MASTER'S DEGREE THESIS

Master of Science in Advanced Mathematics and Mathematical Engineering

Title: Multilinear Algebra for Phylogenetic Reconstruction

Author: Marina Garrote López

Advisor: Marta Casanellas Rius and Jesús Fernández Sánchez

Department: Departament de Matemàtica Aplicada I

Academic year: 2014-2015

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Universitat Politècnica de Catalunya
Facultat de Matemàtiques i Estadística

Master's Degree Thesis

Multilinear Algebra for Phylogenetic Reconstruction

Marina Garrote López

Advisors: Marta Casanellas Rius
and Jesús Férnandez Sánchez

Departament de Matemàtica Aplicada I - UPC

Abstract

Key words: Phylogenetic tree, phylogenetic invariants, general Markov model, joint distribution, tensor

MSC 2010: 92D15, 92D20, 14P10, 60J20, 62P10

Phylogenetic reconstruction tries to recover the ancestral relationships among a group of contemporary species and represent them in a phylogenetic tree. To do it, it is useful to model evolution adopting a parametric statistic model. Using these models one is able to deduce polynomial relationships between the observed probabilities, known as *phylogenetic invariants*. Mathematicians have recently begun to be interested in the study of these polynomials and have developed techniques from algebraic geometry that have already been used in the study of phylogenetics. Nowadays there exist some phylogenetic reconstruction methods based in these phylogenetic invariants. In this project we study some theoretical results on stochasticity conditions of the parameters of the model and we analyze whether they give some new information to these reconstruction methods. We implement the conditions and analyze the results comparing them with the results provided by the reconstruction method *Erik+2* ([FSC15]). Finally we propose a new reconstruction method based in the same ideas, with different implementation, and with very good results on simulated data.

Contents

1 Introduction **1**

2 Preliminaries **3**
- 2.1 Biological preliminaries 3
- 2.2 Phylogenetic trees 5
- 2.3 Evolutionary models 7
- 2.4 Joint distribution 10
- 2.5 Phylogenetic invariants 11
- 2.6 Flattening 14
- 2.7 Tensors 15

3 Theoretical results **17**
- 3.1 Some operations with tensors 17
- 3.2 Stochasticity Conditions 21

4 Implementation and results on simulated data **29**
- 4.1 Numerical and computational issues 29
- 4.2 Analysis of the Results 34

5 A new method for phylogenetic reconstruction **39**

6 Conclusions **45**

1 Introduction

Strong evidences suggest that all the living organisms share a common ancestor and therefore, are related by evolutionary relationships. These relationships are usually expressed in the form of a phylogenetic tree.

Nowadays there are more and more mathematicians and statisticians who collaborate with biologists in order to solve the major problems of the phylogenetics. Many different areas of mathematics are involved in phylogenetic studies, for instance, statistics, probability, algebra, combinatorics and numerical methods. Even more, recently developed techniques from algebraic geometry have already been used in the study of phylogenetics.

The main goal of phylogenetic reconstruction is recovering the ancestral relationships among a group of current species. Moreover it tries to identify which regions of the DNA sequences of contemporary species contain analogous information and study the evolutionary relationships between these species. Another important aim of Phylogenetics is to recover the evolutionary distance from different species.

In order to reconstruct phylogenetic trees it is necessary to model evolution adopting a parametric statistic model. Using these models one is able to deduce polynomial relationships between the parameters of the model, known as *phylogenetic invariants*. Mathematicians have recently begun to be interested in the study of these polynomials and the geometry of the algebraic varieties that arise in this setting. Furthermore they have started to use these phylogenetic invariants to reconstruct phylogenetic trees.

The framework of this project is to understand the relationship between phylogenetics and these algebraic techniques to recover phylogenetic trees from real data. Our main goal is to study and analyze the characterizations of stochasticity of the points in the algebraic varieties mentioned above provided in [ART12], and use them to infer new methods for phylogenetic inference.

This memoir is divided into 2 parts. In the first part, we recall the basic definitions and results on phylogenetics and multilinear algebra that will be used throughout the work. The second part contains our personal contribution and our suggestions for a new method of phylogenetic reconstruction.

First of all we explain concepts that are already known. We explain what *phylogenetic trees* are from the mathematical standpoint and we present several *evolutionary*

models for these trees. Once we have studied the models we will explain what *phylogenetic invariants* are. Moreover define *joint distributions* and its representation as a tensor. We will define some operations among tensors that will be useful, and their meaning in terms of phylogenetic trees. After that we will understand and prove some results about the stochasticity of the parameters of the general Markov model in a tree. One of these results (Theorem 3.2.4 of [ART12]) has been restated and the proof rewritten since the original Theorem contains an error in the proof (see Remark 3.2.5). In Corollary 3.2.9 we present a new equality (phylogenetic invariant) that characterizes when data arises from some topology. This part is developed in Chapter 2 and Chapter 3.

The main goal of the second part of the project is to implement the theoretical results proved in Chapter 3 and see if these conditions of stochasticity of the parameters give some useful information for phylogenetic inference. In Chapter 4 we explain this implementation and we analyze the results and compare them with the reconstruction method *Erik+2* ([FSC15]). We will discuss if these results provide new information to *Erik+2*.

Finally in Chapter 5, we take all this into account and propose a new reconstruction method. It is based on the ideas exposed in Chapter 3 but with a different implementation. This new method has been tested and has obtained very good results which will be analyzed.

2
Preliminaries

2.1 Biological preliminaries

Phylogenetics is the study of relationships between different species or biological entities. It studies how species evolve and where contemporary species come from. According to the theory of the biological evolution developed by Darwin (s.XIX), all species of organisms evolve through the natural selection of small variations that increase the individual's ability to compete, survive, and reproduce. We can model these specialization processes with phylogenetic trees (see Fig.2.1). The nodes of this tree represent different species and every branch is an evolutionary process between two species. The leaves of the tree are contemporary species and the root of the tree is the common ancestor of all the species represented on the tree.

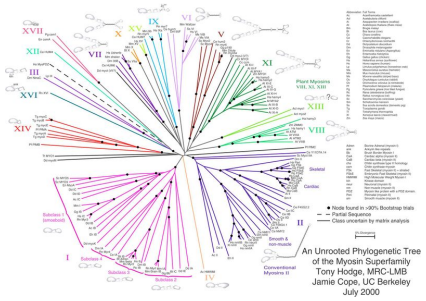

Figure 2.1: A phylogenetic tree.

Genetic information of each individual is encoded in the DNA of the nucleus of its cells. DNA molecules are composed of simpler units called *nucleotides* and consist of two anti-parallel strands of nucleotides coiled around each other to form a double helix. Each nucleotide is composed of a phosphate, a sugar and a basis. According to the bases, nucleotides are called adenine (`A`), cytosine (`C`), guanine (`G`) and thymine (`T`). A base-pair is one of the pairs $\texttt{A}-\texttt{T}$ or $\texttt{C}-\texttt{G}$. The nucleotides on a base-pair are complementary in the sense that in the double helix adenine connects with the thymine and the guanine with cytosine. According to this symmetry, we store a DNA molecule as an ordered sequence of $\texttt{A}, \texttt{C}, \texttt{G}$ and `T` (see Fig. 2.2).

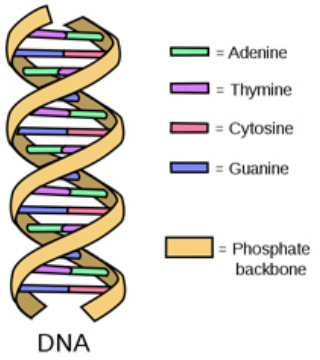

Figure 2.2: DNA molecule.

The heredity information in a genome is thought to be contained in the genes. But the DNA sequences of a same gene may not be the same for different species. They contain similar parts but they can also contain some other parts that we can not compare. For that reason the first problem is identifying which part of the DNA sequences of different species we can compare. This information is collected in an *alignment*. A sequence alignment is a way of arranging the sequences of DNA to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. We can represent the alignment with a table whose rows are the species DNA sequences and whose columns correspond to nucleotides that have evolved from the same nucleotide of the common ancestor of all the species (see Table 2.1). Alignments are used in many contexts, in phylogenetics among them, to see relationships between some species and to reconstruct the phylogenetic tree that relates them. Changes in DNA sequences of different species are given by substitutions, insertions or deletions. In the two latter cases, a nucleotide is inserted or deleted from a given position as compared with the other sequence. In most commonly used evolutionary models, insertions and deletions are not considered and incorporating them would highly increase the complexity of the model. So in this work we will assume that mutations in different alignments are just substitutions. Therefore the alignments we will deal with have the same length and contain no gaps.

Gorilla Gorilla	AACTTCGAGGCTTACCGCTG
Homo Sapiens	AACGTCTATGCTCACCGATG
Pan Troglodytes	AAGGTCGATGCTCACCGATG

Table 2.1: A multiple sequence alignment of DNA sequences of *Homo Sapiens* (Human), *Pan Troglodytes* (Chimpanzee) and *Gorilla Gorilla* (Gorilla).

2.2 Phylogenetic trees

The basic object in a phylogenetic model is a tree $\mathcal{T}$ that contains the evolutionary relationships among a given set of species. In this section we introduce some concepts that allow us to deal with these phylogenetic trees following the approach in [AR04], [AR05] and [Cas12].

2.2.1 Definition (Basic notions of trees) A *tree* $\mathcal{T}$ is a connected graph with no cycles. The *degree* of a vertex is the number of edges incident to it. The vertices of degree 1 are called *leaves* and the set of leaves is denoted by $L(\mathcal{T})$. All the other vertices, which have degree at least 2, are the *interior nodes* of the tree and are designated by the set $Int(\mathcal{T})$. $E(\mathcal{T})$ is the set of the edges of the tree. If all nodes in $Int(\mathcal{T})$ have degree 3, then $\mathcal{T}$ is called a *trivalent tree*.

2.2.2 Definition (Rooted tree) A tree is called a *rooted tree* if one vertex has been labelled as "root", and the edges are oriented away from it.

2.2.3 Definition (Phylogenetic tree) Let X denote a finite set of labels. Then a *phylogenetic tree* is a pair $(\mathcal{T}, \phi)$ where $\mathcal{T}$ is a tree and $\phi : X \to L(\mathcal{T})$ is a one-to-one correspondence.

In a phylogenetic tree, the set X represents a set of living species and the tree $\mathcal{T}$ shows the ancestral relationships among them. Every edge represents an evolutionary process between two species and if it is rooted, then the root represents the common ancestor to the set of species X. For our purposes, usually X will be taken as the set $\{1, 2, \ldots, n\}$.

Another important concept in Phylogenetics is the length of the edges, called *branch length*, that represents the evolutionary distance between different species by the number of nucleotide changes per position that have occurred along the evolutionary process related to the edge.

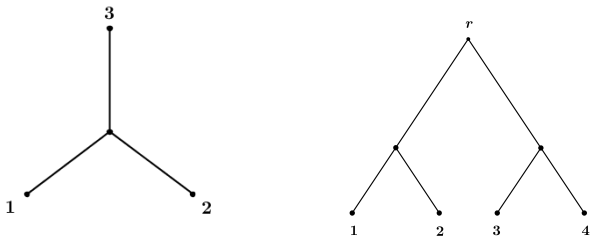

Figure 2.3: At the left an unrooted 3-leaf tree. At the right a rooted phylogenetic tree.

2.2.4 Definition (Tree topology) The *tree topology* of a phylogenetic tree is the topology of the tree as a labelled graph.

That is, two phylogenetic trees $\mathcal{T}_1$ and $\mathcal{T}_2$, with the same set of labels X at the leaves, have the same topology if there is a one-to-one correspondence φ between their vertices that respects adjacency and their leaf labelling. If they are rooted trees and $r_1,\ r_2$ are their roots respectively then we need to impose $\varphi(r_1) = (r_2)$.

2.2.5 Remark For the remainder, we denote by T_n the set of all possible possible unrooted trivalent tree topologies for n-leaf trees. Note that the n has to be greater or equal than 3 and that $\mid T_3 \mid = 1$, which corresponds to the tree represented in Figure 2.3. We will denote the three possible topologies of T_4 by $T_{12|34}$, $T_{13|24}$ and $T_{14|23}$, see Figure 2.5.

We finish this section with an example that illustrates all these definitions.

2.2.6 Example In the Figure 2.3 we can see a rooted phylogenetic tree with the root r placed at the top node. The 4 leaves of tree have been labeled with the set $X = \{1, 2, 3, 4\}$.

Some trees topologically equivalent to the one represented above are pictured in Figure 2.4. If we consider the two trees on the left as unrooted trees then all of them have the same topology

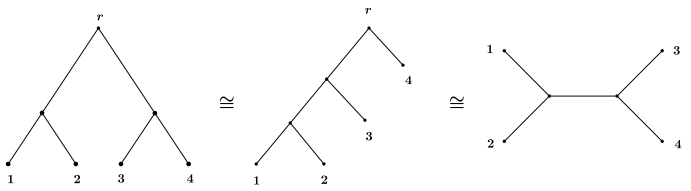

Figure 2.4: Some phylogenetic trees with the same topology as unrooted graphs.

