates are sorted in ascending order of the $n = 50$ KS statistics.

Chapter 7

Applications of Gaussian tree constraints

This chapter draws together all the previous chapters by implementing two novel examples that make use of Gaussian tree constraints and the associated methodology. The first is using the Romance language functional data set described in Section 3.3 and is based upon the paper Shiers et al. [2014]. The second is an example from biology, which expands upon the final example described in Shiers et al. [2016] that relates to functional growth curves of yeast species. In both instances, we are interested in whether a GLTM is a suitable graphical description for the relationships between languages and species respectively. We use Gaussian tree constraints to assess probabilistically whether a tree model is appropriate given the observed data. Applications of the suite of tools for Gaussian tree constraints are not found elsewhere in the literature.

7.1 Application of Gaussian tree constraints to acoustic linguistic functional data

Recall from Section 3.3 that we are considering a linguistic data set comprising phonetic functional data from five Romance languages: French, Italian, Portuguese, and two forms of Spanish (American and Iberian). The observations are spectrograms of speakers saying numbers in their respective languages. Here the evolutionary dependencies between spoken numbers is studied

with each extant language treated as a possible leaf vertex of a graphical model. Using the novel combination of separable-CVA as an approximation to separable-CFA (as described in Section 3.6) provides an ordered basis on which to project the high dimensional data to a lower dimension. Each dimension of the projected data set accounts for a particular combination of phonetic variation and each dimension is considered independently, thus allowing differing evolutionary relationships for different aspects of the speech. Using a selection of the distributional techniques described in Section 6 we are able to quantify the probability of tree-compatibility.

7.1.1 Application of CVA as an approximation to CFA

As motivated, the separable-CVA is used to approximate the separable-CFA of the Romance languages data to achieve a dimension reduction based on components which maximise between- to within-language variability. This is deemed a suitable approximation to make as the functional spectrograms have been sufficiently densely sampled during discretisation. This is backed-up visually by the smoothness of the plots in Figure 3.4.

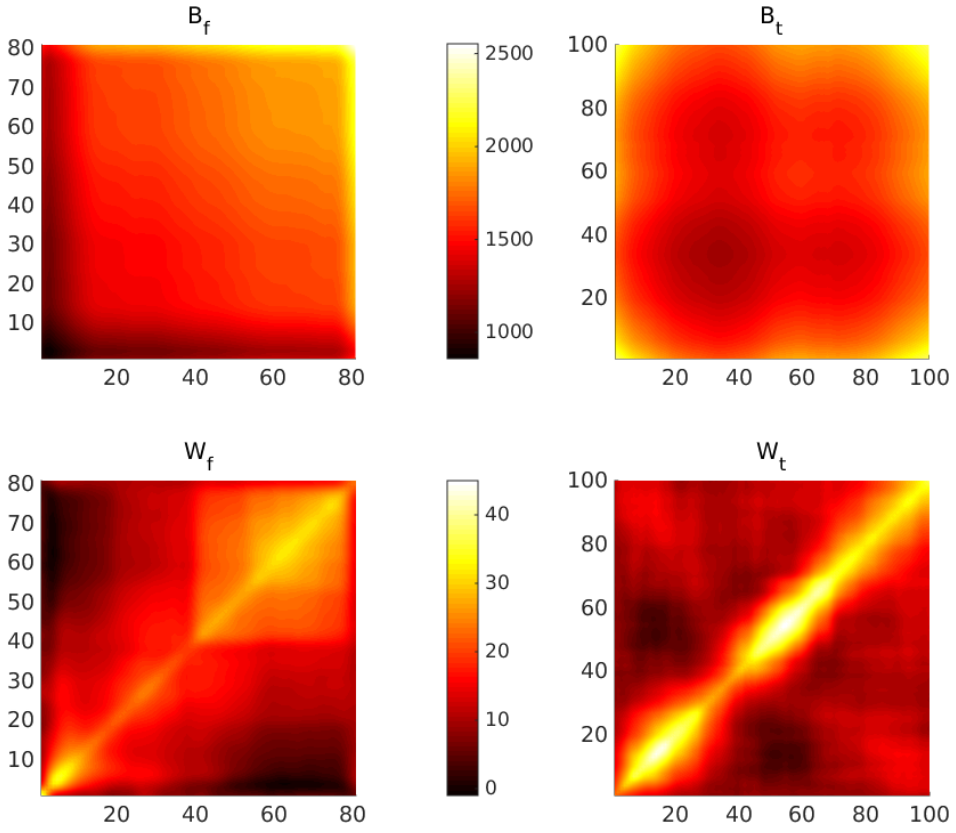


FIGURE 7.1: Sample between-language and within-language covariances of speech data for frequency and time directions.

Figure 7.1 displays the interpolated between-language and within-language covariances of the frequency and time dimensions respectively. Reassuringly the covariances are almost exclusively positively correlated and furthermore the between-language variation is larger than the within-language variation. Other features of note are the ridges along the diagonals of the within-language covariances indicating that similar times and frequencies are highly positively correlated, which is to be expected. The high correlations in the corners of the between-language covariance $\boldsymbol{B}_t$ are to be expected given beginnings and ends of words are likely to be very quiet and hence very similar. In the within-group frequency covariance $\boldsymbol{B}_f$ there is a slight block structure associated with the last 40 points. These relate to some of the recordings being at lower audio sampling rates capturing fewer details in the higher frequencies. The effect of excluding these observations is investigated by rerunning the analyses performed in Section 7.1.4 with the results reported in Section 7.1.6.

Before implementing separable-CVA, we test an assumption of CVA, which is that within-language covariances are equal. In an empirical sense we clearly need to allow for some sampling variation. One appropriate method is to use Box's M statistic [Box, 1949] which can be compared with a chi-squared distribution to potentially reject the homogeneity of covariances. For frequency and time directions in turn, we perform the test on a pair-wise and 5-way basis, the former approach comparing each of the within-language covariances with the average of the remaining four. Using the `MATLAB` package `MBoxtstwod` [Trujillo-Ortiz et al., 2004], none of the covariances were found to be significantly different at the 0.01 level and thus the CVA assumption of homogeneous within-group covariances is not rejected. We can now proceed with the separable-CVA with some comfort that the homogeneous within-groups covariance assumption appears to be reasonable. This is not the only assumption of associated with CVA. Multivariate Gaussianity is also assessed in Section 7.1.6.

Recall that we are using CVA with the aim of projecting the spectrograms to a lower dimensional space while retaining important linguistic information regarding the differences between languages. When selecting a dimension r to project to, it is unusual to have anything but an arbitrary albeit sensible method for selecting r. However, in some acoustic contexts (e.g. Hadjipantelis et al. [2012]) thresholds can be proposed based on sounds which are audible to humans. Otherwise, equivalent techniques to those employed with PCA [Jolliffe, 2002] can be used. For

example, by considering $\frac{\lambda_1+\lambda_2+\ldots+\lambda_r}{\lambda_1+\lambda_2+\ldots+\lambda_s}$ (the proportion of variability taken up by the first r canonical variates), a threshold can be specified to determine a value for r. For this linguistic study Figure 7.2 shows the cumulative variation explained by selecting particular numbers of components. Note that by construction each subsequent dimension accounts for less variability than the previous. Figure 7.2 indicates that the time and frequency projections perform extremely well as close to 97% of the between- to within-language variances can be explained by just a single component. By definition, the combined time and frequency projection is less efficient yet it manages to capture over 94% in the first component. This indicates that the separability assumption in this instance does appear a plausible one. In order to select a dimension, we can retain only those dimensions that contribute at least 0.1% contribution of explained variance. In this instance, this leads to selecting the first $r = 9$ canonical variates, a dimension reduction from 8100 to 9. These account for almost 97.5% of the variability. Each one of these 9 components $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_9$ accounts for some mode of variability between languages. Although the earlier components have high explanatory power, the latter components may isolate directions of variability which are of more interest from a linguistic perspective.

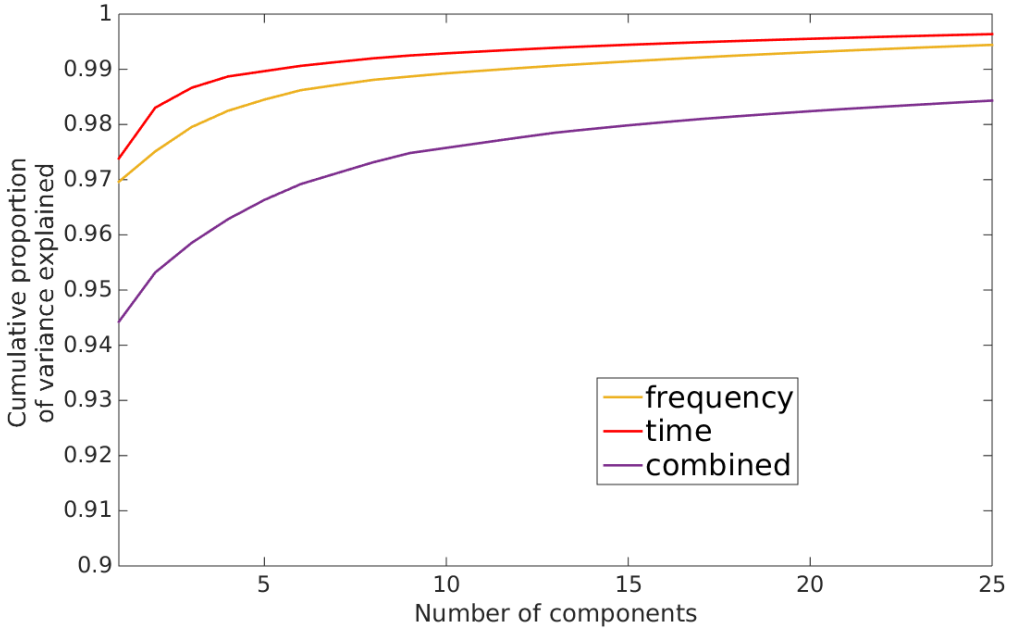


FIGURE 7.2: Cumulative variation explained by number of components. The explanatory power of the first component in terms of between- to within-language combined variability is over 94%.

It is clear from the description given in Section 3.6 that separable-CVA is not implemented on the belief that it reflects some underlying data generating process. Instead, separable-CVA is

simply a practical tool that produces a meaningful representation of the original data in a lower dimension. To demonstrate the effectiveness of projection of the spectrograms to even two dimensions, the projections of the means of the word observations are plotted in Figure 7.3 for all combinations using the first four components. We plot ellipses for each language where the direction of the major and minor axes are obtained as eigenvectors from a PCA of the projected means, where the centre point is the mean of projections, and the radii relate to 95% confidence intervals based upon the standard deviation in the directions of the axes (under Gaussian assumptions). The results of the separable-CVA are encouraging: there are clear groupings in all of the projections. In some, there is overlap between languages such as Italian and Spanish (Iberian) for dimensions 1 and 2, and again in dimensions 1 and 4. It is the projection to dimensions 2 and 3 that appears to perform the best at separating the groups, which indicates that the proportion of variance accounted for is not the only relevant factor in projections. This is consistent with our understanding of CVA. Thus it is possible that the most efficient direction for CVA achieves good within-language projection and good between-language separation for 9 of the 10 pairings but at the cost of projecting a single pairing close together.

Given that the acoustic data set is undoubtedly noisy and a reduction from 8100 dimensions to just 2 is large, this demonstrates the effectiveness of separable-CVA at selecting components which discriminate on a group basis. Note that whilst CVA operates on all languages simultaneously rather than in a pairwise manner, this does not necessarily imply languages in close proximity post-projection share particular acoustic features. However, a particular projection can be examined in more detail, for example through studying Hadamard products for each projection as demonstrated in Section 7.1.3.

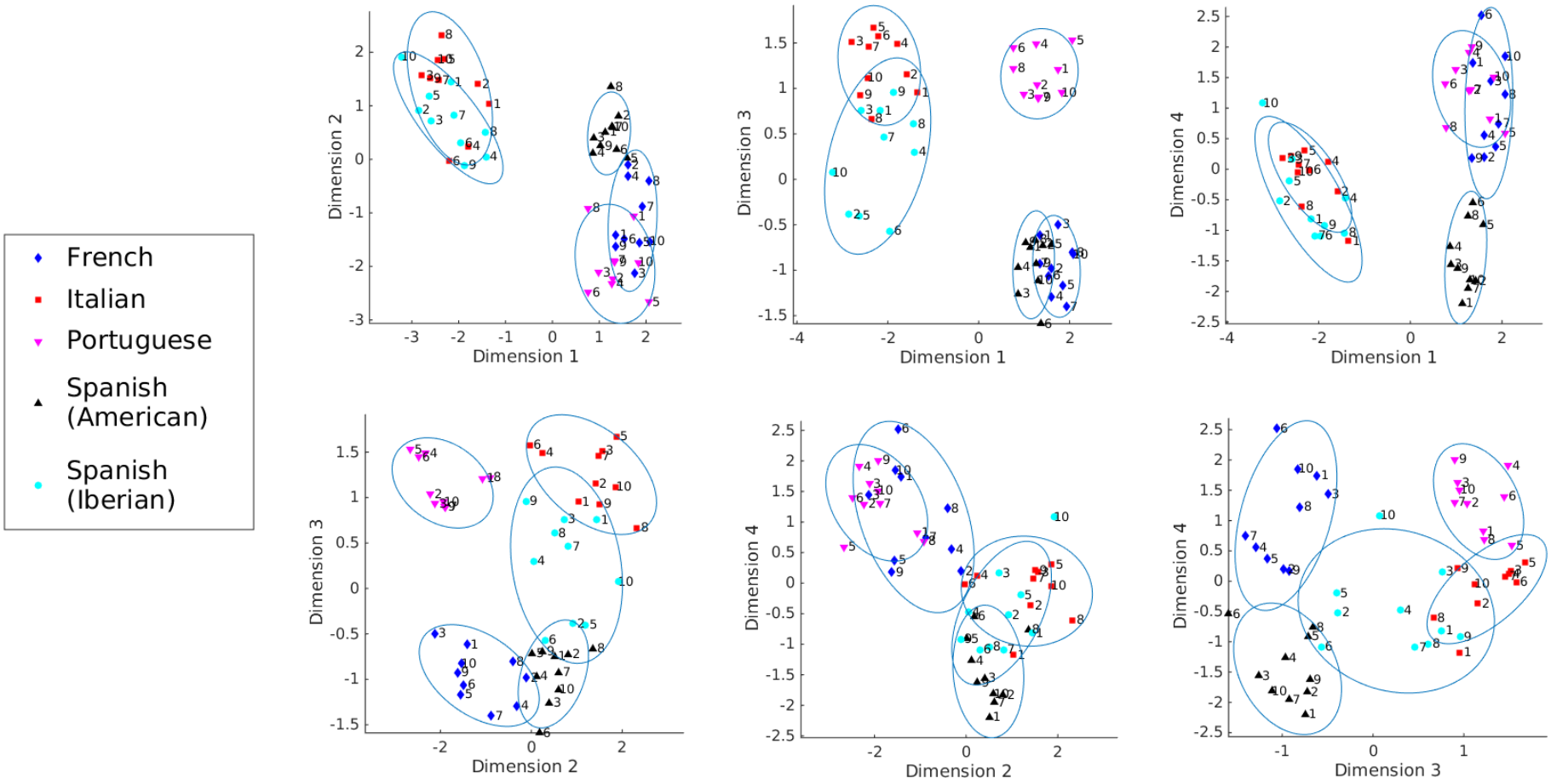


FIGURE 7.3: Two dimensional separable-CVA projection of means of word observations. The details of which linguistic factors are contributing to the projections are explored in Section 7.1.4.

7.1.2 Constructing a suitable covariance statistic

In pursuit of assessing tree-compatibility of the $r = 9$ projections of the Romance data using the positivity constraint $\sigma_{ij}\sigma_{ik}\sigma_{jk} \geq 0$, it is clear that a sample covariance of the scores must be constructed. Recall that the relationships of interest in this study are at the language level and thus between-language covariances (each 5×5) are the appropriate statistics to produce, one for each of the r components. This can be achieved through r covariance matrices of size 5×5 (one for each of the components), or alternatively as one large matrix $5r \times 5r$ with particular elements extracted when needed. The former provides simpler notation and so is adopted here, although it is worth noting that the latter is more generic and allows for scaling up to future multivariate tree constraint testing.

One approach to calculating the entries of these matrices is to treat the mean score of each word in a language as an observation and then measure the distance from the overall word mean projection. Then using appropriate weights, a between-language covariance matrix can be estimated as follows. Let $\bar{y}^i_d = \frac{1}{m_{\bullet d}} \sum_{l=1}^{n_l} m_{ld}\bar{y}^i_{ld}$, $m_{\bullet d} = \sum_l^{n_l} m_{ld}$ where recall m_{ld} is the number of samples of word d in language l, and $\bar{y}^i_{ld} = \boldsymbol{c}_i\bar{x}_{ld}$ the projection of the mean of word d of language l using component $\boldsymbol{c}_i$. Then for component i the between-groups cross-covariance for the projected data has the following form:

$$\boldsymbol{\Sigma}_{\boldsymbol{Y_i}} = [\sigma^i_{l,l'}] \text{ where } \sigma^i_{l,l'} = \sum_{d=1}^{n_d} \frac{\sqrt{m_{ld}}\sqrt{m_{l'd}}(\bar{y}^i_{ld} - \bar{y}^i_d)(\bar{y}^i_{l'd} - \bar{y}^i_d)}{n_d - 1} \tag{7.1.1}$$

where, as before, n_d denotes the number of unique words. Note that this between-group covariance differs from that used in the CVA — this is of the projected data and the word means are used to provide an observational summary of the data. This is a valid construction in the sense that (7.1.1) is an inner product (see Istratescu [1987] for instance). Furthermore, for the cross-covariance to be meaningful, equivalent statistics must be compared, in this case per language word means. The sample matrices $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y_i}}$ will be rank deficient if $n_l \geq n_d$. Also, observe that if for at least one word d the number of observations is unequal across languages then the weighted word mean $\bar{y}^i_d$ differs from the unweighted version. This relaxes a zero-sum condition on the rows or columns of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y_i}}$ permitting the covariance matrix to be full rank. In the alternate case of

a balanced observational design, full rank can be achieved through an alternative construction (for example adding the unweighted word means back to each language-word mean).

Now component-by-component covariances can be used to indicate adherence to a Gaussian tree model using the tripod tree positivity constraint on all $\binom{5}{3}$ selections of languages. Each component captures a different combination of variability. Thus it is not unexpected that some components may show violations of the constraint whereas others may indicate tree-compatibility. Irrelevant of whether a component is tree-compatible, we may be interested as to how to interpret the specific component from a linguistic perspective. We illustrate a method for exploring this in more detail in Section 7.1.3.

7.1.3 Investigating particular projections

To develop an insight into which aspects of the languages are being identified by the separable-CVA and screen for any spurious projections. This can be done at the component level by studying the associated eigenvector. However, while this can give some broad insight, without reference to the original (unprojected data) it can fail to isolate the features of the data that are contributing to each dimension. To establish whether there are any particular frequency ranges or periods of time that are being highlighted by the CVA it is useful to consider two languages at a time in the dimension of choice.

To illustrate one method of identifying key features separating languages we consider the following language pairs and dimensions:

- Spanish American and Spanish Iberian in the first dimension
- Spanish American and Portuguese in the second dimension
- French and Italian in the third dimension
- Spanish American and Portuguese in the fourth dimension

These particular components were selected as examples as they appear to have been effectively separated in the dimension of interest as can be seen from the plots in Figure 7.3. To explore the selected language pairs in detail, we consider the Hadamard (entrywise) products of the

relevant component with the mean language spectrogram of interest. That is, we take the matrix representation of language l mean spectrogram $\bar{\boldsymbol{X}}_l$ and perform the Hadamard product with $\boldsymbol{C}_i$ the matrix representation of component c_i:

$$\bar{\boldsymbol{H}}_l^i = \boldsymbol{C}_i \circ \boldsymbol{X}_l \text{ where } \bar{\boldsymbol{H}}_l^i[f,t] = \boldsymbol{C}_i[f,t] \times \boldsymbol{X}_l[f,t]$$

The resulting matrix $\bar{\boldsymbol{H}}_l^i$ indicates the contribution of each frequency-time point to the overall co-ordinates produced by the component projections. We can do this for any two languages in the same dimension and can visualise the differences between the two matrices by plotting the absolute difference. This helps to highlight areas of the frequency and time directions which are contributing to the between language separation. With expert linguistic analysis we can obtain preliminary interpretations of phonetic features associated with projections. We now illustrate this using the four examples listed above.

Example 7.1.1. Spanish American and Spanish Iberian in the first dimension

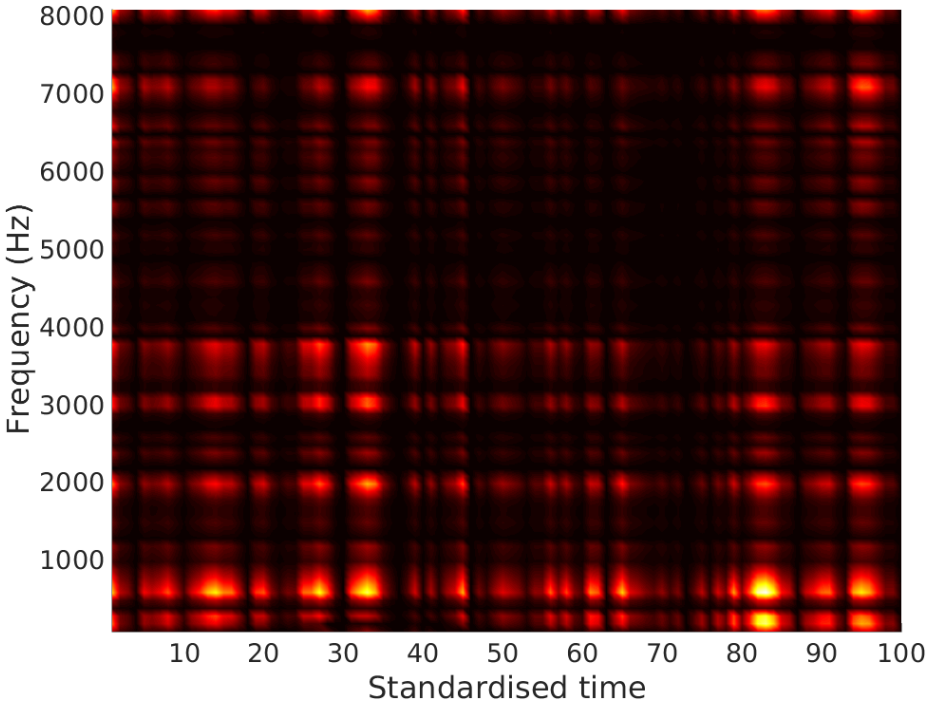


FIGURE 7.4: Absolute differences of Hadamard products for Spanish American and Spanish Iberian in the first dimension.

To read Figure 7.4, we are looking for areas of the plot that have the highest values (those indicated by lighter colours). A grid like pattern can be seen on the graph with particular areas standing out. This is often easier to read from the perspective plots (Figure 7.5 and Figure 7.6).

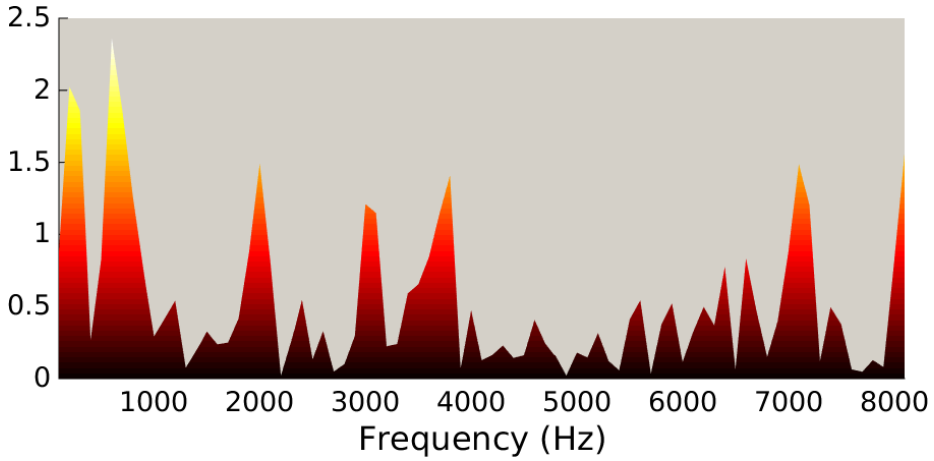


FIGURE 7.5: Frequency perspective of Figure 7.4.

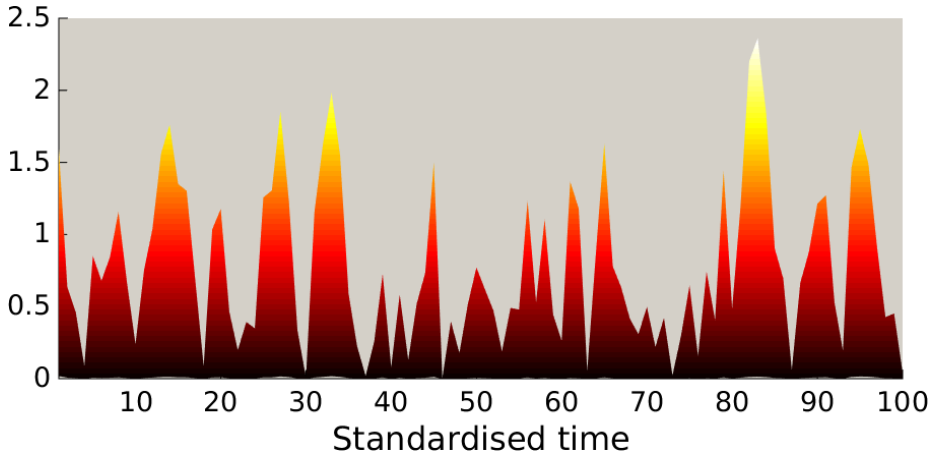


FIGURE 7.6: Time perspective of Figure 7.4.

We can see that the frequencies 0-1000Hz, 2000Hz, 3000-4000Hz 7000Hz and 8000Hz exhibit the highest values. This suggests that it is these frequencies that are contributing significantly to the overall projection differences between American and Iberian Spanish. The 0-1000Hz range likely relates to the first formant F1 being different, due to vowel differences. The variation at the highest frequencies (6000Hz+), are likely to be due to idiosyncratic differences in speakers (since humans cannot readily control speech frequencies in that range) or in the recordings (equipment or recording location).

From the time perspective there are many regions throughout the range that appear to be contributory, which suggests that the frequencies are of more interest. This makes sense intuitively as the difference in the two variants of Spanish would seem more likely to appear in pitch than in time differences given that the words being uttered are the same orthographically.

Example 7.1.2. Spanish American and Portuguese in the second dimension

In this example, we illustrate the importance of investigating projections in detail. Here we consider what appears to be an efficient projection that separates Spanish American and Portuguese. However, on closer inspection using Hadamard products we can see from Figure 7.7 that the projection alone is deceptive; there is a lone region that is responsible for the majority of the projection and its location is in a corner of the time-frequency space suggesting that this is identifying something spurious. This is confirmed in the subsequent perspective plots. Thus, although CVA is effective at separating grouped observations it can sometimes pick up on sections of the data that are uninteresting or artefacts. Even within the same component it is possible to have a mix of genuine and inconsequential projections depending on how the eigenvectors align with the data observations for different languages. The only way to screen for oddities such as this is to look in more detail at the make-up of the projections at the stage of interpretation of components or language-pair separations, the latter being performed here.

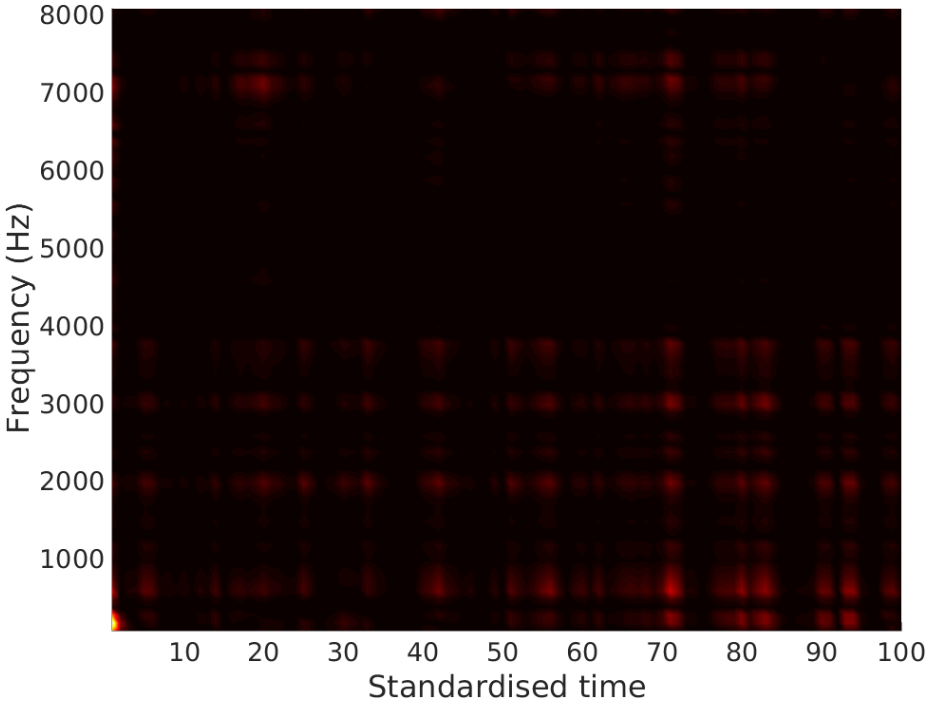


FIGURE 7.7: Absolute differences of Hadamard products for Spanish American and Portuguese in the second dimension.

Example 7.1.3. French and Italian in the third dimension

Once again the grid structure is apparent in the graphical representation of the differences in Hadamard products (Figure 7.10). In contrast to Figure 7.4, the time dimension appears to have distinct periods of lower contributory power.

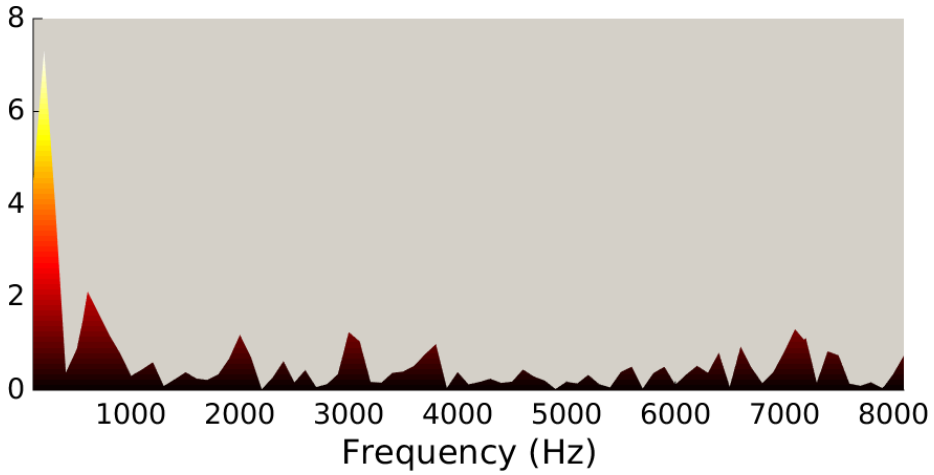


FIGURE 7.8: Frequency perspective of Figure 7.7.

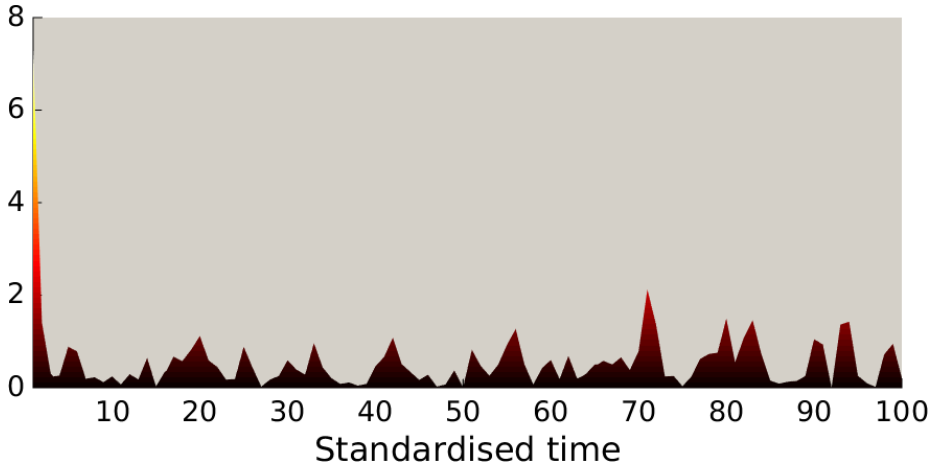


FIGURE 7.9: Time perspective of Figure 7.7.

As before, the details of important frequency and time points is more apparent from the perspective plots. The frequency plot is less clear than in the Figure 7.5 but we can broadly say that the higher frequencies appear to be of less interest, whereas the ranges 0-400Hz, 1500-2000Hz, 3000 and in particular the extended range 3500Hz-5000Hz all appear to significant. The first three formants for vowel sounds tend to be found in the range 0-5000Hz. This suggests that it is these frequencies that are contributing significantly to the overall projection differences between French and Italian.

From the time perspective, once again the description appears less relevant as there are no particular stand out regions. It is more notable that the range approximately 35-55 seems to play less of a role than the rest of the standardised time.