Finally, in Figure 2.5 the three possible topologies of T_4 are represented.

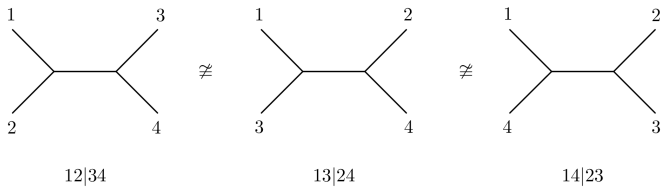

Figure 2.5: The three topologies of T_4:$T_{12|34}$, $T_{13|24}$ and $T_{14|23}$.

One major goal in Phylogenetics is, given an alignment of DNA sequences for n different species, infer which of the T_n topologies explains best the evolution of this set of species. Another goal in Phylogenetics is to infer the branch lengths on this tree (evolutionary distance), but we will not deal with this problem in this work.

2.3 Evolutionary models

Evolution is usually modeled adopting a parametric statistical model. That is, evolution is assumed to be a stochastic process, in which nucleotides mutate randomly over time according to certain probabilities. In order to model an evolutionary process between species we need to assume some hypothesis. We assume that,

- Nucleotides in the DNA sequence are independent and identically distributed (iid). This means that the states at each position in the sequence evolve independently of the other positions and according to the same evolutionary process.
- The DNA mutations occurs randomly.
- Evolutionary processes in different edges only relay in the common node so they are independent.

Assuming these hypothesis we associate a discrete random variable X_i to each node i of $\mathcal{T}$ such that X_i can take κ different states. We denote by $\mathcal{K}$ this set of states. Usually $\mathcal{K}$ is the set of the four nucleotides in DNA, which are denotes by their first letter. Therefore $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and $\kappa = 4$. Since DNA sequences of the contemporary species are known, we say that the random variables at the leaves are observed. However we do not have any information about the ancestral species, that is why the random variables at the interior nodes are hidden. For a tree $\mathcal{T}$ with leaves $1, 2, \ldots, n$, $X = (X_1, X_2, \ldots, X_n)$ represents the joint distribution vector of the leaves. Each column of an alignment is an observation of this vector of random variables.

We introduce now the parameters of a model in a rooted tree $\mathcal{T}$. The vector $\pi = (\pi_1, \ldots, \pi_\kappa)$ is the distribution of X_r, the random variable associated to the root r, and satisfies that all entries are nonnegative and $\sum_i \pi_i = 1$. If $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ we interpret these entries as the probabilities that an arbitrary site in the DNA sequence at the root is occupied by the corresponding base, or, equivalently, as the frequencies with which we would expect to observe these bases in a sequence at the root. A second set of parameters is associated to the evolutionary process that occurs in every edge. For each edge e we associate a $\kappa \times \kappa$ matrix M_e, called *substitution* or *transition* matrix.

2.3.1 Definition (Substitution or Transition Matrix) A *transition matrix* is a $\kappa \times \kappa$ matrix M_e associated to each edge of a phylogenetic tree. Every entry is the conditional probability $P(x|y, e)$ that the state y at the parent node of e being substituted by the state x at its child, during the evolutionary process along the edge e. Since each row contains the probabilities of the κ possible changes that can occur in an evolutionary process, the rows of M_e sum to 1. These matrices M_e are also called *Markov matrices* or *row stochastic matrices*.

2.3.2 Remark If $\kappa = 4$ and $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ then the (i, j)-entry of M_e stands for the conditional probability that if nucleotide i occurs at one site of the DNA sequence in the parent vertex on the edge e, then nucleotide j occurs at the descendant vertex at the same site. In this case, the transition matrices have the form

$$M_e = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} P(\mathtt{A}|\mathtt{A},e) & P(\mathtt{C}|\mathtt{A},e) & P(\mathtt{G}|\mathtt{A},e) & P(\mathtt{T}|\mathtt{A},e) \\ P(\mathtt{A}|\mathtt{C},e) & P(\mathtt{C}|\mathtt{C},e) & P(\mathtt{G}|\mathtt{C},e) & P(\mathtt{T}|\mathtt{C},e) \\ P(\mathtt{A}|\mathtt{G},e) & P(\mathtt{C}|\mathtt{G},e) & P(\mathtt{G}|\mathtt{G},e) & P(\mathtt{T}|\mathtt{G},e) \\ P(\mathtt{A}|\mathtt{T},e) & P(\mathtt{C}|\mathtt{T},e) & P(\mathtt{G}|\mathtt{T},e) & P(\mathtt{T}|\mathtt{T},e) \end{pmatrix} \end{array}.$$

The probabilistic model we have described is a Markov process in the following sense.

2.3.3 Definition (Markov process) A *Markov process* is a random phenomenon that complies the Markov property which says that "the process has no memory". This means that the probability distribution of the future value of a variable depends on its present value, but is independent from the history of the variable.

In other words, in a Markov process the probability that a change occurs in a particular state given that the system is in state i is the same as the probability of the same change, given the entire history of states ending in state i.

The model we have explained above of molecular evolution occurring through random nucleotides substitutions satisfies the Markov assumption, since the probabilities of the possible state changes on any given edge depend only on the state at the ancestral node. Besides, we only have observations of the random variables at the leaves so ours is a *hidden* Markov process.

According to the shape of the transition matrices one has different *models*.

2.3.4 Definition (General Markov model) The *General Markov model* (GMM) is the model with no restriction neither in π nor the transition matrices M_e. Then $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ such that $\sum_i \pi_i = 1$, and

$$M_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ j_e & k_e & l_e & m_e \\ n_e & o_e & p_e & q_e \end{pmatrix}, \text{ where } \begin{cases} a_e + b_e + c_e + d_e = 1, \\ e_e + f_e + g_e + h_e = 1, \\ j_e + k_e + l_e + m_e = 1, \\ n_e + o_e + p_e + q_e = 1. \end{cases}$$

This model will be important in the next chapters. Now we present some other models, which are more restrictive than the GMM.

2.3.5 Definition (Jukes-Cantor model) This is the most restricted model since it adds several additional assumptions. At the same time is really simple. First of all it assumes that all bases occurs with equal probability in the ancestral sequence. Therefore the root distribution vector is

$$\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right).$$

It assumes that the probability of any mutation during an evolutionary process is the same, but different to the probability of no mutation. Then the matrices are

$$M_e = \begin{pmatrix} a_e & b_e & b_e & b_e \\ b_e & a_e & b_e & b_e \\ b_e & b_e & a_e & b_e \\ b_e & b_e & b_e & a_e \end{pmatrix}, \quad \text{where } b_e = \frac{1-a_e}{3}. \tag{2.1}$$

2.3.6 Definition (Strand symmetric model) Another model that has a particular interest is the Strand symmetric model that reflects the double strand symmetry of DNA molecules. As we have explained, in the DNA molecule nucleotides are linked in pairs $\mathtt{A}-\mathtt{T}$ and $\mathtt{C}-\mathtt{G}$, so Strand symmetric model contemplates this fact and assumes the following restrictions $j_e = h_e$, $k_e = g_e$, $l_e = f_e$, $m_e = e_e$, $n_e = d_e$, $o_e = c_e$, $p_e = b_e$, $q_e = a_e$ (see Definition 2.3.4), $\pi_A = \pi_T$ and $\pi_C = \pi_G$. Therefore, the matrices are

$$M_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ h_e & g_e & f_e & e_e \\ d_e & c_e & b_e & a_e \end{pmatrix},$$

with sum of rows equal to 1.

2.3.7 Definition (Kimura models) Kimura 3-parameter is a model introduced by *M. Kimura* in 1981 [Kim81].This model assumes that the base frequencies at the root are equal. It is more general than *Jukes Cantor model* since it has three free parameters. The transition matrices are

$$M_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ b_e & a_e & d_e & c_e \\ c_e & d_e & a_e & b_e \\ d_e & c_e & b_e & a_e \end{pmatrix},$$

where $a_e = 1 - b_e - c_e - d_e$ and the root distribution is assumed to be uniform.

A more restricted model is the Kimura 2-parameter model, which adds another assumption, $b_e = d_e$.

2.3.8 Example The following Figure 2.6 represents the modeled phylogenetic tree, where the $X_i's$ are random variables associated to the leaves, $M_i's$ are the transition matrices, and π_r is the root distribution. Let $\mathcal{K}$ be the set of possible states for X_i. As we have seen, the parameters of a statistical model in a phylogenetic tree depend on the chosen model on the tree. For instance, if the model of this tree is the General Markov model, we have 3×4 free parameters for each substitution matrix and 3 free parameters for the vector π_r. Therefore, this model has $3 \cdot 4 \cdot 6 + 3 = 75$ free parameters.

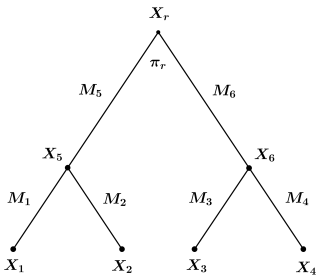

Figure 2.6: Statistical model on a rooted phylogenetic 4-leaved tree.

2.4 Joint distribution

We fix now an evolutionary model $\mathcal{M}$ on a n-leaf tree $\mathcal{T}$ rooted at a node r. Let $\mathcal{K}$ ($\kappa = |\mathcal{K}|$) be the set of states of the random variables X_i associated to the nodes. In what follows we can describe how to compute the joint probability of observing states $x_1, x_2, \ldots, x_n$ at the leaves according to the Markov process we have described.

We denote by $p_{x_1,\ldots,x_n}$ the joint distribution at the leaves of a rooted phylogenetic tree $\mathcal{T}$, which means that $p_{x_1,\ldots,x_n}$ is the probability that the random variables $X_1, \ldots, X_n$ of the leaves take the states $x_1, \ldots, x_n$:

$$p_{x_1,\ldots,x_n} = \text{Prob}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

We define P as a κ^n-dimensional vector, whose entries are the joint probabilities $p_{x_1\ldots x_n}$,

$$P = (p_{x_1,\ldots,x_n})_{x_1,\ldots,x_n \in \mathcal{K}}.$$

Since the evolutionary processes follow a Markov process they are independent and just depend on a common node we can express $p_{x_1,\ldots,x_n}$ in terms of the entries of the substitution matrices.

$$p_{x_1,\ldots,x_n} = \sum_{x_r,(x_v)_{v\in \text{Int}(\mathcal{T})}} \prod_{e\in E(\mathcal{T})} M_e(x_{a(e)}, x_{d(e)}), \tag{2.2}$$

where $x_r \in \mathcal{K}$ is a state of the root, $x_{a(e)} \in \mathcal{K}$ is a state of the parent node of the edge e, and $x_{d(e)} \in \mathcal{K}$ is the state of the descendant node of the edge e. If e is a terminal edge ending at the leaf i then $x_{d(e)} = x_i$. Every entry of P can be seen as a polynomial with the parameters of the model $\mathcal{M}$ as variables.

2.4.1 Example We compute now the joint distribution p_{x_1,x_2,x_3,x_4} of the tree represented in Figure 2.3.8. Using equation (2.2) we get

$$p_{x_1,x_2,x_3,x_4} = \sum_{x_r\in\mathcal{K}} \sum_{x_5\in\mathcal{K}} \sum_{x_6\in\mathcal{K}} \pi_{x_r} \cdot M_5(x_r, x_5) \cdot M_1(x_5, x_1) \cdot M_2(x_5, x_2) \cdot \\ \cdot M_6(x_r, x_6) \cdot M_3(x_6, x_3) \cdot M_4(x_6, x_4).$$

2.5 Phylogenetic invariants

It is known that there exists many algebraic relations among the entries of the joint distribution P (see [Cas12], [CFS10], [Eri05] and [AR07]). To study these relations from an algebraic point of view we regard $P = (p_{x_1,\dots,x_n})_{x_1,\dots,x_n}$ as a vector in $\mathbb{C}^{\kappa^n}$.

Let $\mathcal{T}$ be a rooted phylogenetic tree and $\mathcal{M}$ an evolutionary model. Let r be the root of $\mathcal{T}$ and X_i the random variables associated to the n leaves that can take κ different states from $\mathcal{K}$. Unless noted otherwise be kept notation throughout the work.

Since components of P are polynomials in the model parameters we can associate to the tree a polynomial map $\varphi_{\mathcal{T}}^{\mathcal{M}} : \mathbb{R}^d \to \mathbb{R}^{\kappa^n}$ that maps any d-tuple of parameters to a distribution vector of the κ^n possible observations at the leaves.

More precisely, we define the map

$$\begin{array}{rcl} \varphi_{\mathcal{T}}^{\mathcal{M}} : \mathbb{C}^d & \longrightarrow & \mathbb{C}^{\kappa^n} \\ (\pi, \{M_e\}_{e\in E(\mathcal{T})}) & \mapsto & P = (p_{x_1,x_1,\dots,x_1}, p_{x_1,x_1,\dots,x_2}, p_{x_1,x_1,\dots,x_3}, \dots, p_{x_\kappa,x_\kappa,\dots,x_\kappa}), \end{array} \tag{2.3}$$

where d is the number of free parameters of the model and each coordinate $p_{x_1\dots x_n}$ is expressed in terms of the root distribution π and the transition matrices M_e according to the expression (2.2).

2.5.1 Remark Notice that to read the parameters as probabilities, we should restrict to nonnegative real numbers. Analogously, the points in the image of $\varphi_{\mathcal{T}}^{\mathcal{M}}$ represent join distribution only if they lie in the standard $(\kappa^n - 1)$-simplex. However, in order to use techniques from algebraic geometry, we abandon temporally these restrictions and work over the complex field.

We will consider *complex parameters* and complex parametrization map in general, but we will refer to *stochastic parameters* to the ones coming from the original probabilistic model (that is, all the components of π and the entries of the transition matrices M_i are ≥ 0).

2.5.2 Remark [AR03] It can be proved that if we root the tree $\mathcal{T}$ at a different node r' (call this tree $\mathcal{T}'$) then, for any set of parameters $\pi, \{M_e\}_{e\in E(\mathcal{T})}$, there exist parameters $\pi', \{M'_e\}_{e\in E(\mathcal{T}')}$ such that

$$\varphi_{\mathcal{T}}^{\mathcal{M}}(\pi, \{M_e\}_e) = \varphi_{\mathcal{T}'}^{\mathcal{M}}(\pi', \{M'_e\}_e).$$

This means that the root position cannot be inferred from the joint distribution at the leaves. This phenomenon is usually known as the *non-identifiability* of the root position. For this reason, from now on, we will deal with unrooted trees when addressing the problem of topology reconstruction.

We construct now an algebraic variety in $\mathbb{C}^{\kappa^n}$ that contains the set of image points of $\varphi_{\mathcal{T}}^{\mathcal{M}}$. But first, we recall some basic results from Algebraic Geometry.

2.5.3 Definition (Algebraic variety) An *algebraic variety* $\mathcal{V}$ in $\mathbb{C}^n$ is the set of solutions to a system of polynomial equations: $\mathcal{V} = \{p \in \mathbb{C}^n \mid f_1(p) = 0, \ldots, f_r(p) = 0\}$ for some polynomials $f_1, \ldots, f_r$ on n variables.
The set of algebraic varieties in $\mathbb{C}^n$ form the closed sets of the *Zariski topology.*

2.5.4 Lemma Given any subset S of in $\mathbb{C}^n$ the set of polynomials vanishing on all the points in S forms an ideal $I(S)$ in $\mathbb{C}[x_1, \ldots x_n]$ called the *ideal of* S.

2.5.5 Theorem (Hilbert's Basis Theorem) Every ideal $I \subseteq \mathbb{C}[x_1, ..., x_n]$ can be generated by a finite set of polynomials $f_1, \ldots, f_m$.

We return to phylogenetic trees. Let $\mathcal{T}$ be a phylogenetic tree with n leaves and let T be its topology with the notation kept as in (2.3).

2.5.6 Definition (Phylogenetic variety) The *phylogenetic variety* associated to a tree T and a model $\mathcal{M}$, denoted by $\mathcal{V}_{\mathcal{M}}(T)$, is the smallest algebraic variety containing the image $\operatorname{Im}\varphi_T^{\mathcal{M}}$. Equivalently $\mathcal{V}_{\mathcal{M}}(T)$ is the Zariski closure of $\operatorname{Im}\varphi_T^{\mathcal{M}}$.

2.5.7 Remark The image set $\operatorname{Im}\varphi_T^{\mathcal{M}}$, itself, is not, in general, an algebraic variety. But it defines a dense open subset in the smallest algebraic variety $\mathcal{V}_{\mathcal{M}}(T)$ containing it, in the Zariski topology. The ideal $I(\operatorname{Im}\varphi_T^{\mathcal{M}})$ coincides with the ideal of the variety $\mathcal{V}_{\mathcal{M}}(T)$. We will denote it by $I_{\mathcal{M}}(T)$. As pointed out in Remark 2.5.2, this variety is independent of the node chosen as root in $\mathcal{T}$.

2.5.8 Definition (Invariants, phylogenetic invariants) Given a tree topology T with n leaves and an evolutionary model $\mathcal{M}$, the polynomials in $I_{\mathcal{M}}(T)$ are called *invariants* of T. If f is a polynomial in $I_{\mathcal{M}}(T)$ which does not belong to $I_{\mathcal{M}}(T')$ for some other tree topology T' on n leaves, then f is called a *phylogenetic invariant* of T.

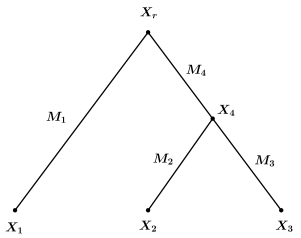

Figure 2.7: 3-leaf rooted tree.

2.5.9 Example Let $\mathcal{T}$ be the 3-leaf tree of Figure 2.7. Suppose $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and every transition matrix M_e associated to the edges is a Jukes cantor matrix. Let $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $M_e = \begin{pmatrix} b_e & a_e & a_e & a_e \\ a_e & b_e & a_e & a_e \\ a_e & a_e & b_e & a_e \\ a_e & a_e & a_e & b_e \end{pmatrix}$.

We compute now the joint distribution at the leaves. Since the parametrization is symmetric under renaming bases we can arrange these joint distributions in 5 groups.

First of all suppose we observe the same state $x \in \mathcal{K}$ at the three leaves. Then

$$p_{x,x,x} = \frac{1}{4} b_1 \cdot (b_4 b_2 b_3 + 3 a_4 a_2 a_3) + \frac{3}{4} a_1 \cdot (a_4 b_2 b_3 + 2 a_4 a_2 a_3 + b_4 a_2 a_3)\,.$$

If the observation at leaves 1 and 2 is x but the one at leaf 3 is y, with $x \neq y$, then

$$\begin{aligned} p_{x,x,y} =& \frac{1}{4} b_1 \cdot (b_4 b_2 a_3 + a_4 a_2 b_3 + 2 a_4 a_2 a_3) + \frac{1}{4} a_1 \, (a_4 b_2 a_3 + b_4 a_2 b_3 + 2 a_4 a_2 a_3) + \\ &+ \frac{2}{4} a_1 \, (a_4 b_2 a_3 + a_4 a_2 b_3 + a_4 a_2 a_3 + b_4 a_2 a_3)\,. \end{aligned}$$

Otherwise if the observations are equal in the second and third leaves (or in the first and third) but different from the first (or second) leaf, then the joint distributions are.

$$\begin{aligned} p_{x,y,y} =& \frac{1}{4} b_1 \cdot (b_4 a_2 a_3 + a_4 b_2 b_3 + 2 a_4 a_2 a_3) + \frac{1}{4} a_1 \, (b_4 b_2 b_3 + 3 a_4 a_2 a_3) + \\ &+ a_1 \, (a_4 b_2 b_3 + b_4 a_2 a_3 + 2 a_4 a_2 a_3)\,, \end{aligned}$$

$$\begin{aligned} p_{x,y,x} =& \frac{1}{4} b_1 \cdot (b_4 a_2 b_3 + a_4 b_2 a_3 + 2 a_4 a_2 a_3) + \frac{1}{4} a_1 \, (a_4 a_2 b_3 + b_4 b_2 a_3 + 2 a_4 a_2 a_3) + \\ &+ \frac{2}{4} a_1 \, (a_4 a_2 b_3 + a_4 b_2 a_3 + a_4 a_2 a_3 + b_4 a_2 a_3)\,. \end{aligned}$$

Finally, if the three observed states are different this joint probability can be computed as

$$\begin{aligned} p_{x,y,z} = & \frac{1}{4} b_1 \cdot (b_4 a_2 a_3 + a_4 a_2 a_3 + a_4 b_2 a_3 + a_4 a_2 b_3) + \frac{1}{4} a_1 \cdot (b_4 b_2 a_3 + a_4 a_2 b_3 + 2 a_4 a_2 a_3) + \\ & + \frac{1}{4} a_1 \, (b_4 a_2 b_3 + a_4 b_2 a_3 + 2 a_4 a_2 a_3) + \frac{1}{4} a_1 \, (b_4 a_2 a_3 + a_4 b_2 a_3 + a_4 a_2 b_3 + a_4 a_2 a_3)\,. \end{aligned}$$

Therefore we have seen that there are many linear relations among these joint distributions, which are invariants of this model.

$$\begin{aligned} p_{\mathtt{AAA}} = p_{\mathtt{CCC}} = p_{\mathtt{GGG}} = p_{\mathtt{TTT}}, \\ p_{\mathtt{AAC}} = p_{\mathtt{AAG}} = \cdots = p_{\mathtt{TTC}} = p_{\mathtt{TTG}}, \\ p_{\mathtt{CAA}} = p_{\mathtt{GAA}} = \cdots = p_{\mathtt{CTT}} = p_{\mathtt{GTT}}, \\ p_{\mathtt{ACA}} = p_{\mathtt{AGA}} = \cdots = p_{\mathtt{TCT}} = p_{\mathtt{TGT}}, \\ p_{\mathtt{ACG}} = p_{\mathtt{ACT}} = \cdots = p_{\mathtt{TAC}} = p_{\mathtt{TCG}}. \end{aligned}$$

In the next section we will see how to produce phylogenetic invariants for the GMM.

2.6 Flattening

In this section we explain how we can see the joint distribution vector P as a matrix which depends on P and a bipartition of the leaves. We also describe the phylogenetic invariants that we obtain from this matrix.

2.6.1 Definition Given a set X a *bipartition* $A \mid B$ *of* X are two sets A and B, with $|A|, |B| \geq 2$ such that $X = A \cup B$ and $A \cap B = \emptyset$.

2.6.2 Definition (Flattening) Let $A|B$ be a partition of the leaves of a tree $\mathcal{T}$ and let $\tilde{X}_A$ and $\tilde{X}_B$ be the random variables associated to A and B. Then $\tilde{X}_A$ and $\tilde{X}_B$ can take $a := \kappa^{|A|}$ and $b := \kappa^{|B|}$ states respectively. Given a vector $P \in \mathbb{C}^{\kappa^n}$ we define the *flattening* $Flatt_{A|B}(P)$ as the $a \times b$ matrix whose entries are the joint distributions of the observations of $\tilde{X}_A$ and $\tilde{X}_B$:

$$Flatt_{A|B}(P) = \begin{array}{c} \\ \text{States of } \tilde{X}_A \end{array} \begin{array}{c} \text{States of } \tilde{X}_B \\ \begin{pmatrix} p_{u_1v_1} & p_{u_1v_2} & \cdots & p_{u_1v_b} \\ p_{u_2v_1} & p_{u_2v_2} & \cdots & p_{u_2v_b} \\ \vdots & \vdots & \ddots & \vdots \\ p_{u_av_1} & p_{u_av_2} & \cdots & p_{u_av_b} \end{pmatrix} \end{array}.$$

2.6.3 Example Let $\mathcal{T}$ be the 4-leaf tree presented at Figure 2.6. Then, the $Flatt_{12|34}(P)$ is the 16×16 matrix:

$$Flatt_{12|34}(P) = \begin{array}{c} \\ \text{States at leaves 1 and 2} \end{array} \begin{array}{c} \text{States at leaves 3 and 4} \\ \begin{pmatrix} p_{\texttt{AAAA}} & p_{\texttt{AAAC}} & p_{\texttt{AAAG}} & \cdots & p_{\texttt{AATT}} \\ p_{\texttt{ACAA}} & p_{\texttt{ACAC}} & p_{\texttt{ACAG}} & \cdots & p_{\texttt{ACTT}} \\ p_{\texttt{AGAA}} & p_{\texttt{AGAC}} & p_{\texttt{AGAG}} & \cdots & p_{\texttt{AGTT}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{\texttt{TTAA}} & p_{\texttt{TTAC}} & p_{\texttt{TTAG}} & \cdots & p_{\texttt{TTTT}} \end{pmatrix} \end{array}.$$

2.6.4 Theorem [AR03] Let $P = \varphi_{\mathcal{T}}(\pi, \{M_e\}_{e \in E(\mathcal{T})})$ where $T = T_{12|34}$. Then the $(\kappa+1) \times (\kappa+1)$ minors of $Flatt_{12|34}(P)$ vanish, equivalently $Flatt_{12|34}(P)$ has rank $\leq \kappa$. Moreover $Flatt_{13|24}(P)$ and $Flatt_{14|23}(P)$ have rank κ^2 for general P.

2.6.5 Remark Theorem 2.6.4 implies that $(\kappa+1) \times (\kappa+1)$ minors of $Flatt_{12|34}(P)$ are phylogenetic invariants for the $T_{12|34}$ tree.

2.7 Tensors

There is a more algebraic way of viewing the joint distribution at the leaves of a phylogenetic tree, which will be really useful in this work.