Example 7.1.4. Spanish American and Portuguese in the fourth dimension

Here the relevant frequency regions are relatively easy to read straight from the plot (Figure 7.13) with 3000-4000Hz appearing to be a key region. This range is in the region of the third formant and could correspond to differences in lip rounding between speakers of the languages. We can also see a small artefact in bottom left hand corner similar to that in Figure 7.7, but it does not appear to be dominating in this example.

The frequency perspective plot does not provide much more clarity in this situation. The time perspective is fairly consistent with ridges throughout the range with notable peaks around 35,

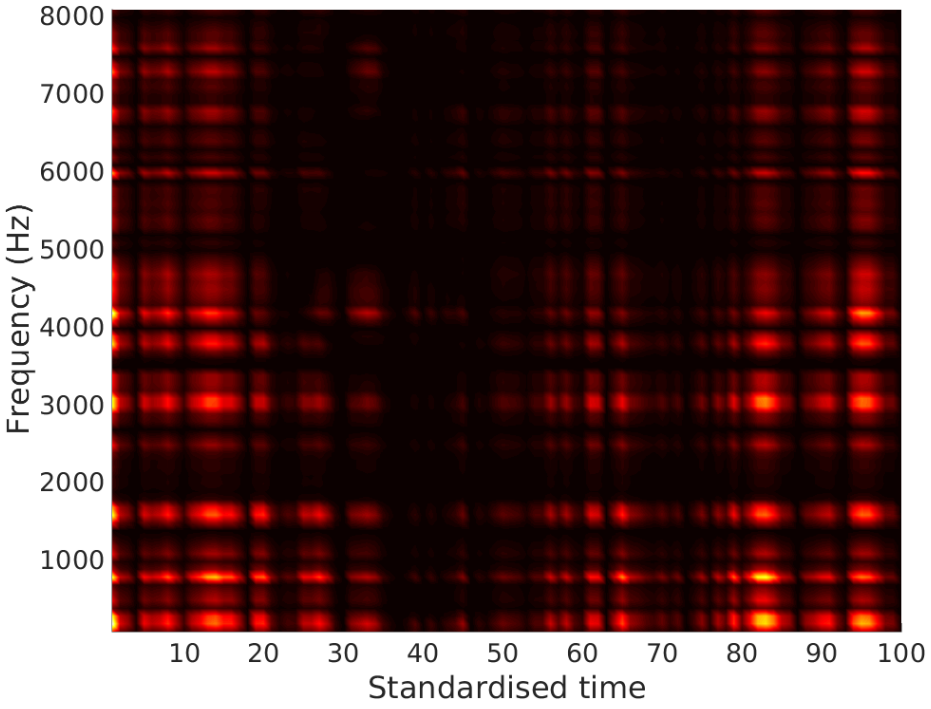


FIGURE 7.10: Absolute differences of Hadamard products for French and Italian in the third dimension.

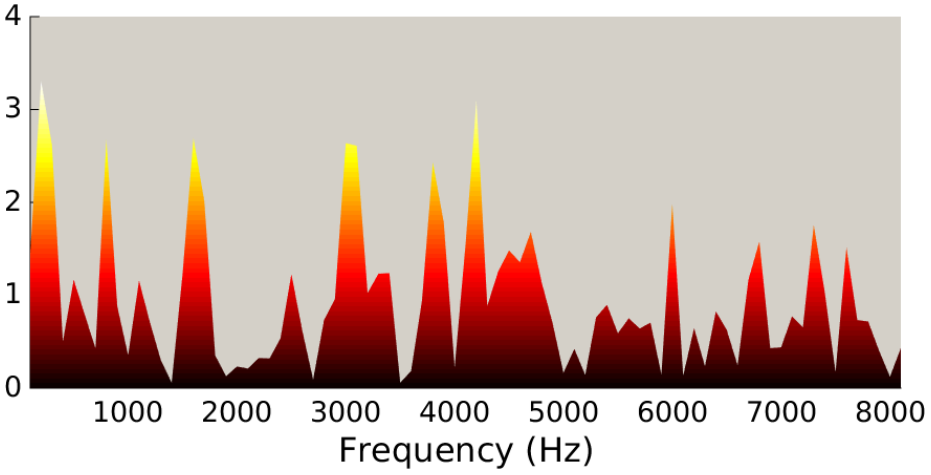


FIGURE 7.11: Frequency perspective of Figure 7.10.

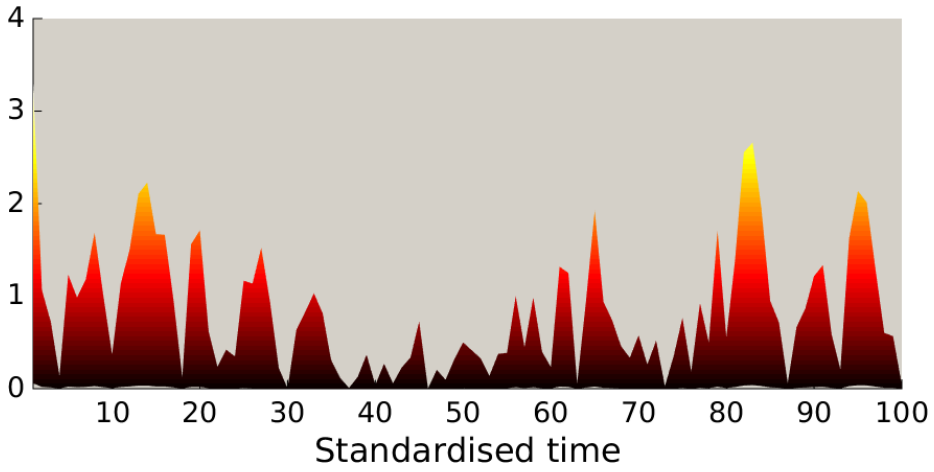


FIGURE 7.12: Time perspective of Figure 7.10.

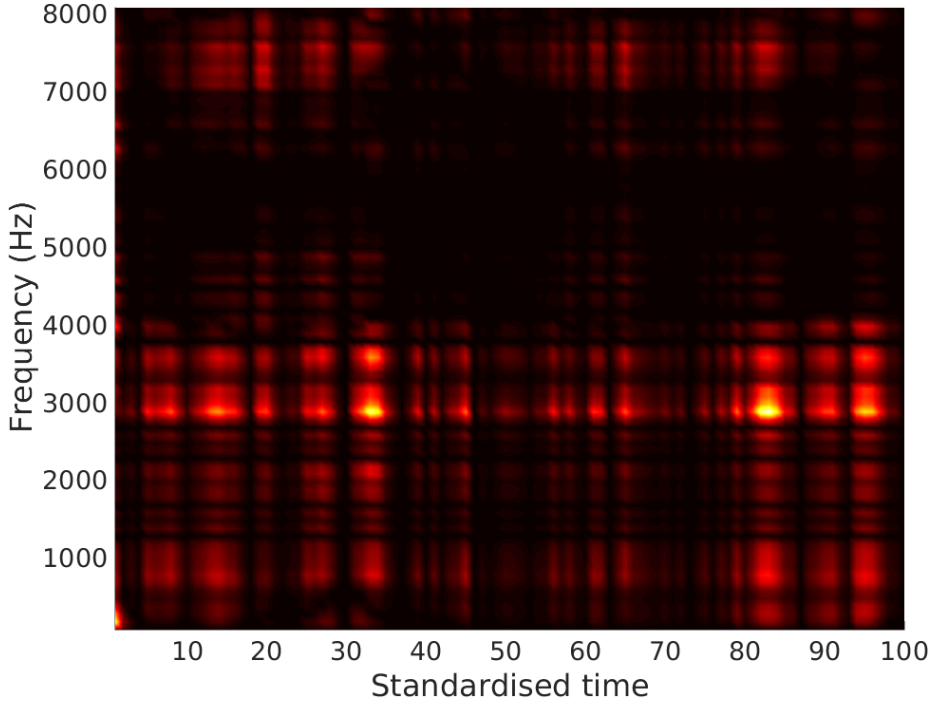


FIGURE 7.13: Absolute differences of Hadamard products for Spanish American and Portuguese in the fourth dimension.

85 and 95. In general it appears that the time aspect is less remarkable than the frequency. This could be a genuine attribute of the spectrograms or could be a result of the time standardisation and warping in the original registration of the data that could have removed the larger differences in the time dimension.

The task of interpreting components is often challenging and conclusions are often somewhat

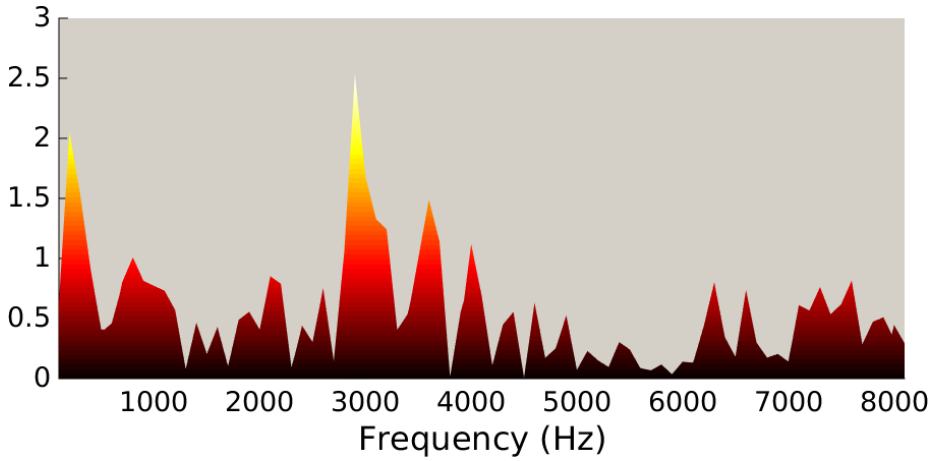


FIGURE 7.14: Frequency perspective of Figure 7.13.

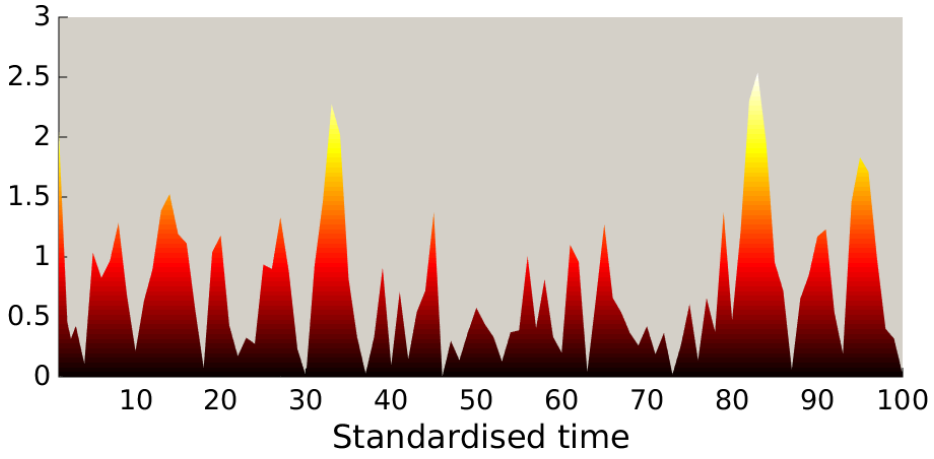


FIGURE 7.15: Time perspective of Figure 7.13.

speculative. However, the methods proposed here are useful tools that assist in assigning meanings to differences in languages in each dimension. Linguistic interpretation should be considered as preliminary in order to frame further more detailed analyses. Further investigations are required to determine whether these are general features of the languages or simply of the particular data set studied. However, in either case, this type of exploratory data analysis provides a helpful starting point before any more detailed analysis has taken place, especially if the question of tree model suitability is being considered.

7.1.4 Assessing tree-compatibility of Romance languages

We begin with a straight application of the semi-algebraic Gaussian tree constraints. We consider the positivity constraint (5.1.1) and the tripod constraints (5.4.2) for each of the $r = 9$ dimensions and report whether or not the inequalities are satisfied. Each constraint is tested

against every selection of three languages, there being 10 total choices of 3 out 5 languages. The overall constraint is reported to hold only if all 10 constraints hold as can be seen in Table 7.1. Recall from 5.4.3 that the positivity constraint is embedded in the tripod constraint so the tripod constraint only holds if the positivity constraint does.

TABLE 7.1: Results of testing point estimates against the positivity and tripod constraints for each of the selected 9 components.

	Positivity		**Tripod**	
Component	**Satisfied?**	**# satisfied**	**Satisfied?**	**# satisfied**
1	Yes	10	No	5
2	Yes	10	No	7
3	Yes	10	No	5
4	Yes	10	No	3
5	Yes	10	No	0
6	Yes	10	No	3
7	Yes	10	No	4
8	Yes	10	No	1
9	Yes	10	No	2

Applying the positivity constraint to each of the 9 component covariance matrices results in all of the components $(\boldsymbol{c}_1, \ldots, \boldsymbol{c}_9)$ adhering to the positivity constraint. Yet, none of the components satisfied the tripod inequalities. Out of the 10 combinations of languages, the number of language triples that satisfied the tripod constraint ranges from 0 to 7. At face value we might declare that a tree model is not an appropriate model for the languages. A slightly more nuanced conclusion could be that the linguistic features reflected in component 5 are less likely to adhere to a tree model than those in component 2. This would be based on the fact that component 2 satisfies the tripod constraint for 7 of the 10 induced tripod trees, whereas component 5 satisfies none. However, this analysis is rather blunt as we are still relying on point estimates. As described in Section 6 we can make use of the inverse-Wishart distribution to provide a posterior probability that the constraints are satisfied and give a more appropriate result than the crude application above. The results are reported in Table 7.2 for a simulation of 10^5 samples.

We can see that there is a range of posterior probabilities of adherence to the positivity constraint despite the previous results indicating adherence. On the other hand, the tripod constraints do not span a large range: most of the components are entirely rejected for tree-compatibility. The possibility that a latent Gaussian tree model is an appropriate model seems low for any of the components. Employing an inverse-Wishart allows us to get a much better understanding of the

TABLE 7.2: Results of simulation from inverse-Wishart posterior declaring posterior probabilities of adherence against the positivity and tripod constraints for each of the selected 9 components.

	Positivity		Tripod	
Component	**Probability**	**# satisfied**	**Probability**	**# satisfied**
1	1.000	10	0.021	5
2	0.990	10	0.002	7
3	0.850	10	0.000	5
4	0.620	10	0.000	3
5	0.242	10	0.000	0
6	0.174	10	0.000	3
7	0.228	10	0.000	4
8	0.110	10	0.000	1
9	0.153	10	0.000	2

tree-compatibility. In this instance, it appears that only the first component could reasonably be considered as a possible candidate to be modelled as a Gaussian latent tree albeit at an already low threshold of 0.02.

Let us now consider component 1 as a potential tree and take this exploratory analysis one step further by also proposing a preliminary tree. Consider the correlation matrix corresponding to the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{Y_i}}$. For each entry ρ_{jk} by making the transformation:

$$d_{jk} = -\log((\rho_{jk} + 1)/2)$$

it is possible to apply one of the many existing tree reconstruction methods in the literature. For example, considering the first component, the UPGMA algorithm [Michener and Sokal, 1957] produces a tree with topology as shown in Figure 7.16. Observe the similarity to the projection in Figure 7.3 for the first dimension, where Italian and Iberian Spanish are close in proximity.

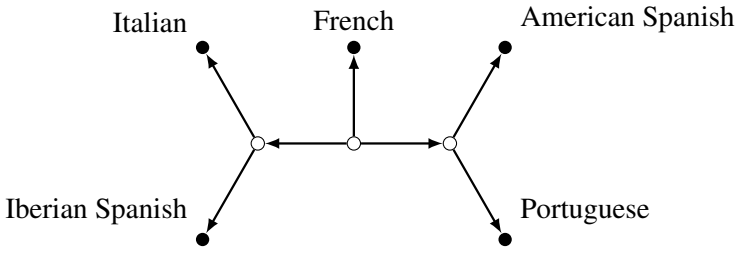


FIGURE 7.16: Topology of UPGMA generated tree for the first component.

7.1.5 Exploratory tetrad analysis example: linguistics

A more rigorous assessment of potential trees can be performed for the linguistic data set using an ETA as described in Section 6.3.

Following on from the previous analysis, we entertain the possibility that component 1 could plausibly be modelled as a latent Gaussian tree. We sample 10^5 covariance matrices from the inverse-Wishart posterior for the sample covariance $\hat{\Sigma}_1$. Considering the quintet tree (an example of which can be seen in Figure 5.2) there are 15 ways to permute the leaf labels and so 15 binary trees to test. In order to test a particular configuration of labels we require a minimal-sized set of quartets Q that defines the quintet tree (i.e. a minimal sufficient list of quartets that uniquely define the tree). Recall that Grünewald et al. [2008, Theorem 2.4] provides the minimum number of quartets required; in the case of the quintet tree this is two. Furthermore, Semple and Steel [2003, Theorem 6.8.8] provides a method for constructing minimal-sized sets of quartet trees that define binary phylogenetic trees.