Let $\mathcal{W} := \mathbb{C}^\kappa$ be a vector space. We identify the canonical basis of $\mathcal{W}$ with the set $\mathcal{K}$. Then the natural basis of $\mathcal{W} \otimes \overset{n)}{\ldots} \otimes \mathcal{W}$ is $\{x_1 \otimes \ldots \otimes x_n\}_{x_1,\ldots,x_n \in \mathcal{K}}$. For instance if $\mathcal{K} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$, the natural basis of $\mathcal{W} \otimes \mathcal{W} \otimes \mathcal{W}$ is $\{\mathtt{A} \otimes \mathtt{A} \otimes \mathtt{A},\ \mathtt{A} \otimes \mathtt{A} \otimes \mathtt{C},\ \ldots,\ \mathtt{T} \otimes \mathtt{T} \otimes \mathtt{T}\}$.

The joint distribution $P = (p_{x_1 \ldots x_n})_{x_1 \ldots x_n \in \mathcal{K}}$ can be thought as a $\kappa \times \overset{n)}{\ldots} \times \kappa$ tensor in $\mathcal{W} \otimes \overset{n)}{\ldots} \otimes \mathcal{W}$ whose coordinates in the natural basis above are $P = (p_{x_1 \ldots x_n})_{x_1 \ldots x_n \in \mathcal{K}}$.

$$P = \sum_{x_1 \ldots, x_n \in \mathcal{K}} p_{x_1,\ldots,x_n} x_1 \otimes \ldots \otimes x_n.$$

For the remainder it will be convenient to write $P(x_1, \ldots, x_n)$ for the component $p_{x_1,\ldots,x_n}$.

Each factor of this tensor product corresponds to one leaf, so in order to make leaves apparent in this tensor product we denote it as $\mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4$ ($\mathcal{W}_i = \mathcal{W}$). If we view the vector of joint distribution P as a tensor in $\mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4$, then the flattening $Flatt_{12|34}(P)$ is the image of P via the isomorphism

$$\begin{array}{ccccc} \mathcal{W}_1 \otimes \mathcal{W}_2 \otimes \mathcal{W}_3 \otimes \mathcal{W}_4 & \cong & Hom(\mathcal{W}_1 \otimes \mathcal{W}_2, \mathcal{W}_3 \otimes \mathcal{W}_4) & \cong & M_{\kappa^2 \times \kappa^2}(\mathbb{C}) \\ P & & \longmapsto & & Flatt_{12|34}(P) \end{array}$$

where $M_{\kappa^2 \times \kappa^2}(\mathbb{C})$ is the space of all $\kappa^2 \times \kappa^2$ matrices with complex entries.

2.7.1 Remark From now on, given a vector $v \in \mathbb{C}^\kappa$, $v(i)$ will be the i-th coordinate of v, $\{e_1, \ldots, e_\kappa\}$ will be the canonical base of $\mathbb{C}^\kappa$ and $\mathbf{1} = (1, \cdots, 1)$. Moreover we will call an *n-tensor* to the tensors $P \in \mathbb{C}^\kappa \otimes \overset{n)}{\ldots} \otimes \mathbb{C}^\kappa$.

We will define now the product of a tensor by a vector or a matrix.

2.7.2 Definition ($P *_i v$, l-th slice, i-th marginalization, $P *_i M$) Given an n-tensor P, an integer $i \in \{1, \ldots n\}$ and a vector $\mathbf{v} \in \mathbb{C}^\kappa$, we define a $(n-1)$-tensor $P *_i \mathbf{v}$ as follows,

$$(P *_i \mathbf{v})(j_1, \ldots, j_{i-1}, j_{i+1}, \ldots, j_n) = \sum_{j_i=1}^{\kappa} v(j_i) P(j_1, \ldots, j_i, \ldots, j_n).$$

We define also the *l-th slice of P in the i-th index* by

$$P_{\ldots l \ldots} = P *_i \mathbf{e}_l,$$

The *i-th marginalization of P* is defined as

$$P_{\ldots + \ldots} = P *_i \mathbf{1}.$$

Given a $\kappa \times \kappa$ matrix M, we define the *n-tensor* $P *_i M$ by

$$(P *_i M)(j_1, \ldots, j_n) = \sum_{l=1}^{\kappa} P(j_1, \ldots, j_{i-1}, l, j_{i+1}, \ldots, j_n) M(l, j_i). \tag{2.4}$$

2.7.3 Remark From now on, we consider the 2-tensors as $\kappa \times \kappa$ matrices via the isomorphism,

$$P = \sum P(j_1, j_2) e_{j_1} \otimes e_{j_2} \leftrightarrow (P(j_1, j_2))_{j_1, j_2},$$

where the rows of the matrix are indexed by the first component and columns by the second.

We illustrate these definitions with an example.

2.7.4 Example Let P be a complex 3-tensor in $\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \mathbb{C}^2$ whose components are,

$$P(0,0,0) = 0.01, \quad P(0,0,1) = 0.21, \quad P(0,1,0) = 0.3, \quad P(1,0,0) = 0.125,$$
$$P(1,1,0) = 0.09, \quad P(1,0,1) = 0.13, \quad P(0,1,1) = 0.11, \quad P(1,1,1) = 0.25.$$

And let $\mathbf{v} = (\frac{1}{4}, \frac{3}{4}) \in \mathbb{C}^2$. The entries of the 2-dimensional tensor $\bar{P} = P *_2 v$ are

$$\bar{P}(0,0) = \frac{1}{4} \cdot 0.01 + \frac{3}{4} \cdot 0.3 = 0.2275,$$
$$\bar{P}(0,1) = \frac{1}{4} \cdot 0.21 + \frac{3}{4} \cdot 0.11 = 0.135,$$
$$\bar{P}(1,0) = \frac{1}{4} \cdot 0.125 + \frac{3}{4} \cdot 0.09 = 0.09875,$$
$$\bar{P}(1,1) = \frac{1}{4} \cdot 0.13 + \frac{3}{4} \cdot 0.025 = 0.39875.$$

The 1-*th slice* of P in the third index is $\tilde{P} = P_{\cdot\cdot 1} = P *_3 \mathbf{e}_1$, and has components $\tilde{P}(x,y) = (P *_3 \mathbf{e}_1)(x,y) = P(x,y,1)$, i.e.

$$\tilde{P}(0,0) = 0.01, \ \tilde{P}(0,1) = 0.3, \ \tilde{P}(1,0) = 0.125, \ \tilde{P}(1,1) = 0.09.$$

And the *i-th marginalization* $\hat{P} = P_{+\cdot\cdot} = P *_1 \mathbf{1}$ has entries $\hat{P}(x,y) = P(0,x,y) + P(1,x,y)$, i.e.

$$\hat{P}(0,0) = 0.126, \ \hat{P}(0,1) = 0.34, \ \hat{P}(1,0) = 0.39, \ \hat{P}(1,1) = 0.135.$$

Finally, if $M = \begin{pmatrix} 0.25 & 0.75 \\ 0.55 & 0.45 \end{pmatrix}$, the components of $\tilde{P} = P *_2 M$ are

$$P(0,0,0) = 0.01 \cdot 0.25 + 0.3 \cdot 0.55 = 0.853,$$
$$P(0,0,1) = 0.21 \cdot 0.25 + 0.11 \cdot 0.55 = 0.113,$$
$$P(0,1,0) = 0.01 \cdot 0.75 + 0.3 \cdot 0.45 = 0.143,$$
$$P(1,0,0) = 0.125 \cdot 0.25 + 0.09 \cdot 0.55 = 0.08,$$
$$P(1,1,0) = 0.125 \cdot 0.75 + 0.09 \cdot 0.45 = 0.134,$$
$$P(1,0,1) = 0.13 \cdot 0.25 + 0.25 \cdot 0.55 = 0.17,$$
$$P(0,1,1) = 0.21 \cdot 0.75 + 0.11 \cdot 0.45 = 0.207,$$
$$P(1,1,1) = 0.13 \cdot 0.75 + 0.25 \cdot 0.45 = 0.21.$$

3 Theoretical results

3.1 Some operations with tensors

In this section we show some technical results related to marginalizations and slices of tensors that arise from stochastic parameters of the general Markov model on a tree $\mathcal{T}$.

3.1.1 Lemma Let P be a 3-tensor in the image of parameters for the General Markov model, $P = \varphi(\pi, \{M_1, M_2, M_3\})$, where $\mathcal{T}$ is the 3-leaf tree of Figure 2.3. Then, the three possible marginalization of P are given by

$$\begin{array}{l} P_{..+} = P *_3 1 = M_1^t diag(\pi) M_2, \\ P_{.+.} = P *_2 1 = M_1^t diag(\pi) M_3, \\ P_{+..} = P *_1 1 = M_2^t diag(\pi) M_3. \end{array} \tag{3.1}$$

Proof We compute the 3rd marginalization of P, $P_{..+}$. By definition

$$P_{..+}(j_1, j_2) = (P *_3 \mathbf{1})(j_1, j_2) = \sum_{j_3=1}^{\kappa} 1 \cdot P(j_1, j_2, j_3). \tag{3.2}$$

Since $P = \psi(\pi, \{M_1, M_2, M_3\})$, we have

$$P(j_1, j_2, j_3) = \sum_{i=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) M_3(i, j_3). \tag{3.3}$$

Substituting in (3.2) we obtain

$$\begin{aligned} P_{..+}(j_1, j_2) = \sum_{j_3=1}^{\kappa} 1 \cdot P(j_1, j_2, j_3) = \sum_{j_3=1}^{\kappa} \sum_{i=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) M_3(i, j_3) = \\ = \sum_{i=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) \left(\sum_{j_3=1}^{k} M_3(i, j_3) \right) = \\ = \sum_{i=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2). \end{aligned}$$

The elements of this sum are written in terms of the vector π, the j_1-th column of M_1 and the j_2-th column of M_2. Equivalently this is the product of the j_1-th row of M_1^T, the diagonal matrix $diag(\pi)$ and the j_2-th column of M_2. Therefore this expression for all j_1 and j_2 becomes

$$P_{..+} = P *_3 1 = M_1^T diag(\pi) M_2. \tag{3.4}$$

Similarly we can compute the expressions of $P_{.\cdot+}$ and $P_{+..}$.

□

3.1.2 Lemma Under the same conditions of Lemma 3.1.1, the slices of P are

$$\begin{aligned} P_{..i} &= P *_3 e_i = M_1^T diag(M_3 e_i) diag(\pi) M_2, \\ P_{.i.} &= P *_2 e_i = M_1^T diag(M_2 e_i) diag(\pi) M_3, \\ P_{i..} &= P *_1 e_i = M_2^T diag(M_1 e_i) diag(\pi) M_3. \end{aligned} \tag{3.5}$$

Proof We check the expression of $P_{..i}$ in a similar way to previous Lemma. Again by definition, using (3.3) and taking into account that e_i is the i-th canonical vector, we have

$$\begin{aligned} P_{..i}(j_1, j_2) = (P *_3 e_i)(j_1, j_2) = \sum_{j_3=1}^{\kappa} e_i(j_3) P(j_1, j_2, j_3) = P(j_1, j_2, i) = \\ = \sum_{m=1}^{\kappa} \pi_m M_1(m, j_1) M_2(m, j_2) M_3(m, i). \end{aligned}$$

Similar arguments to those on Lemma 3.1.1 show that elements of this sum are the vector π, the i-th column of M_3 (i.e. the vector $M_3 e_i$) and the j_1-th column of M_1 and j_2-th column of M_2. This is the same as the product of the j_1-th row of M_1^T, the matrices $diag(\pi)$, $diag(M_3 e_i)$ and the j_2-th column of M_2. Finally for all pairs j_1, j_2 this expression becomes

$$\begin{aligned} P_{..i} = & \begin{pmatrix} M_1(1,1) & M_1(2,1) & M_1(3,1) & M_1(4,1) \\ M_1(1,2) & M_1(2,2) & M_1(3,2) & M_1(4,2) \\ M_1(1,3) & M_1(2,3) & M_1(3,3) & M_1(4,3) \\ M_1(1,4) & M_1(2,4) & M_1(3,4) & M_1(4,4) \end{pmatrix} \times \\ & \times \begin{pmatrix} M_3(1,i) \cdot \pi_1 & 0 & 0 & 0 \\ 0 & M_3(2,i) \cdot \pi_2 & 0 & 0 \\ 0 & 0 & M_3(3,i) \cdot \pi_3 & 0 \\ 0 & 0 & 0 & M_1(4,i) \cdot \pi_4 \end{pmatrix} \times \\ & \times \begin{pmatrix} M_2(1,1) & M_2(1,2) & M_3(1,3) & M_4(1,4) \\ M_2(2,1) & M_2(2,2) & M_3(2,3) & M_4(2,4) \\ M_2(3,1) & M_2(3,2) & M_3(3,3) & M_4(3,4) \\ M_2(4,1) & M_2(4,2) & M_3(4,3) & M_4(4,4) \end{pmatrix} = \\ = & P *_3 e_i = M_1^T diag(M_3 e_i) diag(\pi) M_2. \end{aligned}$$

Using the same arguments we find analogous expressions for $P_{.i.}$ and $P_{i..}$.

□

3.1.3 Example Consider the 3-leaf tree of Figure 2.3, $\mathcal{K} = \{0, 1\}$ and the Jukes Cantor model, with matrices $M_e = \begin{pmatrix} a_e & b_e \\ b_e & a_e \end{pmatrix}$ and $\pi = (\pi_0, \pi_1) = \left(\frac{1}{2}, \frac{1}{2}\right)$. Suppose

$$\begin{array}{lll} a_1 = 0.61, & a_2 = 0.7, & a_3 = 0.58, \\ b_1 = 0.39, & b_2 = 0.3, & b_3 = 0.42. \end{array}$$

Then,

$$\begin{array}{l} p_{111} = p_{000} = \pi_0(a_1a_2a_3) + \pi_1(b_1b_2b_3) = 0.1484, \\ p_{101} = p_{010} = \pi_0(a_1b_2a_3) + \pi_1(b_1a_2b_3) = 0.1104, \\ p_{110} = p_{001} = \pi_0(a_1a_2b_3) + \pi_1(b_1b_2a_3) = 0.1236, \\ p_{011} = p_{100} = \pi_0(b_1a_2a_3) + \pi_1(a_1b_2b_3) = 0.1176. \end{array}$$

We define the tensor P that comes from this modeled tree as $P(i, j, k) = p_{i,j,k}$ As an example we compute now a marginalization and a slice of P.

$\tilde{P} = P_{..+} = P *_1 \mathbf{1}$ has components

$$\begin{array}{ll} \tilde{P}(0,0) = 0.226, & \tilde{P}(0,1) = 0.234, \\ \tilde{P}(1,0) = 0.234, & \tilde{P}(1,1) = 0.266. \end{array}$$

And $M_2^T diag(\pi) M_3 = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix} \cdot \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.58 & 0.42 \\ 0.42 & 0.58 \end{pmatrix} = \begin{pmatrix} 0.266 & 0.234 \\ 0.234 & 0.266 \end{pmatrix}$, which corresponds to the entries of tensor $\tilde{P}$ seen as a matrix, see Remark 2.7.3

Also $\hat{P} = P_{.1.} = P *_2 \boldsymbol{e_1}$ has coordinates,components

$$\begin{array}{ll} \hat{P}(0,0) = 0.1484, & \hat{P}(0,1) = 0.1236, \\ \hat{P}(1,0) = 0.1176, & \hat{P}(1,1) = 0.1104. \end{array}$$

And finally,

$$M_1^T diag(M_2 e_1) diag(\pi) M_3 = \begin{pmatrix} 0.61 & 0.39 \\ 0.39 & 0.61 \end{pmatrix} \cdot \begin{pmatrix} 0.7 & 0 \\ 0 & 0.3 \end{pmatrix} \cdot \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix} \cdot \begin{pmatrix} 0.58 & 0.42 \\ 0.42 & 0.58 \end{pmatrix} =$$
$$= \begin{pmatrix} 0.1484 & 0.1236 \\ 0.1176 & 0.1104 \end{pmatrix},$$

which is also equal to the matrix associated to $\hat{P}$.

The above marginalizations extend naturally to tensors that come from 4-leaf trees.

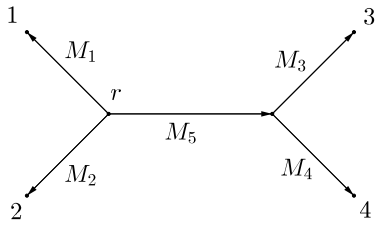

Figure 3.1: Rooted 4-leaf tree $T_{12|34}$.

3.1.4 Corollary Let P be a tensor arising from parameters of the GM(κ) model on $\mathcal{T}$ with tree topology $T_{12|34}$, see Figure 3.1, $P = \varphi_{\mathcal{T}_{12|34}}(\pi; M_1, M_2, M_3, M_4, M_5)$. Then the double marginalizations $P_{+..+}$, $P_{+.+.}$, $P_{.+.+}$ and $P_{.++.}$ can be computed in terms of matrices as follows,

$$\begin{aligned} P_{+..+} &= M_2^T diag(\pi) M_5 M_3, \\ P_{+.+.} &= M_2^T diag(\pi) M_5 M_4, \\ P_{.+.+} &= M_1^T diag(\pi) M_5 M_3, \\ P_{.++.} &= M_1^T diag(\pi) M_5 M_4. \end{aligned}$$

Proof In order to compute all these expressions we need to marginalize the tensor over two different positions. We do the case $P_{+..+}$ and the others are analogous. Firstly we compute $\bar{P} = P_{...+}$ and then we compute $\bar{P}_{+..}$.

$$\begin{aligned} \bar{P} = P_{...+}(j_1, j_2, j_3) &= \sum_{j_4=1}^{\kappa} 1 \cdot P(j_1, j_2, j_3, j_4) = \\ &= \sum_{j_4=1}^{\kappa} \sum_{i=1}^{\kappa} \sum_{m=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) M_5(i, m) M_3(m, j_3) M_4(m, j_4) = \\ &= \sum_{i=1}^{\kappa} \sum_{m=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) M_5(i, m) M_3(m, j_3) \left(\sum_{j_4=1}^{k} M_4(i, j_4) \right) = \\ &= \sum_{i=1}^{\kappa} \sum_{m=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) M_5(i, m) M_3(m, j_3) = \\ &= \sum_{i=1}^{\kappa} \pi_i M_1(i, j_1) M_2(i, j_2) (M_5 M_3)(i, j_3), \end{aligned}$$

which is the tensor that arises from the 3-leaf tree presented in Figure 3.2 with matrices M_1, M_2 and $M_5 M_3$, i.e. $\bar{P}_{T_3} = \varphi_{\mathcal{T}}(\pi, \{M_1, M_2, M_5 M_3\})$. Then, by Lemma 3.1.1, we have,

$$P_{+..+} = \bar{P}_{+..} = M_2^T diag(\pi) M_5 M_3. \tag{3.6}$$

□

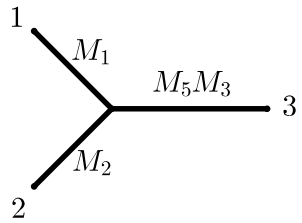

Figure 3.2: 3-leaf tree with transition matrices M_1, M_2 and $M_5 M_3$.

In the following Lemma we will see how, given a tensor in the image of $\varphi_{\mathcal{T}}$ for a 4-leaf tree $\mathcal{T}$, we can produce a new tensor still in Im$\varphi_{\mathcal{T}}$. This is done by multiplying the original tensor with a matrix (in the sense of (2.4)), which has the effect of changing the transition matrix of an exterior edge of the tree.

3.1.5 Remark Notice that, given two $\kappa \times \kappa$ matrices M and N we have

$$N(x,y) = \sum_{l=1}^{\kappa} M(x,l)(M^{-1}N)(l,y), \tag{3.7}$$

for any x, y if M is non singular.

3.1.6 Lemma Let P be a 4-tensor for the general Markov model, $P = \varphi_{\mathcal{T}}(\pi; M_1, \ldots, M_5)$. If M_i is non singular for some i, then the tensor $\bar{P} = P *_i (M_i^{-1}M)$ is the image of the same parameters as P except for M_i which has been replaced by M.

Proof Suppose the tensor P arises from $T_{12|34}$. We can assume $i = 1$ without loss of generality.

$$P *_1 (M_1^{-1}M)(j_1,j_2,j_3,j_4) = \sum_{l=1}^{\kappa} P(l,j_2,j_3,j_4)(M_1^{-1}M)(l,j_1) =$$
$$= \sum_{l=1}^{\kappa}\sum_{m=1}^{\kappa}\sum_{h=1}^{\kappa} \pi_m M_1(m,l)M_2(m,j_2)M_5(m,h)M_3(h,j_3)M_4(h,j_4)(M_1^{-1}M)(l,j_1).$$

By the (3.7) this is equal to

$$\sum_{m=1}^{\kappa}\sum_{h=1}^{\kappa} \pi_m M(m,j_1)M_2(m,j_2)M_5(m,h)M_3(h,j_3)M_4(h,j_4),$$

which is the expression for the position (j_1,j_2,j_3,j_4) of the tensor that arises from $T_{12|34}$ where M_1 has been substituted by M. The computations for $i = 2,3,4$ are equivalent.

□

3.2 Stochasticity Conditions

In this section we will discuss some theoretical results that will allow us to provide some conditions to ensure that a tensor of a joint distribution comes from stochastic parameters.

3.2.1 Definition A set $\{\pi, \{M_e\}_{e\in E(T)}\}$ of stochastic parameters for GM model on a tree T with root r is called *nonsingular* if

(i) At every node j the distribution of the random variable X_j has no zero entry.

(ii) The matrix M_e of every edge e is nonsingular.

3.2.2 Remark For stochastic parameters the condition (i) of the previous definition is equivalent to requiring that the root distribution π_r has no zero entry (assuming (ii)).

The following result has been proved in [ART12]. As we do not use it specifically, we do not include the proof here.

3.2.3 Theorem [ART12] Let P be a (either real or complex) 3-tensor. P arises from nonsingular parameters for the general Markov model with κ parameters on the 3-leaf tree if and only if for some $i \in \{1,2,3\}$ the following conditions hold:

(i) $f_i(P;x) \neq 0$ for an arbitrary vector x,

(iii) $det(P *_i 1) \neq 0$ for $i = 1,2,3$,

where $f_i(P;x) = \det H_x\left((\det(P *_i x))\right)$ and H_x denotes the Hessian operator.

We want to find a similar characterization of P for stochastic parameters. That is, we want to find some conditions that allow us to distinguish when a tensor P is the image of positive real parameters.

3.2.4 Theorem (a) Let $P = \varphi_{\mathcal{T}}(\pi, \{M_1, M_2, M_3\})$ be a 3-tensor with π, $\{M_i\}_i$ with real entries. P is the image of nonsingular stochastic parameters for the general Markov model on the 3-leaf tree if and only if its entries are nonnegative and sum to 1, conditions (i), (ii) and (iii) of Theorem 3.2.3 are satisfied, and

(iii) the matrix

$$det(P_{..+})P_{+..}^T adj(P_{..+})P_{.+.} \tag{3.8}$$

is positive definite, and the following matrices are positive semidefinite

$$\begin{array}{ll} det(P_{..+})P_{i..}^T adj(P_{..+})P_{.+.} & for\ i = 1,\dots,\kappa, \\ det(P_{..+})P_{+..}^T adj(P_{..+})P_{.i.} & for\ i = 1,\dots,\kappa, \\ det(P_{+..})P_{.+.} adj(P_{+..})P_{..i}^T & for\ i = 1,\dots,\kappa. \end{array} \tag{3.9}$$

(b) Moreover, P is the image of nonsingular real positive parameters if and only if its entries are positive and sum to one, conditions (i), (ii), and (iii) are satisfied and

(iii') all matrices in (3.8) and (3.9) are positive definite.

In both cases, the nonsingular parameters are unique up to label swapping.

Proof The proof of this Theorem is essentially the same as in [ART12]. Let P be an arbitrary nonnegative 3-tensor whose components sum to 1. Assuming (i) and (ii) and using Theorem 3.2.3, P is the image of nonsingular parameters. We want to see that condition (iv) is equivalent to these parameters being nonnegative. To this aim we are going to see what expressions in (3.8) and (3.9) means.

Let $\bar{P} = P_{+..}P_{..+}^{-1}P_{.+.}$, using expressions proved in Lemma 3.1.1 we compute

$$\begin{aligned} \bar{P} = P_{+..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(\pi) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1^T diag(\pi) M_3) \\ &= M_3^T diag(\pi) M_3. \end{aligned} \tag{3.10}$$

This is a well defined symmetric matrix since $P_{..+}$ is nonsingular. Since M_3 is real, $\bar{P}$ is a positive definite matrix if and only if

$$x^T \bar{P} x = x^T M_3^T \mathrm{diag}(\pi) M_3 x = (M_3 x)^T \mathrm{diag}(\pi)(M_3 x) > 0,\ \forall x \neq 0.$$

Since M_3 is nonsingular it can be understood as a change of basis and hence $\bar{P}$ is positive semidefinite if and only if the entries of $\mathrm{diag}(\pi)$ are all positive. We clear denominators and obtain an algebraic expression multiplying this matrix by the square of the appropriate nonzero determinant. It follows that (3.8) is positive definite if and only if π is positive.

Using expressions of Lemma 3.1.1 and Lemma 3.5, we have

$$\begin{aligned} P_{i..}^T P_{..+}^{-1} P_{.+.} &= (M_2^T diag(M_1 e_i) M_3)^T (M_1^T diag(\pi) M_2)^{-1} (M_1 diag(\pi) M_3) = \\ &= M_3^T diag(\pi) diag(M_1 e_i) M_3. \end{aligned}$$

This matrix is also symmetric, and it is positive semidefinite if and only if the entries of $diag(\pi)diag(M_1 e_i)$ are nonnegative. Since π is a positive vector we need the ith column of M_1 being nonnegative. Using the matrices $P_{+..}^T P_{..+}^{-1} P_{.i.}$ and $P_{..+}^T P_{+..}^{-1} P_{..i}$ we can also impose the conditions of the i-th column of M_2 and M_3 being nonnegative. This proves (a).