For the dimension of interest, using the sampling distributions given in Section 6.2 and the test statistic (6.3.1) with two degrees of freedom, a p-value can be calculated for each of the 15 non-isomorphic permutations of languages to leaves. To retain an overall significance rate of less than α a Bonferroni correction [Dunn, 1961] is applied such that the significance level is set at $\alpha/15$. For example, correcting to retain overall 0.05 level means that for each test the significance level is $0.00\overline{3}$. Running the ETA we find that none of the 15 language permutations are rejected at the 0.05 level (nor at a more stringent 0.01 level) as reported in Table 7.3.

None of the trees are rejected and there are a number of high and similar p-values. If we were to stop our analysis here we would select tree 2 followed by tree 8 given these have the highest p-values. However, we have not exhausted the T_5 constraints and so we can utilise the final set of semi-algebraic constraints: the tetrad constraints. Once again sampling from the relevant inverse-Wishart posterior 10^5 times, we produce estimates of the probability of tree-compatibility using the full suite of semi-algebraic constraints, i.e. positivity, tripod and tetrad (see 6.1.3). The resulting posterior probabilities of tree-compatibility for the four remaining trees is given in Table 7.4.

TABLE 7.3: Results of test for vanishing tetrads for the first component of the linguistic data set at the 0.05 and 0.01 significance levels. The coding for the trees in column 2 is: 1 = French, 2 = Italian, 3 = Portuguese, 4 = American Spanish, 5 = Iberian Spanish

Tree #	Tree	p-value	Outcome
1	$12\|3\|45$	0.476	Do not reject
2	$12\|4\|35$	0.525	Do not reject
3	$12\|5\|34$	0.491	Do not reject
4	$13\|2\|45$	0.483	Do not reject
5	$13\|4\|25$	0.478	Do not reject
6	$13\|5\|24$	0.387	Do not reject
7	$14\|2\|35$	0.311	Do not reject
8	$14\|3\|25$	0.520	Do not reject
9	$14\|5\|23$	0.376	Do not reject
10	$15\|2\|34$	0.341	Do not reject
11	$15\|3\|24$	0.363	Do not reject
12	$15\|4\|23$	0.389	Do not reject
13	$23\|1\|45$	0.398	Do not reject
14	$24\|1\|35$	0.285	Do not reject
15	$25\|1\|34$	0.451	Do not reject

TABLE 7.4: Posterior probabilities of tree-compatibility using all semi-algebraic constraints for remaining four trees relating to component 1

Tree #	Tree	Probability
1	$12\|3\|45$	0.006
2	$12\|4\|35$	0.012
3	$12\|5\|34$	0.002
4	$13\|2\|45$	0.000
5	$13\|4\|25$	0.000
6	$13\|5\|24$	0.000
7	$14\|2\|35$	0.000
8	$14\|3\|25$	0.000
9	$14\|5\|23$	0.000
10	$15\|2\|34$	0.000
11	$15\|3\|24$	0.000
12	$15\|4\|23$	0.000
13	$23\|1\|45$	0.000
14	$24\|1\|35$	0.000
15	$25\|1\|34$	0.000

The highest probability tree is Tree 2 at 0.012 which is displayed in Figure 7.17. This seems quite low but would not be rejected at the Bonferroni-adjusted 0.01 level. On the other hand, the previously second highest scoring permutation Tree 8 would be rejected at the 0.01 level. So having performed the full set of algebraic and semi-algebraic constraint testing the conclusion of the ETA is that Tree 2 is a plausible GLTM that provides an appropriate description of the conditional independence relationships given the data.

For example, from this particular analysis we could hypothesise that the differences in vowel sounds of Portuguese and French evolved independently conditional on the common ancestor of Spanish and Italian. In combination with expert judgement, such statements can provide a good starting points for further analysis of these features in relation to a specified tree.

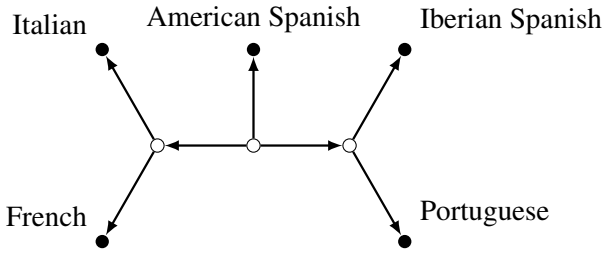


FIGURE 7.17: Topology of highest probability quintet tree for the first component of the Romance data set.

7.1.6 Alternative analyses

We now repeat some of the same methodology but on the data set that has undergone a copula transformation and a reduced data set that has excluded selected observations. This allows us to consider the effect of adopting other reasonable assumptions or adjustments and seeing whether the results of the analyses differ.

7.1.6.1 Assessment of Gaussianity

The use of CVA is based upon the assumption of Gaussianity, as are the derivations of the Gaussian tree constraints. While it can be argued that use of these techniques is valid on the basis that the highest moment of interest is second-order, it may sometimes be preferable to make an adjustment for non-Gaussian data. Given the use of CVA, here we describe a method for assessing whether multivariate Gaussianity is inherent in the data set. There are a number of tests for multivariate normality with differing properties (e.g. type I and type II error rates [Mecklin and Mundfrom, 2005]). Here we chose to use Royston's H-test [Royston, 1983] as it is regarded as having good type I error control and amongst the best power for small sample sizes. These properties are shared with the Henze-Zirkler procedure [Henze and Zirkler, 1990] which is also rated as robust test even if the data has strong correlations [Mecklin and Mundfrom, 2005]. Royston's H-test is a goodness of fit test and an extension of the Shapiro-Wilk test

for univariate normality [Royston, 1982]. We can employ Royston's H-test via the `MATLAB` package `Roystest` [Trujillo-Ortiz et al., 2007] on the 219 observations treating each of the 8100 spectrogram time-frequency points as variables. At the 0.01 level, Gaussianity of the data set is rejected.

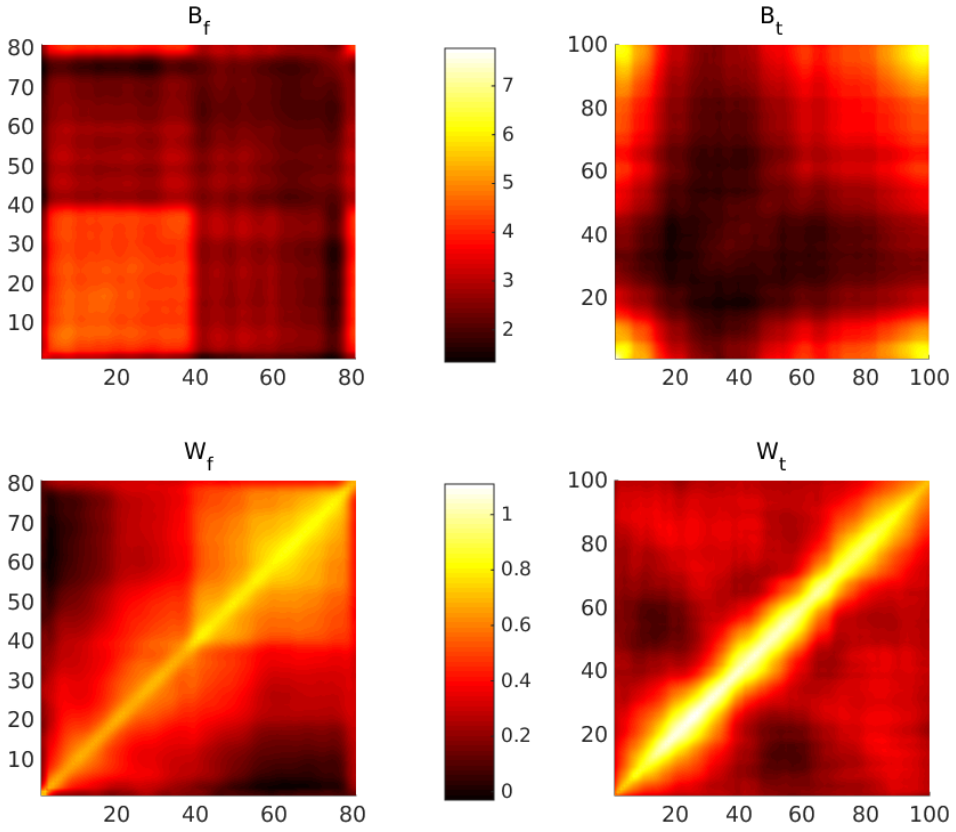


FIGURE 7.18: Sample between-language and within-language covariances of speech data for frequency and time directions.

One approach to addressing the lack of Gaussianity in a data set is to consider the marginal distributions of the data. It is well known that a necessary condition for multivariate Gaussianity is that the univariate marginals must also be normally distributed (see Timm [2007, Chapter 3] for example). While marginal Gaussianity is not sufficient it can aid multivariate Gaussianity. We perform a copula transform on each of the variables using the function `copula.trans` from the `R` package `regpro` [Klemela, 2013], which makes each of the marginal distributions approximately Gaussian. Rerunning Royston's H-test on the transformed data no longer rejects Gaussianity. This indicates that the marginal Gaussianity transformation was in this instance enough to bring the joint density within the limits of normality. Here we compare the results of the analysis when using the transformed data set.

In Figure 7.18 we can see that the block structure in the frequency covariances is even clearer than in the previous analysis (see Figure 7.1). This is considered in the subsequent analysis in Section 7.1.6.2. Otherwise the covariances appear similar albeit scaled differently by the nature of the copula transform. We proceed as before, considering a dimension reduction. In contrast to the previous analysis, the first component only accounts for 12% of the between- to within-language variation. This could mean that the separability assumption is impacting of the efficiency or it could simply be that no single component truly can account for a large proportion of the variance. Considering all the components with greater than 0.75% explanatory power (i.e. the first 15 dimensions) we proceed with testing the semi-algebraic constraints. Simulating 10^5 times from the inverse-Wishart posterior we report the results in Table 7.5.

TABLE 7.5: Results of simulation from inverse-Wishart posterior for first 15 components of copula transformed data.

	Positivity		**Tripod**	
Component	**Probability**	**Satisfied?**	**Probability**	**Satisfied?**
1	0.999	Yes	0.000	No
2	0.825	Yes	0.000	No
3	0.129	Yes	0.000	No
4	0.397	Yes	0.000	No
5	0.363	Yes	0.000	No
6	0.142	Yes	0.000	No
7	0.146	No	0.000	No
8	0.249	Yes	0.000	No
9	0.111	Yes	0.000	No
10	0.409	Yes	0.000	No
11	0.166	Yes	0.000	No
12	0.144	Yes	0.000	No
13	0.123	Yes	0.000	No
14	0.111	Yes	0.000	No
15	0.000	Yes	0.000	No

We can observe that the positivity constraint once again has a range of posterior probabilities associated with it, whereas the tripod constraint is very unlikely to hold according to the simulation. Although the components are likely to reflect different combinations of features so that we cannot automatically compare any two dimensions, it is interesting to note that the eventual similarly low posterior probabilities are found with both analyses.

7.1.6.2 Excluding low sample rate observations

Recall Figure 7.1 and Figure 7.18 where it was noted that there appeared to be some artefact in the frequency covariance plots in the form of a block structure. After closer investigation, it was noted that some of the recordings (from which the spectrograms were created) were recorded at a lower sampling rate. By the Nyquist-Shannon sampling theorem [Sujatha, 2010, Chapter 6] in order to record frequencies at level αHz, the sampling rate needs to be a minimum of 2αHz. Thus when we are considering frequencies of 8000Hz a sampling rate of 16000Hz is required to capture the full range of frequencies.

For example, we compare two female French speakers of the word "cinq" and denote them speakers A and B. Speaker A has been recorded at a sampling rate of 16000Hz whereas speaker B has been recorded at a sampling rate of 11025Hz. We note that in the spectrograms there is a difference in the higher frequencies with very little power beyond 5500Hz for speaker B.

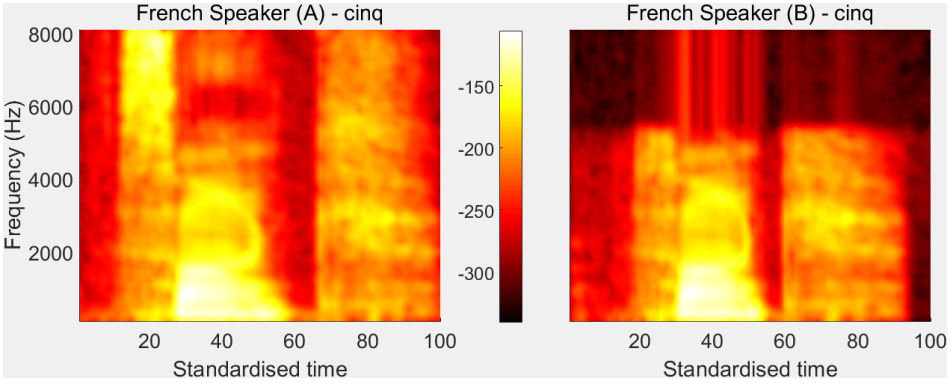


FIGURE 7.19: Spectrograms of two female French speakers saying the word "cinq".

This feature is illustrated even more clearly through the use of frequency analysis. The result of fast Fourier transforms (FFT) using Hann (or Hanning) windows of size 512 samples [Walker, 1996, Chapter 4] are given in Figure 7.20 and Figure 7.21. There is a clear drop-off in power for the latter plot. Note the difference in axes ranges and scales.

We identify 38 recordings that are lacking in the higher frequencies. To assess the robustness of the original conclusions we re-run the analysis excluding these observations. Reviewing the between- and within-language covariance we can see that there is much less of a block structure particularly for the frequency direction (Figure 7.22).

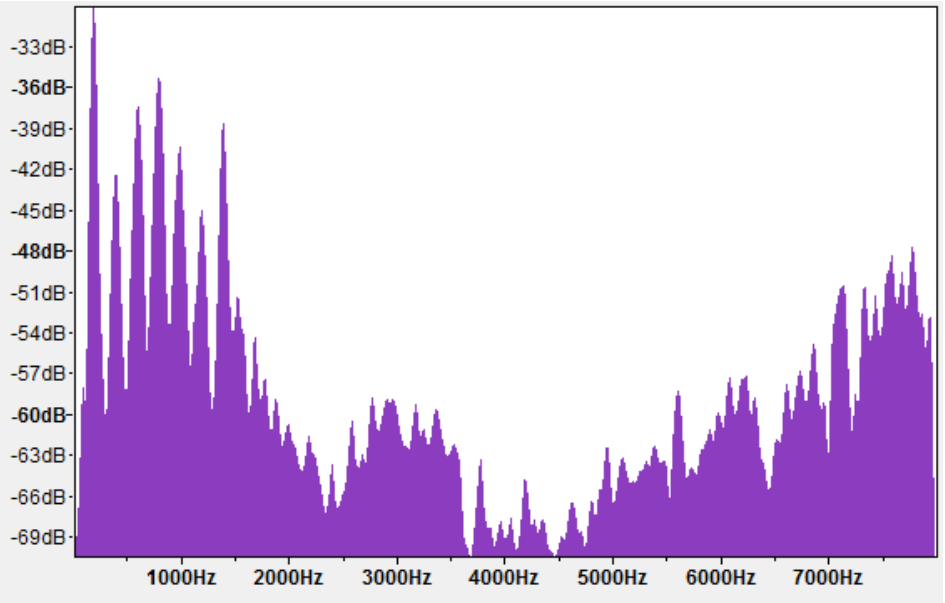


FIGURE 7.20: Frequency analysis for French speaker A.

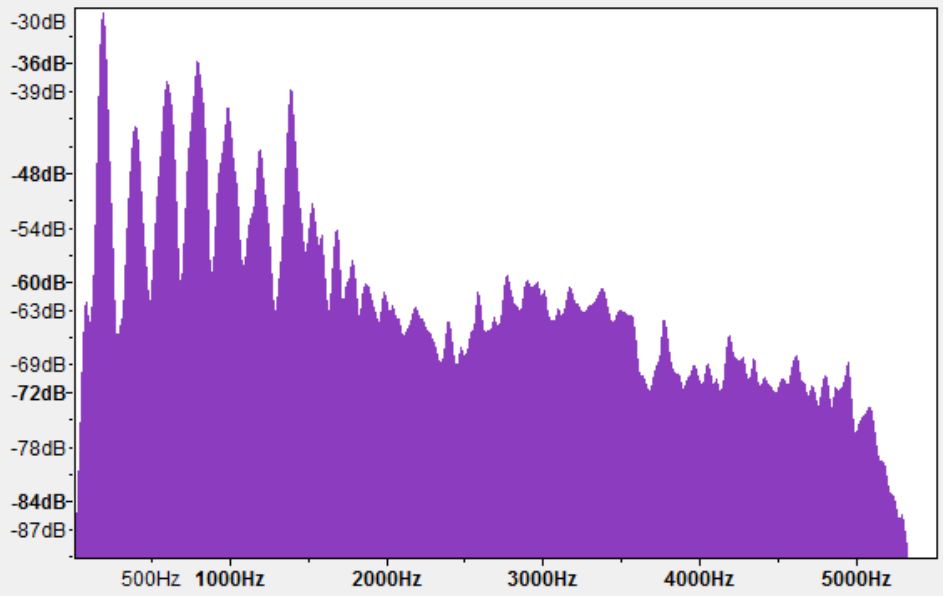


FIGURE 7.21: Frequency analysis for French speaker B.