If the matrices of (3.8) and (3.9) are positive definite, we can repeat this proof but requiring positiveness of the parameters. This proves (b).

In order to clear denominators and obtain an algebraic expression we multiply all these matrices by the square of the appropriate nonzero determinant which does not change the sign and gives us the expressions (3.8) and (3.9).

□

3.2.5 Remark In the paper [ART12], Theorem 3.2.4 is announced for general tensors P. They assume that $P = \varphi_{\mathcal{T}}(\pi, \{M_1, M_2, M_3\})$ where $\pi, M_1 M_2, M_3$ are complex. But this is not true, we provide here a counterexample. If M_3 is not real, $diag(\pi)$ being positive does not imply $\bar{P} = M^T diag(\pi) M$ (see 3.10) being positive definite. For instance for $\kappa = 2$ if we consider the matrices

$$D = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \; M = \begin{pmatrix} \frac{2+i}{4} & \frac{2-i}{4} \\ \frac{2-i}{4} & \frac{2+i}{4} \end{pmatrix},$$

then

$$M^T D M = \frac{1}{16} \begin{pmatrix} 3 & 5 \\ 5 & 3 \end{pmatrix},$$

is not positive definite. Moreover the reverse implication is neither true. For instance

$$\bar{P} = M^T D M = \begin{pmatrix} 8 & 0 \\ 0 & 8 \end{pmatrix},$$

where

$$D = \begin{pmatrix} -1 & 0 \\ 0 & 4 \end{pmatrix}, \; M = \begin{pmatrix} 2i & -2i \\ 1 & 1 \end{pmatrix}.$$

In this case $\bar{P}$ is positive definite but D is not positive.

Assuming now that an n-tensor P arises from nonsingular parameters on a tree, we would like to give some semialgebraic conditions that are satisfied if and only if P comes from stochastic parameters. If we consider marginalizations of P to three variables and using Theorem 3.2.4, we can give conditions that hold when the root distribution and the product of matrices associated to any path from an interior node to a leaf are stochastic. Nevertheless we need some extra conditions to guarantee matrices of the interior edges being stochastic.

The following result gives us a condition for all parameters of the 12|34 tree being stochastic.

3.2.6 Theorem [ART12] Let P be a 4-tensor. Suppose P arises from nonsingular real parameters for $GM(\kappa)$ model on $T_{12|34}$. If the marginalizations $P_{+\dots}$ and $P_{\dots+}$ arise from stochastic parameters and, moreover, the $\kappa^2 \times \kappa^2$ matrix

$$det(P_{+..+})det(P_{.+.+})Flatt_{13|24}\big(P *_2 (adj(P^T_{+..+})P^T_{.+.+})) *_3 (adj(P_{.+.+})P_{.++.})\big) \quad (3.11)$$

is positive semidefinite, then P arises from stochastic parameters.

Proof The root r is placed at the interior node near leaves 1 and 2 as we can see in Figure 3.1. Let M_i, $i = 1, 2, 3, 4$ be the complex matrix associated to the edges leading to leaves, M_5 the matrix on the internal edge and π the root distribution. The rows of these matrices sum to 1. We define the matrices

$$\begin{array}{l} N_{32} = P^T_{+..+} = M_3^T M_5^T diag(\pi) M_2, \\ N_{31} = P^T_{.+.+} = M_3^T M_5^T diag(\pi) M_1, \\ N_{14} = P_{.++.} = M_1^T diag(\pi) M_5 M_2, \\ N_{13} = P_{.+.+} = M_1^T diag(\pi) M_5 M_3. \end{array} \quad (3.12)$$

We define now a tensor $\bar{P}$ that is arising from the same parameters as P except that M_2 has been replaced by M_1 (see Lemma 3.1.6).

$$\bar{P} = P *_2 N_{32}^{-1} N_{31} = P *_2 M_2^{-1} M_1.$$

Similarly we can define

$$\tilde{P} = \bar{P} *_3 N_{13}^{-1} N_{14} = \bar{P} *_3 M_3^{-1} M_4, \quad (3.13)$$

that is a tensor arising from the same parameters as $\bar{P}$ except that M_4 has been replaced by M_3 (by Lemma 3.1.6).

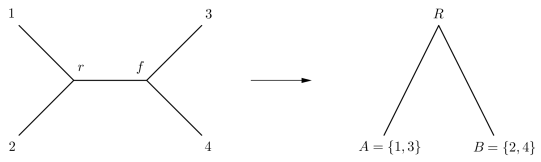

Figure 3.3: *Left:* 4-leaved tree *Right:* Split A={1,3}, B={2,4}.

Let $A = \{1, 3\}$ and $B = \{2, 4\}$be a bipartition of the leaves, and $R = \{r, f\}$ the set of the nodes that are shared by the induced subtrees. Since $13|24$ is not a split of the underlying tree, we can write the $\kappa^2 \times \kappa^2$ flattening matrix of the tensor P as

$$Flatt_{13|24}(P) = M_A^T diag(\pi(R)) M_B$$

where $\pi(R)$ is the distribution of R and M_A and M_B are the transition matrices from R to A and B respectively. Therefore

$$\begin{aligned} M_A =& (P(A = (x_i, x_j)|R = (x_u, x_v)))_{\substack{(x_i,x_j) \in \mathcal{K}^2 \\ (x_u,x_v) \in \mathcal{K}^2}} \\ =& (P(X_1 = x_i, X_3 = x_j | r = x_u, f = x_v))_{x_i,x_j,x_u,x_v \in \mathcal{K}} \end{aligned}$$

Then $P(A = (x_i, x_j)|R = (x_u, x_v)) = M_1(x_u, x_i) M_3(x_v, x_j)$ and we can deduce

$$M_A = (M_1 \otimes M_3).$$

Therefore

$$Flat_{13|24}(P) = (M_1 \otimes M_3)^T D (M_2 \otimes M_4), \tag{3.14}$$

where D is the diagonal matrix that contains the κ^2 entries of $diag(\pi)M_5$. Since $\tilde{P}$ arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4 we can write

$$Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D (M_1 \otimes M_4).$$

Since the 3-marginalization arise from stochastic parameters, M_1 and M_4 are nonsingular and π has positive entries. Thus $M_1 \otimes M_4$ is also nonsingular. All principal minors of $Flat_{13|24}(\tilde{P})$ are nonnegative if and only if $Flat_{13|24}(\tilde{P})$ is positive semidefinite. Then we have to require that D has nonnegative entries and so, since π has positive entries we can ensure that M_5 has nonnegative entries. If we multiply $Flat_{13|24}(\tilde{P})$ by the square of the appropriate nonzero determinant we clear denominators and obtain the algebraic expressions stated in the Theorem.

□

3.2.7 Remark The theoretical results that we have proved in this chapter allow us to provide the algebraic description of the model, given by phylogenetic invariants, together with a semialgebraic description of the points with stochastic sense. In other words, as well as finding polynomials vanishing on the image of the parametrization map, we have found polynomial inequalities sufficient to characterize the stochastic image.

Recall that a subset of $\mathbb{C}$ is called *semialgebraic set* if it is generated by a finite sequence of polynomial equations of the form $P(x_1, ..., x_n) = 0$ and inequalities of the form $Q(x_1, ..., x_n) > 0$, or any finite union of such sets.

The conditions of matrices being positive definite/semidefinite can be expressed as semialgebraic conditions using Sylvester's criterion, which claims that a real symmetric matrix is positive definite (or positive semidefinite) if and only its *leading* principal minors are positive (or nonnegative).

On the other hand, the replacements of inverses in (3.13) by adjoint matrices in (3.11) is not only done in order to have semialgebraic conditions, but also to avoid dealing with badly conditioned matrices.

3.2.8 Remark Recall that the tensor $\tilde{P}$ constructed in (3.13) (in the proof of Theorem 3.2.6) arises from the same parameters that P except that M_2 has been replaced by M_1 and M_3 by M_4. Then $\tilde{P}$ so it is the joint distribution of the tree presented in Figure 3.4. Observe that this tree is symmetric with respect to the interior edge, then we can state the following result.

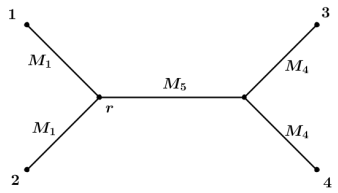

Figure 3.4: 4-leaf tree with transition matrices M_1, M_1, M_2, M_2 and M_5.

3.2.9 Corollary Let P be a 4-tensor whose components sum to 1. Suppose that $P = \varphi_{\mathcal{T}}(\pi, M_1, M_2, M_3, M_4, M_5)$ with $\mathcal{T} = T_{12|34}$. Let $\tilde{P}$ be constructed as in (3.13). Then,

$$Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P}), \tag{3.15}$$

and

$$Flatt_{12|34}(\tilde{P}) \neq Flatt_{13|24}(\tilde{P}). \tag{3.16}$$

In particular

$$\begin{aligned} &det(P_{+..+})det(P_{.+.+})Flatt_{13|24}\big(P *_2 (adj(P^T_{+..+})P^T_{.+.+})) *_3 (adj(P_{.+.+})P_{.++.})\big)) = \\ =&det(P_{+..+})det(P_{.+.+})Flatt_{14|23}\big(P *_2 (adj(P^T_{+..+})P^T_{.+.+})) *_3 (adj(P_{.+.+})P_{.++.})\big) \end{aligned}$$

gives rise to 256 phylogenetic invariants of degree 17.

Proof Using (3.14), and the fact that in $\tilde{P}$ M_2 has been replaced by M_1, and M_3 by M_4, we have

$$Flat_{13|24}(\tilde{P}) = (M_1 \otimes M_4)^T D(M_1 \otimes M_4) = Flat_{14|23}(\tilde{P}). \tag{3.17}$$

In contrast,

$$Flatt_{12|34}(\tilde{P}) = \bar{M}_1^T diag(\pi)\tilde{M}_4,$$

where

$$\bar{M}_1(x_i, (x_j, x_k)) = M_1(x_i, x_j)M_1(x_i, x_k),$$

$$\tilde{M}_4(x_i, (x_j, x_k)) = \sum_{l=1}^{\kappa} M_5(x_i, x_l)M_4(x_l, x_j)M_4(x_l, x_k).$$

which is, in general, not equal to (3.17).

The matrix equality $Flat_{13|24}(\tilde{P}) = Flat_{14|23}(\tilde{P})$ provides 16×16 equalities between entries. By (3.11) these entries are algebraic expressions of the components of P. Moreover, because of (3.16), they are phylogenetic invariants.

Finally, regarding at (3.11), we infer the degree of these expressions in the components of P the two determinants have degree 4 each, which makes degree 8. The components of the tensors $adj(P^T_{+..+})P^T_{.+.+}$ and $adj(P_{.+.+})P_{.++.}$ have degree 4. The $*$ operation adds degrees, so we obtain a tensor of degree $1+4+4=9$ before applying the $Flat_{13|24}(\cdot)$. All together gives a tensor with components of degree $8+9=17$.

□

4 Implementation and results on simulated data

4.1 Numerical and computational issues

Given a 4-tensor that arises from real nonsingular parameters for the general Markov model it is theoretically enough to verify the conditions of the Theorem (3.2.6) to ensure that this tensor comes from stochastic parameters. But, in practice, are these conditions sufficient? And, do they provide new information to recover the topology of a 4-leaf tree using these conditions? In this chapter we will try to answer these questions proposing an equivalent set of sufficient conditions that are useful on approximated data. Since real data are not exactly the image of nonsingular stochastic parameters for the GMM, instead of checking whether all the matrices that we have obtained in Theorem 3.2.4 and Theorem 3.2.6 are symmetric and positive definite (or positive semidefinite) we will determine how far these matrices are from being symmetric positive definite (or positive semidefinite). To compute these distances we will use the next Theorem (see [Hig88] for a complete proof). But first we need a definition.

4.1.1 Definition For any square matrix B its *polar decomposition* is the unique matrix decomposition of the form $B = UH$ where $U^TU = Id$ and $H = H^T$ is positive semidefinite.

4.1.2 Theorem [Hig88] Let $A \in \mathbb{R}^{n\times n}$, and let $B = \dfrac{A + A^T}{2}$ be the *symmetric part of* A and $C = \dfrac{A - A^T}{2}$ be the *skew-symmetric part*. If $B = UH$ is the polar decomposition of B, then $X = \dfrac{B + H}{2}$ is the nearest (in the Frobenius norm) matrix to A being positive semidefinite. Moreover, the Frobenius distance from X to A is given by

$$\delta_F(A) = \sqrt{\sum_{\lambda_i(B)<0} \lambda_i(B)^2 + \|C\|_F^2},$$

where $\lambda_i(B)$ are the eigenvalues of B, and $\|\cdot\|_F$ is the Frobenius norm.

Our major goal is to find out if given a tensor P that comes from a 4-leaf tree we can infer the tree topology and, at the same time, ensure that it comes from stochastic parameters using the results that we have proved in Chapter 3.

Notation Given a tensor P, we denote by $\tilde{P}$ the tensor constructed in the proof of Theorem 3.2.4.

4.1.3 Remark The conditions of $P_{+\ldots}$ and $P_{\ldots+}$ in Theorem 3.2.6 coming from stochastic parameters just guarantee π being stochastic. Since this is independent from the tree topology (because the root can be placed in another interior node, see Remark 2.5.2) we will not use these conditions.

Recall that in Theorem 3.2.6, assuming a certain tree topology, we found a condition on $\tilde{P}$ reflecting the stochasticity of the the transition matrix associated to the interior edge. This condition was given in terms of some matrix being symmetric positive semidefinite. We claim that if we apply the same construction but assuming a different tree topology we will not obtain a symmetric positive semidefinite matrix.

4.1.4 Proposition Let $\mathcal{T}$ be a 4-leaf tree and let P be a distribution in the image by $\varphi_{\mathcal{T}}$ of real stochastic parameters. The following conditions are satisfied:

(i) If $\mathcal{T}$ has tree topology equal to $T_{12|34}$, then the matrices $Flatt_{13|24}(\tilde{P})$ and $Flatt_{14|23}(\tilde{P})$ are symmetric and positive semidefinite.

(ii) If $\mathcal{T}$ has tree topology equal to $T_{13|24}$, then the matrices $Flatt_{12|34}(\tilde{P})$ and $Flatt_{14|23}(\tilde{P})$ are not symmetric positive semidefinite.

(iii) If $\mathcal{T}$ has tree topology equal to $T_{14|23}$, then the matrices $Flatt_{12|34}(\tilde{P})$ and $Flatt_{13|24}(\tilde{P})$ are not symmetric positive semidefinite.

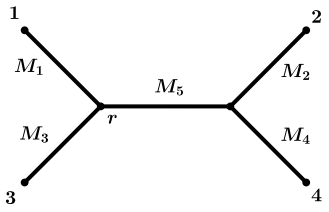

Figure 4.1: 4-leaf tree with tree topology $T_{13|24}$.

Proof

(i) Is a consequence of Corollary 3.2.9.

(ii) is shown by means of an example. (iii) will follow by symmetry. Let $\mathcal{T}$ be the tree of Figure 4.1, and take the set of nucleotides as the space of possible states. Let $\pi = (0.22, 0.26, 0.24, 0.28)$ be the root distribution and take transition matrices:

$$M_1 = \begin{pmatrix} 0.7 & 0.15 & 0.10 & 0.05 \\ 0.07 & 0.75 & 0.16 & 0.02 \\ 0.12 & 0.08 & 0.68 & 0.12 \\ 0.05 & 0.08 & 0.07 & 0.8 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0.82 & 0.05 & 0.12 & 0.01 \\ 0.11 & 0.6 & 0.07 & 0.22 \\ 0.07 & 0.14 & 0.75 & 0.04 \\ 0.12 & 0.14 & 0.10 & 0.64 \end{pmatrix},$$

$$M_3 = \begin{pmatrix} 0.67 & 0.11 & 0.09 & 0.13 \\ 0.06 & 0.75 & 0.14 & 0.05 \\ 0.07 & 0.15 & 0.63 & 0.15 \\ 0.04 & 0.08 & 0.16 & 0.72 \end{pmatrix}, \quad M_4 = \begin{pmatrix} 0.71 & 0.13 & 0.1 & 0.06 \\ 0.13 & 0.63 & 0.14 & 0.10 \\ 0.12 & 0.06 & 0.80 & 0.02 \\ 0.03 & 0.09 & 0.11 & 0.77 \end{pmatrix},$$

$$M_5 = \begin{pmatrix} 0.59 & 0.16 & 0.12 & 0.13 \\ 0.12 & 0.66 & 0.08 & 0.14 \\ 0.07 & 0.16 & 0.73 & 0.04 \\ 0.18 & 0.10 & 0.08 & 0.64 \end{pmatrix}.$$

Using (2.2) we can compute the vector of joint distributions at the leaves, which is a 4^4-vector. As we have said, it can also be regarded as a $4 \times 4 \times 4 \times 4 \times 4$ tensor P. With this tensor P we compute

$$N_{32} = P^T_{+..+}, \quad N_{31} = P^T_{.+.+},$$
$$N_{14} = P_{.++.}, \quad N_{13} = P_{.+.+},$$

and the tensors

$$\bar{P} = P *_2 N_{32}^{-1} N_{31},$$
$$\tilde{P} = \bar{P} *_3 N_{13}^{-1} N_{14}.$$

In this case the two flattenings relative to the wrong topologies are $Flatt_{12|34}(\tilde{P})$ and $Flatt_{14|23}(\tilde{P})$. We have computed the distance of these matrices to the set of symmetric positive semidefinite matrices, and we have obtained:

$$\delta_F(Flatt_{12|34}(\tilde{P})) = 5.77899 \times 10^{-10},$$

$$\delta_F(Flatt_{14|23}(\tilde{P})) = 7.13323 \times 10^{-10}.$$

These numbers are close to zero and it might seem that we simply have obtained them because of numerical errors. But if $T = T_{12|34}$ the same computations gives us

$$\delta_F(Flatt_{13|24}(\tilde{P})) = \delta_F(Flatt_{14|23}(\tilde{P})) = 5.53549 \times 10^{-17}.$$

which has a really different magnitude order. Therefore we conclude that $Flatt_{12|34}(\tilde{P})$ and $Flatt_{14|23}(\tilde{P})$ are not symmetric positive semidefinite.

4.1.5 Remark The small values obtained in the preceding proof suggest to take the logarithm to design a reconstruction method.

Computation of $\tilde{P}$

Given a tensor P, in (3.13) we have defined

$$\tilde{P} = \left(P *_2 N_{32}^{-1} N_{31}\right) *_3 N_{13}^{-1} N_{14},$$

where

$$N_{32} = P_{+..+}^T, \quad N_{31} = P_{.+.+}^T,$$
$$N_{14} = P_{.++.}, \quad N_{13} = P_{.+.+}.$$

If P is a distribution arising from some stochastic parameters on $T_{12|34}$, then the same tensor $\tilde{P}$ could also be constructed as a product of different matrices:

$$\begin{aligned}\tilde{P} &= P *_2 (N_{32}^{-1} N_{31})) *_3 (N_{13}^{-1} N_{14}) = \\ &= (P *_2 (N_{42}^{-1} N_{41})) *_3 (N_{13}^{-1} N_{14}) = \\ &= (P *_2 (N_{32}^{-1} N_{31})) *_3 (N_{23}^{-1} N_{24}) = \\ &= (P *_2 (N_{42}^{-1} N_{41})) *_3 (N_{23}^{-1} N_{24}),\end{aligned}$$

where

$$N_{ij} = M_i \text{diag}(\pi) M_5 M_j \text{ for } i < j \text{ and } N_{ji} = N_{ij}^T.$$

In this case $\tilde{P}$ corresponds to the tree with M_1 and M_4 instead of M_2 and M_3 (see figure 3.4). But different replacements could also have been considered and would also work in the proof of Theorem 3.2.6. For instance,

$$\tilde{P} = P *_2 (N_{32}^{-1} N_{31})) *_4 (N_{14}^{-1} N_{13}),$$

corresponds to the tree where M_2 has been replaced by M_1, and M_4 by M_3. And this tensor can also be obtained in 4 different ways as long as P is a distribution from the tree with topology $T_{12|34}$.

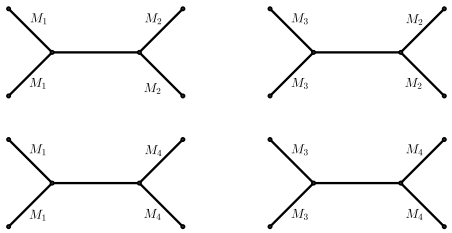

Figure 4.2: 4-leaf trees symmetric with respect to the interior edge.

In summary, the tensor $\tilde{P}$ could be taken as the tensor arising from any of the trees of Figure 4.2, and we have 4 different ways of computing each of these tensors. In the following table all these tensors $\tilde{P}$ are computed and grouped if they arise from the same tree.

Computations of $\tilde{P}$

$(P *_2 N_{32}^{-1}N_{31}) *_3 N_{13}^{-1}N_{14}$	$(P *_2 N_{32}^{-1}N_{31}) *_3 N_{23}^{-1}N_{24}$
$(P *_2 N_{42}^{-1}N_{41}) *_3 N_{13}^{-1}N_{14}$	$(P *_2 N_{42}^{-1}N_{41}) *_3 N_{23}^{-1}N_{24}$
$(P *_2 N_{32}^{-1}N_{31}) *_4 N_{14}^{-1}N_{13}$	$(P *_2 N_{32}^{-1}N_{31}) *_4 N_{24}^{-1}N_{23}$
$(P *_2 N_{42}^{-1}N_{41}) *_4 N_{14}^{-1}N_{13}$	$(P *_2 N_{42}^{-1}N_{41}) *_4 N_{24}^{-1}N_{23}$
$(P *_1 N_{31}^{-1}N_{32}) *_3 N_{13}^{-1}N_{14}$	$(P *_1 N_{31}^{-1}N_{32}) *_3 N_{23}^{-1}N_{24}$
$(P *_1 N_{41}^{-1}N_{42}) *_3 N_{13}^{-1}N_{14}$	$(P *_1 N_{41}^{-1}N_{42}) *_3 N_{23}^{-1}N_{24}$
$(P *_1 N_{31}^{-1}N_{32}) *_4 N_{14}^{-1}N_{13}$	$(P *_1 N_{31}^{-1}N_{32}) *_4 N_{24}^{-1}N_{23}$
$(P *_1 N_{41}^{-1}N_{42}) *_4 N_{14}^{-1}N_{13}$	$(P *_1 N_{41}^{-1}N_{42}) *_4 N_{24}^{-1}N_{23}$

(4.1)

$$\begin{array}{lll} N_{12} = P_{..++}, & N_{13} = P_{.+.+}, & N_{14} = P_{.++.}, \\ N_{23} = P_{+..+}, & N_{24} = P_{+.+.}. & N_{34} = P_{++..}, \end{array} \text{ and } N_{ij} = N_{ji}^T \text{ if } i > j. \tag{4.2}$$

Moreover checking whether M_5 has nonnegative entries in the proof of Theorem 3.2.4 could also be done verifying whether the matrix

$$Flatt_{13|24}(\tilde{P}) = Flatt_{14|23}(\tilde{P}), \tag{4.3}$$

is positive semidefinite where $\tilde{P}$ is some of the 16 tensors corresponding to Figure 4.2.

When all this is applied to a tensor P obtained from real data, this tensor P is not the image of stochastic parameters on $T_{12|34}$ anymore, all the 16 tensors of the Table above are different, and the equality of 4.3 does not hold. In this case, there are up to 32 different ways of checking that M_5 have nonnegative entries.

Implementation
We deal with multiple sequence alignments of DNA sequences of 4 species and we try to reconstruct the tree topology of their phylogenetic tree. From each alignment we compute a tensor P with the relative frequencies of any possible quadruples of nucleotides as components.

For any tree topology we compute:

(i) The 16 tensors $\tilde{P}$ (For $T = T_{12|34}$ see the Table (4.1)) and the two flattenings matrices ($Flatt_{13|24}(\tilde{P})$ and $Flatt_{14|23}(\tilde{P})$ if $T = T_{12|34}$) of incorrect bipartitions.

(ii) The distance δ_F of the previous 32 matrices to the space of symmetric positive semidefinite matrices, and the mean of all these distances.

Then the output of this method will be three scores, one for each topology $T_{12|34}$, $T_{13|24}$ and $T_{14|23}$, that corresponds to the means of (ii). We will choose the topology with the smaller score. This method will be called the M_5-method.

As an alternative method we will compute

(iii) For a given topology and for any tensor $\tilde{P}$, the distance between the two flattenings relative to the other two topologies, and the mean of these 16 distances. For example, for $T = T_{12|34}$, the score is given by the mean of the values $\|Flatt_{13|24}(\tilde{P}) - Flatt_{14|23}(\tilde{P})\|$ where $\tilde{P}$ is one of the 16 possible tensors of Table (4.1).

This gives us also three scores, one for each topology. The chosen topology will be the one with minimal score. This method will be called the *Flatt*-method.

4.1.6 Remark To avoid numerical problems, in the computation of matrices in (4.2) we will use $\text{adj}(N_{ij})$ instead of N_{ij}^{-1}.

4.2 Analysis of the Results

We test these methods on simulated alignments that correspond to phylogenetic trees of some tree space defined as follows. Take the tree presented in Figure 4.3, with tree topology $T_{12|34}$ and where the branch lengths are characterized in such a way that the exterior edges going to the leaves 1 and 3 have length b and the other three edges have length a. The values of a and b are taken from 0.01 to 1.5.