Proceeding as before, we select a 0.1% cut-off for the explanatory power and project independently into 10 dimensions. Together these components account for approximately 97.5% of the total between-to-within variation. Testing each of the 10 covariance matrices with the first set of Gaussian tree constraints, we find the same overall conclusion as with the analysis of the full data set in Section 7.1.6.1 where all the components of interest are rejected for tree-compatibility. The results are reported in Table 7.6. Comparing the results of the point estimate positivity test (column 3) with those in Table 7.2 and Table 7.5 we can see that although all

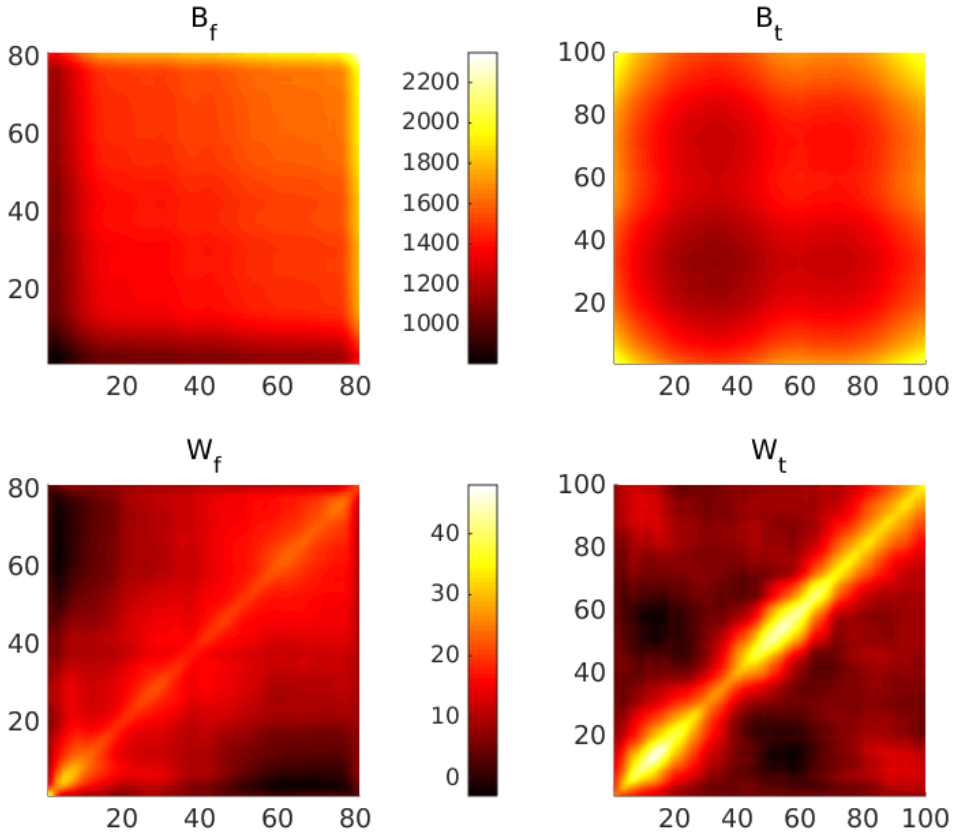


FIGURE 7.22: Sample between-language and within-language covariances of speech data for frequency and time directions.

TABLE 7.6: Results of simulation from inverse-Wishart posterior for first 10 components of copula transformed reduced data.

	Positivity		Tripod	
Component	**Probability**	**Satisfied?**	**Probability**	**Satisfied?**
1	0.918	Yes	0.000	No
2	0.923	No	0.000	No
3	0.604	No	0.000	No
4	0.891	No	0.000	No
5	0.157	No	0.000	No
6	0.623	No	0.000	No
7	0.533	No	0.000	No
8	0.303	No	0.000	No
9	0.209	No	0.000	No
10	0.118	No	0.000	No

but the first component fails the positivity test here, the range of posterior probabilities (column 2) is comparable. This indicates the limited usefulness of Gaussian tree constraints as binary outcome tests.

In the spirit of Section 7.1.6.1, we can perform a copula transform on the reduced data set. Performing the separable-CVA leads to the selection of the first 10 components being retained

for projection. The subsequent analyses produce largely similar results to the previous one, with the rejection of tree-compatibility after the tripod inverse-Wishart constraint testing. This points to there being some robustness in the conclusions of these alternative analyses with respect to decisions made regarding transformations and removal of particular samples.

7.2 Yeast data growth curves

7.2.1 Confirmatory tetrad analysis example: biology

We consider a data set consisting of a set growth curves for seven yeast species. These species have been previously studied in Marcet-Houben and Gabaldón [2009] and there is a purported phylogeny for the named species (see Figure 7.23) in the form of a tree. However, Libkind et al. [2011] conclude that yeast species *Saccharomyces bayanus* is a hybrid involving *Saccharomyces cerevisiae* which would violate the tree assumption. In Warringer et al. [2011] and Liti et al. [2009] traits relating to growth curves have been used to investigate relationships between physiological and genetic structure. Positive correlation between growth-related phenotypic variation and genotypic phylogenetic relationships was reported i.e. some aspects of the growth curves reflected the genetic description of the relationships between yeast species. With a range of phylogenetic results in the literature, it is of interest to carry out a CTA to assess whether the proposed tree structure given in Marcet-Houben and Gabaldón [2009] is reflected in the growth data. This analysis is not used to determine which of the genetic analyses is correct but rather to investigate the plausibility or otherwise of aspects of the conditional independence relationships for growth curves being consistent with the phylogenetic tree presented. We carry out a CTA to assess whether the proposed tree structure in Marcet-Houben and Gabaldón [2009] is reflected in any aspects of the growth data.

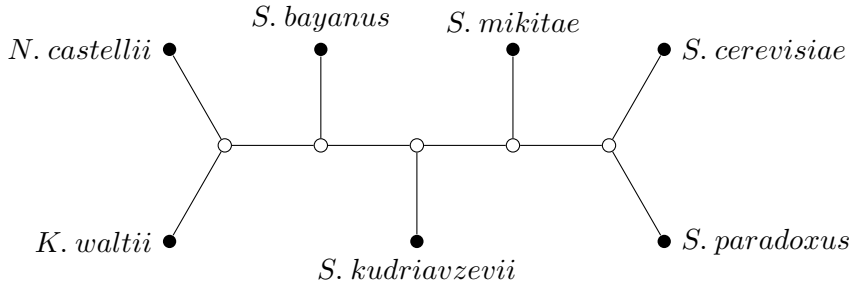


FIGURE 7.23: Septet tree T_7 of yeast species as per Marcet-Houben and Gabaldón [2009].

7.2.1.1 Processing the data set

We consider data consisting of a set of growth curves for seven yeast species: *Kluyveromyces waltii*, *Saccharomyces bayanus*, *Saccharomyces mikatae*, *Saccharomyces kudriavzevii*, *Saccharomyces cerevisiae*, *Saccharomyces paradoxus* and *Naumovozyma castellii* (synonym: *Saccharomyces castellii*). Each was observed in the same 96 environments, each species with at least two replicates. The growth is recorded 288 times approximately every six minutes over a period of just over 26 hours. It is safe to assume that the underlying process is producing functional data as is assumed for other growth data (e.g. Ramsay and Silverman [2005], Gervini and Carter [2014]). This assumption appears reasonable when studying the data, for some examples see some typical linearly interpolated growth plots in Figure 7.24. For this study, we denote the underlying functions as $x_{l,e,r}(t)$ for $t \in [0, t_{max}]$ where $l \in \{1, \ldots, 7\}$ indicates the species, $e \in \{1, \ldots, 96\}$ indicates the environment and $r \in \mathbb{N}$ indicates the replication (ranges between 1 and 4 with a total of 19 replicates cross all species). Theoretically t_{max} could be ∞ but in practice observations will stop at a finite time, in this case approximately 26 hours. The observed data $\boldsymbol{x}_{l,e,r}$ take the form of vectors of length 288 with growth level recorded at each of the times $\boldsymbol{t}_{l,r} = (t_{l,r,1}, \ldots, t_{l,r,288})$ (dependent on the species and replicate). The only exception is the first replicate for *S. Paradoxus* that is missing 86 of the growth recordings (32 to 117) across all of the environments. We address this missing data subsequently. The data set can be expressed as an array of $19 \times 96 \times 288$.

Having made the reasonable assumption that the data is functional, we decide to fit a spline to the data such that the entire growth curve is defined rather than just select points. Of course this will only be an approximation to the true unattainable growth curve. However, the spline should form a suitable representation given the data if the spline is a high enough order, a sufficient

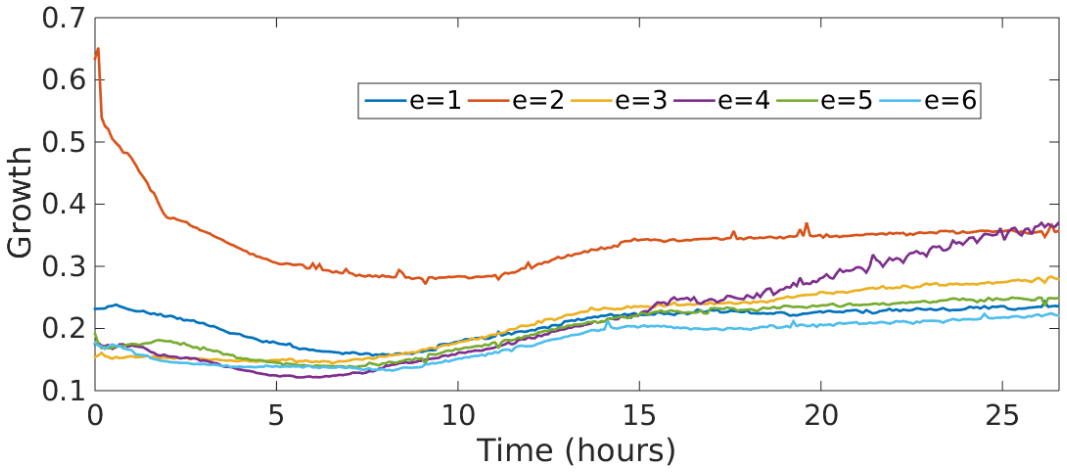


FIGURE 7.24: Examples of growth curves for *S. Bayanus*, replicate $r = 1$, environments $e = 1, \ldots, 6$.

number of knots (locations where the spline pieces are joined) are selected, and appropriate smoothing parameters are chosen. Fitting a spline then allows us to evaluate over all the growth curves at consistent values, that is $\boldsymbol{t}_{l,r} = \boldsymbol{t} \forall l, r$. It should also help reduce observational noise by smoothing across observed data. We choose to use a cubic spline (i.e. a maximum of order 3 polynomials between each knot) with smoothing incorporated. We employ the `MATLAB` function `csaps` to carry out the spline fitting. The spline $S_{l,e,r}$ is constructed by minimising:

$$p_s \sum_{j=1}^{288} |x_{l,e,r,j} - S_{l,e,r}(t_{l,r,j})|^2 + (1 - p_s) \int_{t_{l,r,1}}^{t_{l,r,288}} |D^2 S_{l,e,r}(u)|^2 \mathrm{d}u$$

where D^2 is the second derivative of $S_{l,e,r}$ and p_s is the smoothing parameter which is set to be constant across all species and replicates. If $p_s = 0$ then the $S_{l,e,r}$ becomes the line of best fit through the observed data with respect to squared distance. This is rarely appropriate and undermines the choice of cubic splines. If $p_s = 1$ then $S_{l,e,r}$ then we get the regular cubic spline interpolant. This is usually closer to the desired result, though if the data is noisy then this condition can be too prescriptive. Through some experimentation, a value of $p_s = 0.9$ appeared to provide a good compromise in terms of smoothness of the splines and also following the data sufficiently closely. In Figure 7.25 we can view a comparison of $p_s = 0.9, 0, 1$ for a magnified section of *S. Mikitae*, first replicate, 14th environment ($l = 1, e = 14, r = 1$). This illustrates that $p = 0$ is clearly inappropriate as for this latter section the fitted spline does not even go through the data. We can also see that for $p = 1$ the spline has tracked the smallest bumps in the observations. Given that it is questionable that this jitter-like detail is a feature of the underlying

true growth curves, it is probable that $p = 1$ is causing the spline to over-fit the data.

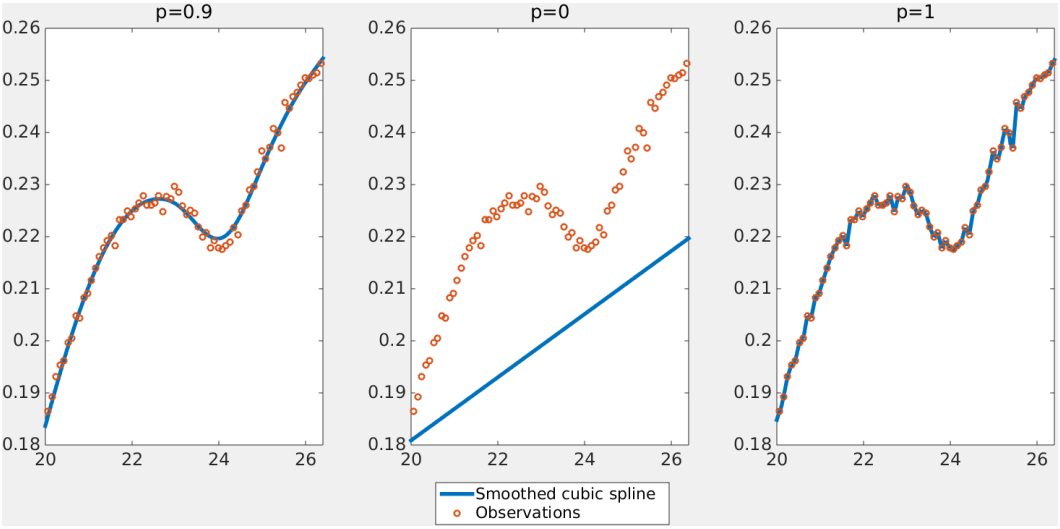


FIGURE 7.25: Comparison of smoothing parameters $p = 0.9, 0, 1$

The spline fitting copes well with missing data simply smoothing over the gap. We are fortunate that the growth curves appear particularly smooth outside of the missing data range. Also, we have two other replicates for the species *S. Mikitae* with which we can compare the fitted spline to check for any dubious spline sections. Figure 7.26 shows the fitted spline for environment 1 for all three replicates of *S. Mikitae*. For $r = 1$ the interpolation seems sensible and this is backed-up by the other replicates not indicating any unusual features in the growth curves. This figure is typical when considering the other environments. Another approach to missing data was tested whereby the splines were fitted for the other replicates and then an average was taken for the missing data range. This was then adjusted by a constant to shift the curve up or down to align with the splines for the observed ranges in replicate 1. However, this did not work well universally and the region of missing data often stood out when the curve shift was too simplistic. Therefore, although we cannot know what the missing data is we can at least make it consistent with our assumption that the growth is smooth and so the spline interpolation method is preferred.

For each of the splines $S_{l,e,r}$ we can now evaluate at consistent values across species, replicates and environments. Ascertaining the minimum of the maximum of the time ranges $t_{min} = \min_{l,r}\{t_{l,r,288}\}$ allows us to know the range over which all observations are defined and can thus be evaluated over. We can then subdivide this range to evaluate the fitted splines at comparable points. These need not be equally spaced points of evaluation. If regions of the growth curves

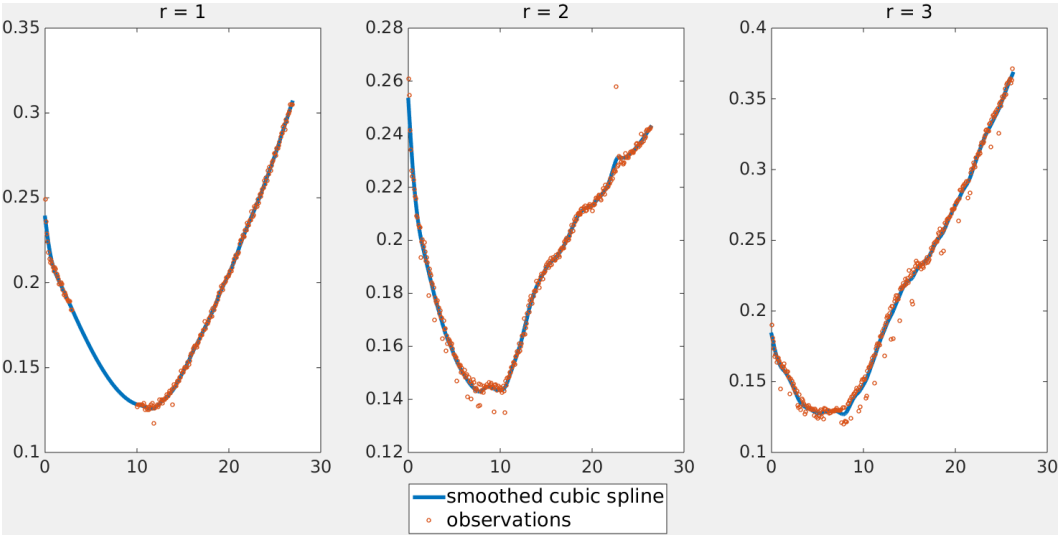


FIGURE 7.26: An example of the smooth cubic spline interpolating the missing data for replicate $r = 1$ and comparisons with $r = 2$ and $r = 3$ that have no missing observations.

are known to consistently fluctuate across observations then more evaluation points can be focused in these ranges to capture more of the detail. However, the growth curves tend to vary in shape but also tend not to have intricate details. Thus we decide to evaluate at regular intervals. Experimenting with differing intervals we opt to evaluate each spline at 81 points, roughly every 20 minutes given that $t_{min} \approx 26.25$ in hours. We then end up with an array of $19 \times 96 \times 81$ where the entries in the columns of the third array dimension are now comparable and hopefully with reduced noise. We denote a specific selection of species, environment and replicate by the 96×81 matrix $\boldsymbol{X}_{l,e,r}$.

Means are calculated for each species and environment pair by averaging across replicates:

$$\bar{\boldsymbol{X}}_{l,e} = \frac{1}{r_l} \sum_{r=1}^{r_l} \boldsymbol{X}_{l,e,r}$$

where r_l is the number of replicates for species l. Similarly for just the mean environment matrix:

$$\bar{\boldsymbol{X}}_{e} = \frac{1}{19} \sum_{l=1}^{7} r_l \bar{\boldsymbol{X}}_{l,e}.$$

We then standardise to remove mean environmental effects:

$$\tilde{\boldsymbol{X}}_{l} = \bar{\boldsymbol{X}}_{l,e} - \bar{\boldsymbol{X}}_{e}.$$

The next stage is to centre the data at zero and perform a PCA (see Section 3.5.1) across species and environments. As the mean environmental effects have been removed we consider all 672 observations together for the PCA. The PCA identifies the core variability of the growth curves. The first four dimensions are found to account for over 99% of variability. For each of the mean species projections in these dimensions, the sample covariance matrix is constructed.

To investigate the meaning of each of the dimensions we can plot the coefficients (effectively normalised eigenvectors) relating to the first four components. These are shown in Figure 7.27. We can see that the eigenvector relating to the first component mostly accounts for variation in the latter half of growth curves. The second eigenvector accounts for variation in the middle of the growth curve with opposite sign from about 18 hours onwards. Coefficients for dimension 3 indicate that between 10 and 20 hours the sign is negative whereas outside of this range the sign is positive. Finally, the fourth dimension eigenvector is approximately sinusoidal crossing $y = 0$ at around 5 hours, 13 hours and 22 hours. Note that in this way, studying the coefficient plot helps us assign a broad interpretation of each principal component.

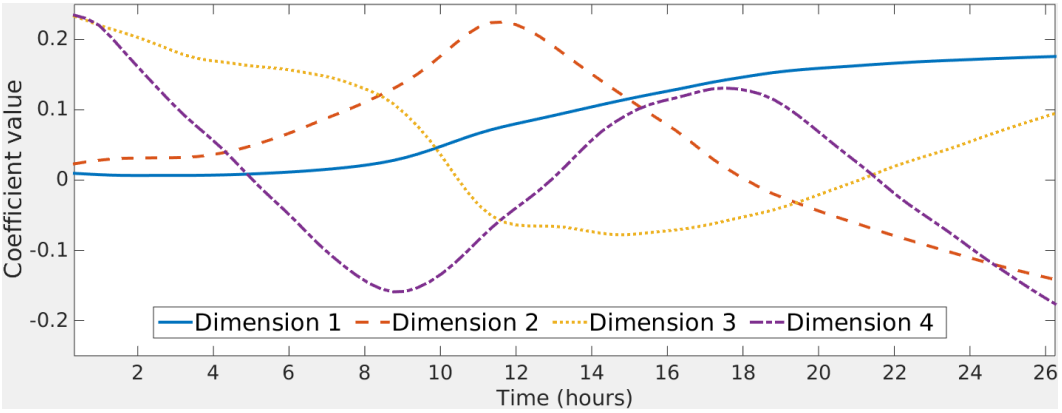


FIGURE 7.27: Interpolated plots of the coefficients relating to the first four principal components.

If we require more detail we can plot the Hadamard product of a coefficient and the centred data. For example, consider the mean observation for each species in the third dimension. Figure 7.28 indicates the mean contribution of each part of the growth curve to the overall score for the mean of each species. The scores for dimension 3 for the seven mean species ranges from -0.07 for *S. Mikitae* to 0.10 for *S. paradoxus*. Yet the plot indicates that for most of the time period neither of these species is the maximum or minimum. Considering *K. Waltii* we can see that from about 10 hours onwards this curve either bounds above or below the rest of the species' curves. However,

given the overall score is -0.05, it is clear that these two ranges are in part cancelling each other out. In conjunction with the third dimension coefficient curve from Figure 7.27 we can ascertain that on average the zero centred *K. Waltii* growth curves were positively values between 10 and 21 hours and negatively valued beyond 21 hours. This is deduced by considering the sign of the third eigenvector in comparison to the sign of the mean contribution. Together these two plots can be used to explore the meaning of any particular dimension of projection.

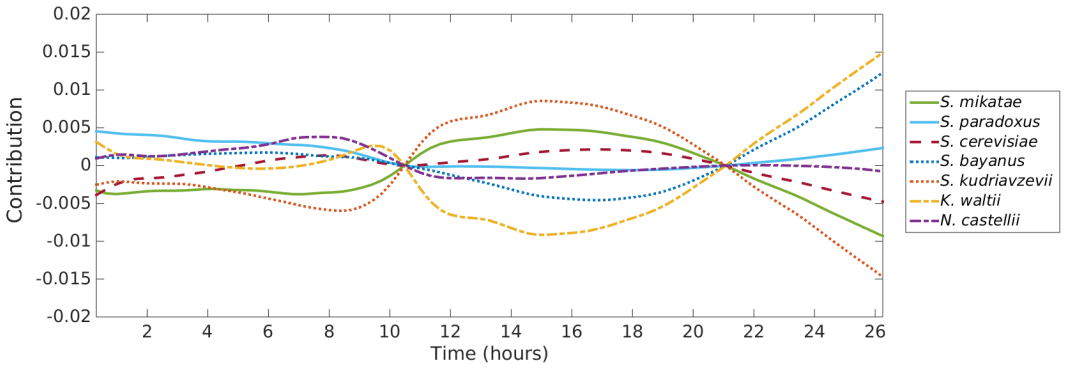


FIGURE 7.28: Interpolated plots of the Hadamard product of the coefficients relating to the third principal components and the mean of each species.

7.2.1.2 Performing a CTA

Before implementing a CTA (see Section 6.3) it is good practice to assess the positivity and tripod constraints. These assessments can often be easier to undertake and if they are found to reject tree-compatibility then there is no need to proceed with the CTA. Using the inverse-Wishart approach specified in (6.1.2) it seems that none of the first four dimensions satisfy both the constraints. The highest posterior probability is 0.008 which is probably low enough to dismiss tree-compatibility. Making that judgement, it would be pointless to proceed with CTA. However, the analysis does not have to stop here as we can also consider subsets of the seven species to assess them for tree-compatibility with respect to the tripod constraints (and by implication the positivity constraints). We assess subsets of 5 and 6 species, thus if just one or two of species are responsible for violating the tree inequalities we will be able to tell.

We find that there are a few subsets of five species for which the posterior tree-compatibility probability increases to a level worth investigating further. The most notable of these is the exclusion of *N. castelli* and *S. paradoxus* which reports posterior probabilities for the first four

components as 0.302, 0.021, 0.184, and 0.029 respectively. The cut-off level is as before subjective, but here we decide to consider the first and third components further via CTA. The results indicate that the first dimension is not compatible with a tree at the 0.01 level, whereas the third dimension does not reject the tree hypothesis having a p-value of 0.337. As discussed in Section 6.3, the selection of quartets for CTA are not unique, but often there is a reason to prefer one choice over another. Here we selected by using the quartets that were informed by the most data observations (i.e. the most replicates). As a robustness check we consider the other combinations and find that the same conclusion are reached at the $\alpha = 0.01$ level. We can now proceed with a final evaluation of the tree shown in Figure 7.29 (induced by the removal of *N. castelli* and *S. paradoxus*). We consider the tetrad semi-algebraic constraints for the third dimension given the topology of the tree.

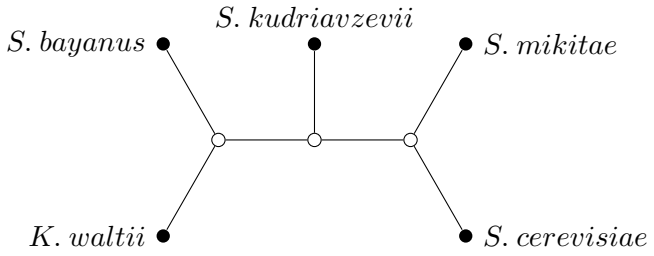


FIGURE 7.29: Quintet tree T_5 of yeast species as per Marcet-Houben and Gabaldón [2009] with *N. castelli* and *S. paradoxus* removed.

By simulating 10^5 samples using the inverse-Wishart distribution, the CTA for T_5-compatibility (see Figure 7.29) gives p-values of 0.721 and 0.955 for the first and third components respectively. To double check these results we repeated the test using the bootstrapping strategy outlined in Bollen and Stine [1992], which is more robust to small sample sizes where the distribution of the test statistics is not known. This technique transforms the data set using the maximum likelihood estimate of the covariance under the null model T_5. Bootstrapped samples are taken and the distribution of the test statistic under the null is then estimated. The results are very similar with p-values of 0.729 and 0.921 respectively. The CTA and inverse-Wishart simulation results both give upper bounds on T_5-compatibility, but on balance we conclude that the first and third components are T_5-compatible. Therefore, the class of GLTMs does appear suitable for modelling the subset of the seven species for some aspects of these yeast species' growth curves. However, for features relating to components 2 and 4, there is some evidence to support

the exploration of a wider model class that would allow for the hybrid hypothesis described in Libkind et al. [2011].

7.3 Discussion

These applications illustrate the implementation of the full set of latent Gaussian tree constraints in practice. They demonstrate that particular constraints are useful for different scenarios or aims. The semi-algebraic constraints tend to perform well as exploratory steps to check the appropriateness of a tree model search whereas often the algebraic constraints are performed as a follow-up to the semi-algebraic constraints. Primarily this is because the associated algebraic methods are more time-consuming to set-up than the semi-algebraic tools and thus this implementation order is sensible. The range of tools adds to the versatility and overall usefulness of having identified the complete correlation space of GLTMs.

The complete semi-algebraic structure of the correlation space has not been utilised elsewhere for assessing tree-compatibility of data (only the positivity constraint has been used previously, see Shiers et al. [2014]). Incorporating a prior (such as the inverse-Wishart) and sampling from the posterior distribution allows for probabilistic conclusions about the model. It provides a more nuanced answer than a simple assessment of inequalities via the plugging in of covariance point estimates, and allows two or more incompatible but plausible trees to be considered relative to one another. Whilst the tetrad constraints have been known previously, their use in combination with the semi-algebraic constraints is novel and allows for a more critical analysis of proposed trees.

In the linguistic application, we have demonstrated a method for isolating and identifying distinguishing aspects of variability in acoustic functional data which may be of evolutionary interest. It shows that it is possible to identify prominent features which render particular components effective for distinguishing the language groups. However, it also highlights the challenge of precise physical interpretation of particular components, a task which appears notably more complex due to having both a time and a frequency dimension. It would be of interest to express

these differences back in the sound domain, although given the difficulties in inverting spectrograms to sound, this is not a trivial task. However, it is the subject of ongoing work, including experiments with other parametric acoustic representations that are more easily inverted.

In the yeast species example, we illustrated how given a candidate tree we can give a thorough assessment of tree-compatibility via the semi-algebraic tripod structure but more specifically using CTA and the tetrad inequalities. The implementation of CTA was noticeably more straightforward compared to ETA once the moment estimators had been constructed as there was only one candidate tree to consider. The interpretation of the principal components was less challenging than the canonical components in the linguistic example. This was in part due to PCA optimisation having a single aim whereas CVA optimisation is relative regarding between- and within-group variation. Furthermore, in these particular examples the 2-dimensional nature of the linguistic data was inherently more challenging than the 1-dimensional growth curves.

An important practical consideration is the scalability of these methods. Techniques employing the semi-algebraic constraints can be adapted to larger number of variables reasonably well. In comparison, ETA does not scale well as the number of variables increases the number of permutations of leaves and the number of trees grows exponentially. Thus ETA is used most appropriately on small data sets or subsets of larger models where extra resolution is required. CTA is more manageable as although it can have a significant initial cost for larger trees, once established implementation is generally viable. The limiting factor for CTA is likely to be the computational costs of constructing covariances of quartets. This can be addressed in part through smart programming that makes use of symmetries and sparseness of the matrices.

A further consideration with tetrad analyses is that the selection of the quartet set used for testing is not unique. Identifying the minimal defining sets can be challenging, but moreover, there may not always be an obvious reason for selecting one minimal defining quartet set over another. For reasonably sized problems structural equation modelling software can be used to automate the process of finding linearly dependent tetrads, and consequently minimal defining quartet sets. In such cases it is recommended that at least a some of these alternative sets are randomly selected to assess the sensitivity and robustness of the results.

Overall, these tree constraint methods are most effective as supplementary tools that can provide an opportunities to reduce model search time and to critique the output of model searches. The

latter point is of particular relevance as although model searches employ a range of selection criteria and weightings they may often induce bias, for example through greedy algorithms. A CTA is based on the unavoidable underlying structure of the space and thus adds a final probabilistic check on proposed models.

Chapter 8

Discussion

One of the main contributions of this thesis is the derivation of the complete set of Gaussian tree constraints. Furthermore, the proposed associated methodology turns these theoretical results into practically useful tools. This is demonstrated through an application to a linguistic acoustic data set originating from speakers of certain Romance languages. Apart from a new suite of tools for assessing tree-compatibility, the example is particularly interesting due to the nature of the data set. The use of acoustic functional data offers a fresh approach to considering the relationships between languages, potentially offering new insights. The nature of this approach means that an extensive range of existing knowledge and tools has been drawn upon. But furthermore, it has been necessary to develop the novel and powerful tools separable-CFA and separable-CVA in order to overcome the common problem of the number of variables exceeding the number of observations. The linguistic example has been supplemented by phylogenetic examples in order to demonstrate the versatility of the techniques and illustrate how the results can easily be carried across to other domains. The approach taken in this thesis is just one way of analysing functional data with hypothesised latent structure. The methods used were governed by what was currently feasible, what could realistically be developed and of course the inferential needs of the problem. These are clearly not the only ways to analyse such problems and had another application been the focus then alternative methods may have been selected.

Looking forward to potential directions of research, an obvious extension of this work is to the multivariate Gaussian domain. However, obtaining the complete description of the space is non-trivial and ongoing work. Furthermore, some constraints in higher dimensions are more difficult

to interpret from a practical point of view as they can involve restrictions across dimensions and across variables simultaneously. Alternatively, a generalisation from Gaussian to other elliptical distributions would be very useful if possible and would relax the level of distributional assumptions that are required.

The linguistic application was the driving force for this thesis and in the process of finding an appropriate way of analysing the data, a whole framework of methodology was constructed. Yet clearly the linguistic application was merely the first venture using this approach and there would be much to be gained from further analyses. For instance, the linguistic data set could be extended in terms of sample sizes, the number of unique words recorded and even the number of languages. Finally, it was the combination of statistical and linguistic expertise that allowed us to make the most progress in understanding our data set. Therefore, it is essential to continue this collaboration in order to extend and improve upon this framework, with the eventual aim of providing new perspective on language relationships that are disputed.

Appendix A

Explicit representations for the G-Wishart

Here we provide explicit representations of the marginal likelihood, the log-likelihood and the derivative of the log-likelihood, the latter indicating that numerical methods are required to determine the maximum likelihoods estimate of δ as $l'(\delta)$ contains the PolyGamma$[z]$ function (see Cuyt et al. [2008] for example). Note that here $|\cdot|$ represents determinant applies to D and U but cardinality when applied to C and S. Elsewhere in the thesis $\det(\cdot)$ represents determinant whereas $|\cdot|$ remains cardinality.

$$p(Z|G) = \frac{p(Z,G)}{p(Z)} = \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{I_G(\delta+n, D+U)}{I_G(\delta, D)} = L(\delta) = \frac{1}{(2\pi)^{\frac{np}{2}}} \frac{\prod_{j=1}^{k} \frac{2^{\frac{(\delta+n+|C_j|-1)|C_j|}{2}} \Gamma_{|C_j|}(\frac{\delta+n+|C_j|-1}{2})}{|D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{2^{\frac{(\delta+|S_j|-1)|S_j|}{2}} \Gamma_{|S_j|}(\frac{\delta+|S_j|-1}{2})}{|D_{S_j}|^{\frac{\delta+|S_j|-1}{2}}}}{\prod_{j=1}^{k} \frac{2^{\frac{(\delta+|C_j|-1)|C_j|}{2}} \Gamma_{|C_j|}(\frac{\delta+|C_j|-1}{2})}{|D_{C_j}|^{\frac{\delta+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{2^{\frac{(\delta+n+|S_j|-1)|S_j|}{2}} \Gamma_{|S_j|}(\frac{\delta+n+|S_j|-1}{2})}{|D_{S_j+U_j}|^{\frac{\delta+n+|S_j|-1}{2}}}} \tag{A.0.1}$$

$$= \frac{1}{(2\pi)^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{2^{\frac{(n+|C_j|-1)|C_j|}{2}} \Gamma_{|C_j|}(\frac{\delta+n+|C_j|-1}{2}) |D_{C_j}|^{\frac{\delta+|C_j|-1}{2}}}{2^{\frac{(\delta+|C_j|-1)|C_j|}{2}} \Gamma_{|C_j|}(\frac{\delta+|C_j|-1}{2}) |D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{2^{\frac{(\delta+|S_j|-1)|S_j|}{2}} \Gamma_{|S_j|}(\frac{\delta+|S_j|-1}{2}) |D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}}}{2^{\frac{(\delta+n+|S_j|-1)|S_j|}{2}} \Gamma_{|S_j|}(\frac{\delta+n+|S_j|-1}{2}) |D_{S_j}|^{\frac{\delta+|S_j|-1}{2}}} \tag{A.0.2}$$

$$= \frac{1}{(2\pi)^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{2^{\frac{n|C_j|}{2}} \Gamma_{|C_j|}(\frac{\delta+n+|C_j|-1}{2}) |D_{C_j}|^{\frac{\delta+|C_j|-1}{2}}}{\Gamma_{|C_j|}(\frac{\delta+|C_j|-1}{2}) |D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{\Gamma_{|S_j|}(\frac{\delta+|S_j|-1}{2}) |D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}}}{2^{\frac{n|S_j|}{2}} \Gamma_{|S_j|}(\frac{\delta+n+|S_j|-1}{2}) |D_{S_j}|^{\frac{\delta+|S_j|-1}{2}}} \tag{A.0.3}$$

$$= \frac{2^{\frac{np}{2}}}{(2\pi)^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{\Gamma_{|C_j|}(\frac{\delta+n+|C_j|-1}{2}) |D_{C_j}|^{\frac{\delta+|C_j|-1}{2}}}{\Gamma_{|C_j|}(\frac{\delta+|C_j|-1}{2}) |D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{\Gamma_{|S_j|}(\frac{\delta+|S_j|-1}{2}) |D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}}}{\Gamma_{|S_j|}(\frac{\delta+n+|S_j|-1}{2}) |D_{S_j}|^{\frac{\delta+|S_j|-1}{2}}} \tag{A.0.4}$$

$$= \frac{1}{\pi^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{\Gamma_{|C_j|}(\frac{\delta+n+|C_j|-1}{2}) |D_{C_j}|^{\frac{\delta+|C_j|-1}{2}}}{\Gamma_{|C_j|}(\frac{\delta+|C_j|-1}{2}) |D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}}} \prod_{j=2}^{k} \frac{\Gamma_{|S_j|}(\frac{\delta+|S_j|-1}{2}) |D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}}}{\Gamma_{|S_j|}(\frac{\delta+n+|S_j|-1}{2}) |D_{S_j}|^{\frac{\delta+|S_j|-1}{2}}} \tag{A.0.5}$$

$$= \frac{1}{\pi^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{|D_{C_j}|^{\frac{\delta+|C_j|-1}{2}} \pi^{\frac{|C_j|(|C_j|-1)}{4}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+n+|C_j|-i-2}{2}} e^{-t} \, \mathrm{d}t}{|D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}} \pi^{\frac{|C_j|(|C_j|-1)}{4}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+|C_j|-i-2}{2}} e^{-t} \, \mathrm{d}t}$$

$$\times \prod_{j=2}^{k} \frac{|D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}} \pi^{\frac{|S_j|(|S_j|-1)}{4}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t}{|D_{S_j}|^{\frac{\delta+|S_j|-1}{2}} \pi^{\frac{|S_j|(|S_j|-1)}{4}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+n+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t} \tag{A.0.6}$$

$$= \frac{1}{\pi^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{|D_{C_j}|^{\frac{\delta+|C_j|-1}{2}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+n+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t}{|D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t} \prod_{j=2}^{k} \frac{|D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t}{|D_{S_j}|^{\frac{\delta+|S_j|-1}{2}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+n+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t} \tag{A.0.7}$$

$$l(\delta) = ln\left(\frac{1}{\pi^{\frac{np}{2}}} \prod_{j=1}^{k} \frac{|D_{C_j}|^{\frac{\delta+|C_j|-1}{2}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+n+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t}{|D_{C_j}+U_{C_j}|^{\frac{\delta+n+|C_j|-1}{2}} \prod_{i=1}^{|C_j|} \int_0^\infty t^{\frac{\delta+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t} \prod_{j=2}^{k} \frac{|D_{S_j}+U_{S_j}|^{\frac{\delta+n+|S_j|-1}{2}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t}{|D_{S_j}|^{\frac{\delta+|S_j|-1}{2}} \prod_{i=1}^{|S_j|} \int_0^\infty t^{\frac{\delta+n+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t} \right) \tag{A.0.8}$$

$$= -\frac{np}{2}\ln(\pi) + \sum_{j=1}^{k}\left(\frac{\delta+|C_j|-1}{2}\ln(|D_{C_j}|) + \sum_{i=1}^{|C_j|}\ln(\int_0^\infty t^{\frac{\delta+n+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t) - \frac{\delta+n+|C_j|-1}{2}\ln(|D_{C_j}+U_{C_j}|) - \sum_{i=1}^{|C_j|}\ln(\int_0^\infty t^{\frac{\delta+|C_j|-i-2}{2}} e^{-t}\,\mathrm{d}t)\right)$$

$$+\sum_{j=2}^{k}\frac{\delta+n+|S_j|-1}{2}\ln(|D_{S_j}+U_{S_j}|) + \sum_{j=2}^{k}\sum_{i=1}^{|S_j|}\ln(\int_0^\infty t^{\frac{\delta+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t) - \sum_{j=2}^{k}\frac{\delta+|S_j|-1}{2}\ln(|D_{S_j}|) - \sum_{j=2}^{k}\sum_{i=1}^{|S_j|}\ln(\int_0^\infty t^{\frac{\delta+n+|S_j|-i-2}{2}} e^{-t}\,\mathrm{d}t) \tag{A.0.9}$$

$$l'(\delta) = \frac{\delta}{2}\sum_{j=1}^{k}\ln(|D_{C_j}|) + \sum_{j=1}^{k}\sum_{i=1}^{|C_j|}\frac{\partial}{\partial\delta}\ln(\int_0^\infty t^{\frac{\delta+n+|C_j|-i-2}{2}}e^{-t}\,\mathrm{d}t) - \frac{\delta}{2}\sum_{j=1}^{k}\ln(|D_{C_j}+U_{C_j}|) - \sum_{j=1}^{k}\sum_{i=1}^{|C_j|}\frac{\partial}{\partial\delta}\ln(\int_0^\infty t^{\frac{\delta+|C_j|-i-2}{2}}e^{-t}\,\mathrm{d}t)$$
$$+ \frac{\delta}{2}\sum_{j=2}^{k}\ln(|D_{S_j}+U_{S_j}|) + \sum_{j=2}^{k}\sum_{i=1}^{|S_j|}\frac{\partial}{\partial\delta}\ln(\int_0^\infty t^{\frac{\delta+|S_j|-i-2}{2}}e^{-t}\,\mathrm{d}t) - \frac{\delta}{2}\sum_{j=2}^{k}\ln(|D_{S_j}|) - \sum_{j=2}^{k}\sum_{i=1}^{|S_j|}\frac{\partial}{\partial\delta}\ln(\int_0^\infty t^{\frac{\delta+n+|S_j|-i-2}{2}}e^{-t}\,\mathrm{d}t) \tag{A.0.10}$$

$$l'(\delta) = \frac{\delta}{2}\sum_{j=1}^{k}\Big(\ln(|D_{C_j}|) - \ln(|D_{C_j}+U_{C_j}|)\Big) + \sum_{j=1}^{k}\sum_{i=1}^{|C_j|}\big(\psi(\delta+n+i-1) - \psi(\delta+i-1)\big)$$
$$+ \frac{\delta}{2}\sum_{j=2}^{k}\Big(\ln(|D_{S_j}+U_{S_j}|) - \ln(|D_{S_j}|)\Big) + \sum_{j=2}^{k}\sum_{i=1}^{|S_j|}\big(\psi(\delta+i-1) - \psi(\delta+n+i-1)\big) \tag{A.0.11}$$

where $\psi(z) = \text{PolyGamma}[z]$.

Appendix B

Summary tables for Section 4.3.4

TABLE B.1: Number of violations for $n = 500$ and simulation of 10^4 repetitions.

# of violations	Tree I	Non-tree I	Tree II	Non-tree II
0	0	2	0	4
1-3	0	1821	0	195
4-6	240	1821	87	1338
7-9	2174	3429	1102	3023
10-12	4137	3003	3447	3331
13-15	2845	1281	3824	1770
16-18	576	217	1458	330
19-20	28	8	82	9

TABLE B.2: Number of violations for $n = 883$ and simulation of 10^4 repetitions.

# of violations	Tree I	Non-tree I	Tree II	Non-tree II
0	0	40	0	20
1-3	0	1377	0	925
4-6	1375	3846	591	3034
7-9	4720	3294	3086	3448
10-12	3118	1238	4066	2043
13-15	735	197	1930	494
16-18	52	8	320	35
19-20	0	0	7	1

TABLE B.3: Number of violations for $n = 1500$ and simulation of 10^4 repetitions.

# of violations	Tree I	Non-tree I	Tree II	Non-tree II
0	0	249	0	141
1-3	0	3599	0	2406
4-6	4208	4428	2124	4054
7-9	4768	1506	4733	2521
10-12	955	208	2565	780
13-15	67	10	531	93
16-18	2	0	47	4
19-20	0	0	0	1

TABLE B.4: Number of violations for $n = 5000$ and simulation of 10^4 repetitions.

# of violations	Tree I	Non-tree I	Tree II	Non-tree II
0	0	2916	0	1910
1-3	0	6461	0	5814
4-6	9658	615	8572	1950
7-9	342	8	1333	310
10-12	0	0	90	16
13-15	0	0	5	0
16-18	0	0	0	0
19-20	0	0	0	0

Bibliography

V. Algazi. On the optimality of the karhunen-loève expansion (corresp.). *Information Theory, IEEE Transactions on*, 15(2):319–321, 1969.

K. Allan. *The Oxford Handbook of the History of Linguistics*. Oxford Handbooks Series. OUP Oxford, 2013.

E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Advances in Applied Mathematics*, 40(2):127–148, 2008.

E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132, 2009.

E. S. Allman, J. A. Rhodes, and A. Taylor. A semialgebraic description of the general Markov model on phylogenetic trees. *SIAM Journal on Discrete Mathematics*, 28(2):736–755, 2014.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

J. A. D. Aston and C. Kirch. Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, 6(4):1906–1948, 2012.

J. A. D. Aston, J.-M. Chiou, and J. P. Evans. Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society Series C. Applied Statistics*, 59(2):297–317, 2010.

J. A. D. Aston, D. Buck, J. Coleman, C. J. Cotter, N. S. Jones, V. Macaulay, N. MacLeod, J. M. Moriarty, and A. Nevins. Phylogenetic inference for function-valued traits: speech sound evolution. *Trends in Ecology & Evolution*, Volume 27(3):160–166, 2012.

A. Atay-Kayis and H. Massam. A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika*, 92(2):317–335, 2005.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

J. Barnard, R. McCulloch, and X.-L. Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312, 2000.

G. P. Basharin, A. N. Langville, and V. A. Naumov. The life and work of AA Markov. *Linear Algebra and its Applications*, 386:3–26, 2004.

P. A. Bekker and J. de Leeuw. The rank of reduced dispersion matrices. *Psychometrika*, 52(1): 125–135, 1987.

P. M. Bentler. Multivariate analysis with latent variables: Causal modeling. *Annual review of psychology*, 31(1):419–456, 1980.

D.P. Bertsekas and W. Rheinboldt. *Constrained Optimization and Lagrange Multiplier Methods*. Computer science and applied mathematics. Elsevier Science, 2014. ISBN 9781483260471. URL https://books.google.co.uk/books?id=j6LiBQAAQBAJ.

P. C. Besse, H. Cardot, and D. B. Stephenson. Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics*, 27(4):673–687, 2000.

K. A. Bollen. *Structural equations with latent variables*. John Wiley & Sons, 2014.

K. A. Bollen and R. A. Stine. Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, 21(2):205–229, 1992.

K. A. Bollen and K.-f. Ting. Confirmatory Tetrad Analysis. In P. Marsden, editor, *Sociological Methodology*, volume 23 of *Sociological Methodology*, chapter 5, pages 147–175. Wiley, 1993.

D. A. W. Bolnick, B. A. S. Shook, L. Campbell, and I. Goddard. Problematic use of Greenberg's linguistic classification of the Americas in studies of Native American genetic variation. *American Journal of Human Genetics*, 75(3):519, 2004.

A. Bouchard-Côté, D. Hall, T. L. Griffiths, and D. Klein. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proceedings of the National Academy of Sciences*, 110(11):4224–4229, 2013.

G. E. P. Box. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36(3/4): 317–346, 1949.

J. L. Brown, Jr. Mean square truncation error in series expansions of random functions. *Journal of the Society for Industrial and Applied Mathematics*, 8(1):28–32, 1960.

P. Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory B*, 17: 48–50, 1974.

N. A. Campbell and W. R. Atchley. The geometry of canonical variate analysis. *Systematic Biology*, 30(3):268–280, 1981.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2008.

M. Casanellas and J. Fernández-Sánchez. Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees. *Molecular Biology and Evolution*, 24(1):288, 2007.

P. E. Castro, W. H. Lawton, and E. A. Sylvestre. Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28(4):329–337, 1986.

B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Molecular Biology and Evolution*, 17(10):1529–1541, 2000.

T. F. Cox and M. A. A. Cox. *Multidimensional scaling*. CRC Press, 2010.

A. Cuyt, F. Backeljauw, and C. Bonan-Hamada. *Handbook of Continued Fractions for Special Functions*. SpringerLink: Springer e-Books. Springer, 2008.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366, 1980.

A. P. Dawid and S. L. Lauritzen. Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317, 1993.

C. de Boor. *A practical guide to splines*. Springer, revised edition, 2001.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281, 1979.

A. Dress, K. T. Huber, J. Koolen, V. Moulton, and A. Spillner. *Basic phylogenetic combinatorics*. Cambridge University Press, Cambridge, 2012.

M. Drton and A. Goia. Correction on: Moments of minors of Wishart matrices. *The Annals of Statistics*, 40(2):1283–1284, 2012.

M. Drton and S. Sullivant. Algebraic statistical models. *Statistica Sinica*, 17(4):1273–1297, 2007.

M. Drton, B. Sturmfels, and S. Sullivant. Algebraic factor analysis: tetrads, pentads and beyond. *Probab. Theory Related Fields*, 138(3-4):463–493, 2007.

M. Drton, H. Massam, and I. Olkin. Moments of minors of Wishart matrices. *The Annals of Statistics*, pages 2261–2283, 2008.

T. E. Duncan, S. C. Duncan, and L. A. Strycker. *An introduction to latent variable growth curve modeling: Concepts, issues, and application*. Routledge Academic, 2013.

M. Dunn, A. Terrill, G. Reesink, R. A. Foley, and S. C. Levinson. Structural phylogenetics and the reconstruction of ancient language history. *Science*, 309(5743):2072–2075, 2005.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961.

J. Dutkowski and J. Tiuryn. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 23(13):149–158, 2007.

I. Dyen, J. Kruskal, and P. Black. Comparative Indo-European Database collected by Isidore Dyen. http://www.wordgumbo.com/ie/cmp/iedata.txt, February 1997.

L. Egghe. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58(5):702–709, 2007.

P. H. C. Eilers and B. D. Marx. Flexible smoothing with B-splines and penalties. *Statistical science*, pages 89–102, 1996.

A. Engström, P. Hersh, and B. Sturmfels. Toric cubes. *Rendiconti del Circolo Matematico di Palermo*, pages 1–12, 2012.

J. S. Farris. The retention index and the rescaled consistency index. *Cladistics*, 5(4):417–419, 1989.

J. Felsenstein. The number of evolutionary trees. *Systematic Biology*, 27(1):27–33, 1978.

J. Felsenstein. Statistical inference of phylogenies. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 146(3):246–272, 1983.

G. Fervaha and G. Remington. Interpreting a multivariate analysis of functional neuroimaging data. *Frontiers in Psychiatry*, 3(52), 2012.

P. Forster and A. Toth. Toward a phylogenetic chronology of ancient Gaulish, Celtic, and Indo-European. *Proceedings of the National Academy of Sciences*, 100(15):9079–9084, 2003.

S. A. Fulop. *Speech Spectrum Analysis*. Signals and communication technology. Springer, 2011.

D. Garcia. Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational statistics & data analysis*, 54(4):1167–1178, 2010.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

D. Gervini and P. A. Carter. Warped functional analysis of variance. *Biometrics*, 70(3):526–535, 2014.

J. Gill, S. Linusson, V. Moulton, and M. Steel. A regular decomposition of the edge-product space of phylogenetic trees. *Advances in Applied Mathematics*, 41(2):158–176, 2008.

L. A. Goodman and J. M. Mirande. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231, 1974.

E. Grabe, G. Kochanski, and J. Coleman. Connecting intonation labels to mathematical descriptions of fundamental frequency. *Language and Speech*, 50(3):281–310, 2007.

J. H. Greenberg. *Method and perspective in anthropology: papers in honor of Wilson D. Wallis*, chapter A quantitative approach to the morphological typology of language, pages 192–220. University of Minnesota Press, 1954. Editor: Spencer, R F.

S. T. Gries. *Quantitative Corpus Linguistics with R: A Practical Introduction*. Taylor & Francis, 2009.

S. Grünewald, P. J. Humphries, and C. Semple. Quartet Compatibility and the Quartet Graph. *The Electronic Journal of Combinatorics*, 15(1):R103, 2008.

P.-Z. Hadjipantelis. Functional data analysis in phonetics. Thesis, December 2013. URL http://wrap.warwick.ac.uk/62527/.

P.-Z. Hadjipantelis, J. A. D. Aston, and J. P. Evans. Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. *The Journal of the Acoustical Society of America*, 131(6):4651–4664, 2012.

W. Hao and G. B. Golding. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC genomics*, 9(1):235, 2008.

S. Harmeling and C. K. I. Williams. Greedy learning of binary latent trees. *Pattern Analysis and Machine Intelligence*, 33(6):1087–1097, 2011.

P. D. Hebert, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis. Identification of birds through DNA barcodes. *PLoS Biology*, 2(10):e312, 2004.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics-Theory and Methods*, 19(10):3595–3617, 1990.

G. Herdan. *Type-token Mathematics: A Textbook of Mathematical Linguistics*. Janua linguarum. series maior. no. 4. Mouton en Company, 1960.

S. H. Holan, C. K. Wikle, L. E. Sullivan-Beckers, and R. B. Cocroft. Modeling complex Phenotypes: Generalized Linear Models Using Spectrogram Predictors of Animal Communication Signals. *Biometrics*, 66(3):914–924, 2010.

L. Horváth and P. Kokoszka. *Inference for Functional Data with Applications*. Springer, New York, 2012.

V. I. Istratescu. *Inner Product Structures: Theory and Applications*. Springer, 1987.

I. T. Jolliffe. *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition, 2002.

N. S. Jones and J. Moriarty. Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of the Royal Society Interface*, 10(78):20120616, 2012.

H. T. Kiiveri. Canonical variate analysis of high-dimensional spectral data. *Technometrics*, 34 (3):321–331, 1992.

J. Kim. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Molecular phylogenetics and evolution*, 17(1):58–75, 2000.

J. Klemela. *regpro: Nonparametric Regression*, 2013. URL http://CRAN.R-project.org/package=regpro. R package version 0.1.0.

L. L. Koenig, J. C. Lucero, and E. Perlman. Speech production variability in fricatives of children and adults: results of functional data analysis. *Journal of the Acoustical Society of America*, 124(5):3158–3170, November 2008.

E. F. K. Koerner. *Linguistic Historiography: Projects & Prospects*. Amsterdam studies in the theory and history of linguistic science. J. Benjamins, 1999.

R. Köhler, G. Altmann, and R. G. Piotrowski. *Quantitative Linguistics: An International Handbook*. Handbooks of Linguistics and Communication Science (HSK). De Gruyter, 2005.

D. Koller, N. Friedman, L. Getoor, and B. Taskar. Graphical Models in a Nutshell. In L. Getoor and B. Taskar, editors, *Introduction to Statistical Relational Learning*. MIT Press, 2007.

H. Kramer. On best approximation of random processes et al.(Corresp.). *IRE Transactions on Information Theory*, 1(6):52–53, 1960.

W. J. Krzanowski. *Principles of multivariate analysis*, volume 3 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1990. A user's perspective, Corrected reprint of the 1988 edition, Oxford Science Publications.

H. Kučera and W. N. Francis. *Computational analysis of present-day American English*. Brown University Press, 1967.

J. Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

H. O. Lancaster. *The Chi-Square Distribution*. John Wiley & Sons, 1969.

P. Lancaster and M. Tismenetsky. *The theory of matrices*. Computer Science and Applied Mathematics. Academic Press Inc., Orlando, FL, second edition, 1985.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press, 1996. Oxford Science Publications.

C. M. Lee, S. Narayanan, and R. Pieraccini. Recognition of negative emotions from the speech signal. In *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, pages 240–243, 2001.

D. Libkind, C. T. Hittinger, E. Valério, C. Gonçalves, J. Dover, M. Johnston, P. Gonçalves, and J. P. Sampaio. Microbe domestication and the identification of the wild genetic stock of lager-brewing yeast. *Proceedings of the National Academy of Sciences*, 108(35):14539–14544, 2011.

G. Lindgren. *Stationary Stochastic Processes: Theory and Applications*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2012.

G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, C. M. Tsai, I. J.and Bergman, D. Bensasson, M. J. T. O'Kelly, A. van Oudenaarden, D. B. Barton, E. Bailes, M. Jones, R. Durbin, and E. J. Louis. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–341, 2009.

P.-L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049, 2013.

C. J. Long, E. N. Brown, C. Triantafyllou, I. Aharon, L. L. Wald, and V. Solo. Nonstationary noise estimation in functional MRI. *Neuroimage*, 28(4):890–903, 2005.

J. C. Lucero and L. L. Koenig. Time normalization of voice signals using functional data analysis. *The Journal of the Acoustical Society of America*, 108(4):1408–1420, 2000.

A. Lüdeling. *Corpus Linguistics*. Number v. 2 in Handbooks of Linguistics and Communication Science (HSK). De Gruyter, 2009.

K. Malmkjaer. *The Routledge Linguistics Encyclopedia*. Taylor & Francis, 2009.

M. Marcet-Houben and T. Gabaldón. The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *PLoS One*, 4(2):e4357, 2009.

J. G. Martinez, K. M. Bohn, R. J. Carroll, and J. S. Morris. A Study of Mexican Free-Tailed Bat Chirp Syllables: Bayesian Functional Mixed Models for Nonstationary Acoustic Time Series. *Journal of the American Statistical Association*, 108(502):514–526, 2013.

C. J. Mecklin and D. J. Mundfrom. A Monte Carlo comparison of the Type I and Type II error rates of tests of multivariate normality. *Journal of Statistical Computation and Simulation*, 75 (2):93–107, 2005.

C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, pages 130–162, 1957.

M. Mizuta. Graphical Representation of Functional Clusters and MDS Configurations. In *Data Analysis, Classification and the Forward Search*, pages 31–37. Springer, 2006.

C. Mooshammer. Acoustic and laryngographic measures of the laryngeal reflexes of linguistic prominence and vocal effort in German. *The Journal of the Acoustical Society of America*, 127(2):1047–1058, 2010.

V. Moulton and M. Steel. Peeling phylogenetic 'oranges'. *Advances in Applied Mathematics*, 33(4):710–727, 2004.

L. Nakhleh, D. A. Ringe, and T. Warnow. Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages. *Language*, 81(2):382–420, 2005.

S. Nelson-Sathi, J.-M. List, H. Geisler, H. Fangerau, R. D. Gray, W. Martin, and T. Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 278(1713):1794–1803, 2011.

G. K. Nicholls and R. D. Gray. Dated ancestral trees from binary trait data and their application to the diversification of languages. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):545–566, 2008.

G. K. Nicholls and R. J. Ryder. Phylogenetic models for semitic vocabulary. In *Proceedings of the 26th International Workshop on Statistical Modelling*, 2011.

J. Nichols. *Linguistic diversity in space and time*. University of Chicago Press, 1992.

S. Nittrouer, R. S. McGowan, P. H. Milenkovic, and D. Beehler. Acoustic measurements of men's and women's voices: a study of context effects and covariation. *Journal of Speech and Hearing Research*, 33(4):761–775, December 1990.

M. P. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press Series. Edinburgh University Press, 1998.

I. Olkin and A. W. Marshall. *Inequalities: Theory of Majorization and Its Applications*. Mathematics in Science and Engineering. Elsevier Science, 2014.

A. J. O'Malley and A. M. Zaslavsky. Domain-level covariance analysis for multilevel survey data with structured nonresponse. *Journal of the American Statistical Association*, 103(484): 1405–1418, 2008.

E. S. Parris and M. J. Carey. Language independent gender identification. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, volume 2, pages 685–688vol. 2, 1996.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Representation and Reasoning Series. Morgan Kaufmann Publishers, 1988.

J. Pearl and L. Xu. Structuring Causal Tree Models with Continuous Variables. In *Proceedings of the Third Annual Conference on Uncertainty in Artificial Intelligence, Seattle, WA, USA, July 10-12, 1987*, 1987.

E. Pépiot. Voice, speech and gender: male-female acoustic differences and cross-language variation in English and French speakers. In *XVèmes Rencontres Jeunes Chercheurs de l'ED 268*, pages à–paraître, 2013.

N. I. Platnick and H. D. Cameron. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Systematic Zoology*, pages 380–385, 1977.

J. H. Pollard. *A Handbook of Numerical and Statistical Techniques: With Examples Mainly from the Life Sciences*. Cambridge University Press, 1979.

G. Price. Romance. In J. Gvozdanović, editor, *Indo-European Numerals*, Mathematical Research, chapter 13. Berlin, 1992.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.

S. J. Ratcliffe, L. R. Leader, and G. Z. Heller. Functional data analysis with application to periodically stimulated foetal heart rate data. I: functional regression. *Statistics in Medicine*, 21(8):1103–1114, April 2002.

S. Ratnasingham and P. D. N. Hebert. BOLD: the barcode of life data system. *Molecular Ecology Notes*, 7(3(i)):355–364, 2007a.

S. Ratnasingham and P. D. N. Hebert. Barcode of Life Data Systems v3, 2007b. URL http://www.boldsystems.org/.

K. Rexová, D. Frynta, and J. Zrzavỳ. Cladistic analysis of languages: Indo-European classification based on lexicostatistical data. *Cladistics*, 19(2):120–127, 2003.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):233–243, 1991.

O. Rieppel. The series, the network, and the tree: changing metaphors of order in nature. *Biology and Philosophy*, 25(4):475–496, 2010.

Jonathan Rougier. A representation theorem for stochastic processes with separable covariance functions, and its implications for emulation.

A. Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.

J. P. Royston. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, pages 115–124, 1982.

J. P. Royston. Some techniques for assessing multivarate normality based on the Shapiro-Wilk W. *Applied Statistics*, pages 121–133, 1983.

G. Sampson. Statistical Linguistics. In William J. Frawley, editor, *International Encyclopedia of Linguistics*. Oxford University Press, 2003.

A. Schleicher. *Die Deutsche Sprache*. Cotta, 1860.

L. Schwaller, S. Robin, and M. Stumpf. Bayesian Inference of Graphical Model Structures Using Trees. *ArXiv e-prints*, April 2015.

C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

R. Settimi and J. Q. Smith. On the geometry of Bayesian graphical models with hidden variables. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 472–479. Morgan Kaufmann, 1998.

R. Settimi and J. Q. Smith. Geometry, moments and Bayesian networks with hidden variables. In *Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics*, pages 293–298. Morgan Kaufmann, 1999.

R. Settimi and J. Q. Smith. Geometry, moments and conditional independence trees with hidden variables. *The Annals of Statistics*, 28(4):1179–1205, 2000.

N. Shiers and J. Q. Smith. Graphical inequality diagnostics for phylogenetic trees. In *Proceedings of 6^{th} European workshop on probabilistic graphical models, Granada, Spain, 19-21 Sep 2012*, pages 291–298, 2012.

N. Shiers, J. A. D. Aston, J. Q. Smith, and J. S. Coleman. Gaussian tree constraints applied to acoustic linguistic functional data. *ArXiv e-prints*, October 2014.

N. Shiers, P. Zwiernik, J. A. D. Aston, and J. Q. Smith. The correlation space of Gaussian latent tree models and model selection without fitting. *ArXiv e-prints*, April 2016.

J. Q. Smith. *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press, 2010.

H. Sørensen, J. Goldsmith, and L. M. Sangalli. An introduction with medical applications to functional data analysis. *Statistics in Medicine*, 2013.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2001.

E. Stanghellini and B. Vantaggi. Identification of discrete concentration graph models with one hidden binary variable. *Bernoulli*, 19(5A):1920–1937, 2013.

B. Streitberg. Lancaster interactions revisited. *The Annals of Statistics*, 18(4):1878–1885, 1990.

B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology*, 12(2):204–228, 2005.

L. E. Sucar. *Probabilistic Graphical Models: Principles and Applications*. Advances in Computer Vision and Pattern Recognition. Springer London, 2015.

C. Sujatha. *Vibration And Acoustics*. McGraw-Hill Education (India) Pvt Limited, 2010.

Seth Sullivant. Algebraic geometry of Gaussian Bayesian networks. *Advances in Applied Mathematics*, 40(4):482–513, 2008.

M. Swadesh. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proceedings of the American philosophical society*, pages 452–463, 1952.

P. Syal and D. V. Jindal. *An Introduction to Linguistics: Language, Grammar and Semantics*. Eastern Economy Edition. PHI Learning, 2007.

R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 2008.

E. C. Teeling and S. B. Hedges. Making the impossible possible: rooting the tree of placental mammals. *Molecular biology and evolution*, 30(9):1999–2000, 2013.

M. Těšitelová. *Quantitative Linguistics*. Linguistic & literary studies in Eastern Europe. Benjamins Pub., 1992.

N. H. Timm. *Applied Multivariate Analysis*. Springer Texts in Statistics. Springer New York, 2007.

A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, and L. Cupul-Magana. *MBoxtstwod: Multivariate Statistical Testing for the Homogeneity of Covariance Matrices Without Data by the Box's M.*, 2004. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=6548.

A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, and L. Cupul-Magana. *Roystest: Royston's Multivariate Normality Test*, 2007. URL http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17811.

T Verma and J Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270. Elsevier Science, 1990.

K. Vij and R. Biswas. *Basics of DNA and evidentiary issues*. Jaypee Brothers Publishers, 2005.

J. S. Walker. *Fast Fourier Transforms, Second Edition*. Studies in Advanced Mathematics. Taylor & Francis, 1996.

Haonan Wang, JS Marron, et al. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

Jonas Warringer, Eniko Zorgo, Francisco A Cubillos, Amin Zia, Arne Gjuvsland, Jared T Simpson, Annabelle Forsmark, Richard Durbin, Stig W Omholt, Edward J Louis, et al. Trait variation in yeast is defined by population history. *PLoS Genetics*, 7(6):e1002111, 2011.

M. Welling. Robust higher order statistics. In *Tenth International Workshop on Artificial Intelligence and Statistics*, pages 405–412. Citeseer, 2005.

L. Wenyin, X. Quan, M. Feng, and B. Qiu. A short text modeling method combining semantic and statistical information. *Information Sciences*, 180(20):4031–4041, 2010.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley New York, 1990.

J. Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928a.

J. Wishart. Sampling errors in the theory of two factors. *British Journal of Psychology. General Section*, 19(2):180–187, 1928b.

Sewall Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

Z. Yang. *Computational Molecular Evolution*. Oxford University Press, 2007.

F. Yao and T. Lee. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):3–25, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

P. Zwiernik and J. Q. Smith. Implicit inequality constraints in a binary tree model. *Electronic Journal of Statistics*, 5:1276–1312, 2011.

P. Zwiernik and J. Q. Smith. Tree cumulants and the geometry of binary tree models. *Bernoulli*, 18(1):290–321, 2012.