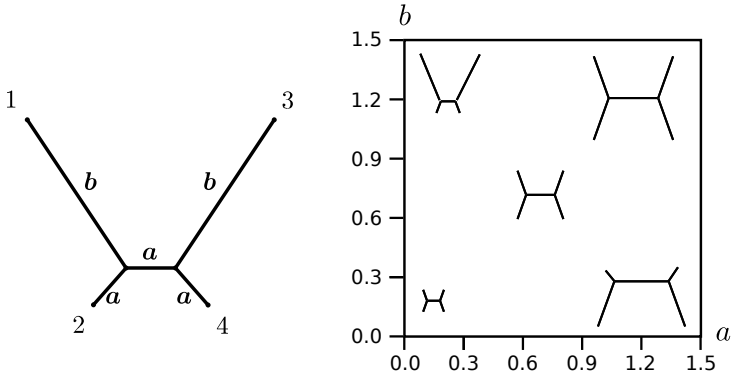

Figure 4.3: At the left, a phylogenetic tree with tree topology $T_{12|34}$ and branch lengths a and b. At the right the tree space with $a, b \in [0, 1.5]$.

We compare the results of the methods with the reconstruction method *Erik+2* [FSC15] developed by *M. Casanellas* and *J. Fernández-Sánchez* which is based in Theorem 2.6.4. It computes the distance of some normalized version of the three flattenings to the set of matrices of rank 4.

For $a = 0.01, 0.05, 0.45, 1.05, 1.45$ and $b \in \{0.01, 0.03, \ldots, 1.49\}$ we have tested these methods on 100 alignments of length 1000 generated on the topology $T_{12|34}$ with transition matrices on the general Markov model. First of all we want to see how often the means of (ii) and (iii) of Implementation for the right topology $T_{12|34}$ are

smaller than for the other topologies. The following graphics show the performance of the methods *Erik+2*, M_5 and *Flatt*.

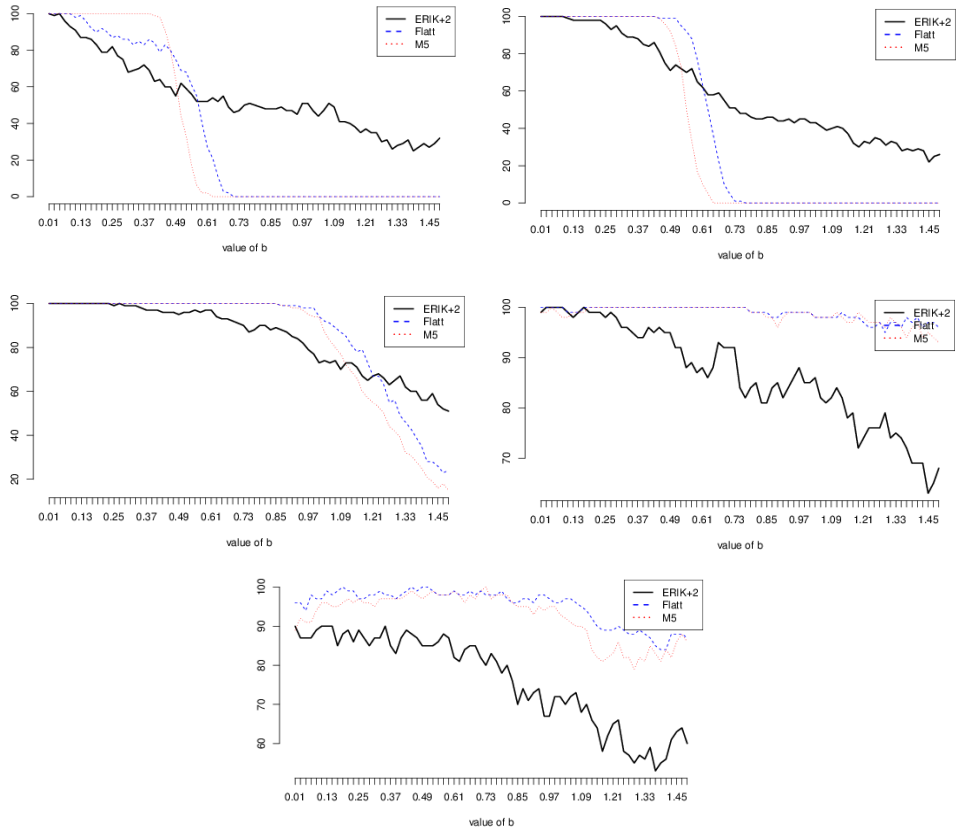

Figure 4.4: On the top: From left to right, graphics with $a = 0.01$, 0.05. In the middle: from left to right, $a = 0.45$, 1.05. At the bottom: $a = 1.45$.

From these graphics we can observe that both the mean M_5 and the *Flatt* method fails when b is much larger than a.

The area where these methods fail is called the *Felsestein zone* (Felsenstein 1978) which corresponds to small values of a and big values of b; in this zone it is said that occurs the phenomenon known as *long branch attraction*. If data has been obtained from a tree with 2 very long exterior edges (compared to the other edges), then reconstruction methods tend to join these edges. Felsenstein identified this phenomenon as a deficiency of the Maximum Parsimony method of phylogenetic reconstruction.

In order to study more closely what happens in the Felsenstein zone and what is the behavior of these means we can observe the following graphics.

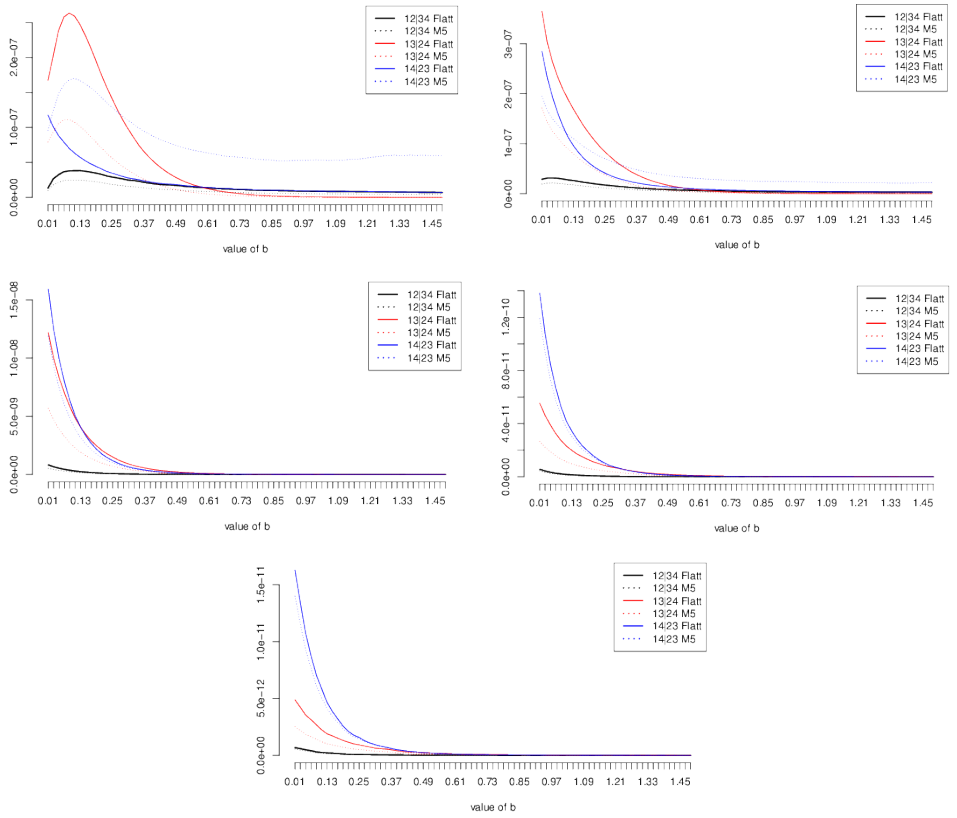

Figure 4.5: On the top: From left to right, graphics with $a = 0.01$, 0.05. In the middle: from left to right $a = 0.45$, 1.05. At the bottom: $a = 1.45$.

The solid lines are the ones corresponding to the means of *Flatt* while broken lines corresponds to M_5. Different colors correspond to different topologies. We see from the graphics that the values of these means are really small and really close for the tree topologies. This explains why the method does not recover the right topology in many cases. Moreover both methods have a similar comportment. Nevertheless when b is small the distance between the three means is bigger than for bigger b's, and so the method in these areas still works correctly (see Figure 5.3). When we increase b's the lines becomes closer and it is more difficult to know which is the right topology. However, for $a = 0.01$ and $a = 0.05$ the red line corresponding to the topology $T_{13|24}$ is lower that the other two (for $b \geq 0.60$). This behavior corresponds to the long branch attraction of the Felsenstein zone.

The next test compares the comportment of *Erik+2* and the mean M_5. Since the outputs of these two methods have different magnitude we can not produce a linear graphic for a given a. The following figures are computed for a fixed pair a, b. The three colors correspond to the three topologies, the x-axis to the scores of *Erik+2* and the y-axis to the scores of M_5. Any dot of the plane corresponds to one alignment.

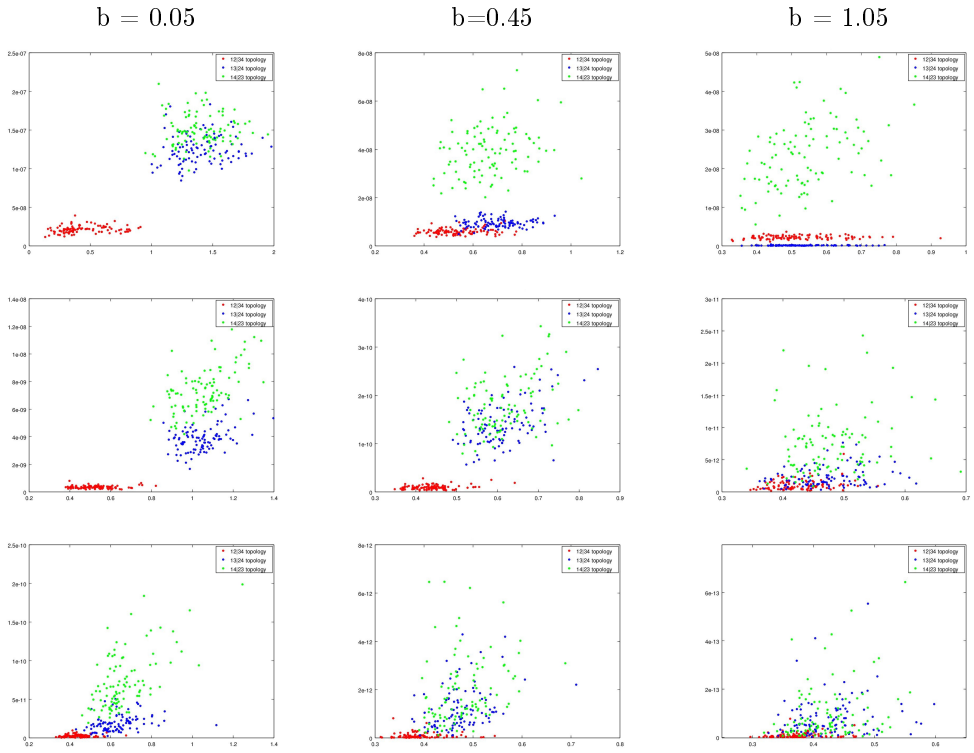

Table 4.1: On the top: $a = 0.05$. In the middle: $a = 0.45$. At the bottom: $a = 1.05$.

For $a = 0.05$ and $b = 0.05$ the red dots, which correspond to the $T_{12|34}$ topology, are far from the rest. That means that both *Erik+2* and M_5 give a smaller score for this topology than for the remainder. When b increases the dots become closer and they start to coalesce. The scores provided by *Erik+2* for the different topologies lie in the same range and do not discriminate among the topologies. For $b = 1.05$ the scores provided by M_5 keep the dots separated vertically, but it is clear that the ones corresponded to $T_{13|24}$ reach smaller values. Therefore in these cases the M_5-method fails.

For all the other values of a, especially when $b = 0.05, 0.45$ the red dots are quite separated from the others, in such a way that both *Erik+2* and M_5 recover the right topology. For $b = 1.45$ although points are quite separated, red dots have a tendency to move to the lower left corner of the plane. All these results confirm that the method is deficient in the areas where b is much bigger than a.

We finish with the analysis of this method with some different graphics. We have observed that for different values of a and b the scores obtained from M_5 take very different values. For that reason, for the three scores $S_{12|34}$, $S_{13|24}$ and $S_{14|23}$ given by M_5 for one alignment we normalize them and compute $\left(\frac{S_{12|34}}{S}, \frac{S_{13|24}}{S}, \frac{S_{14|23}}{S}\right)$ where $S = S_{12|34} + S_{13|24} + S_{14|23}$. These new scores are the barycentric coordinates of points lying on a triangle. The following graphics present 100 of these points corresponding to 100 alignments for fixed a and b.

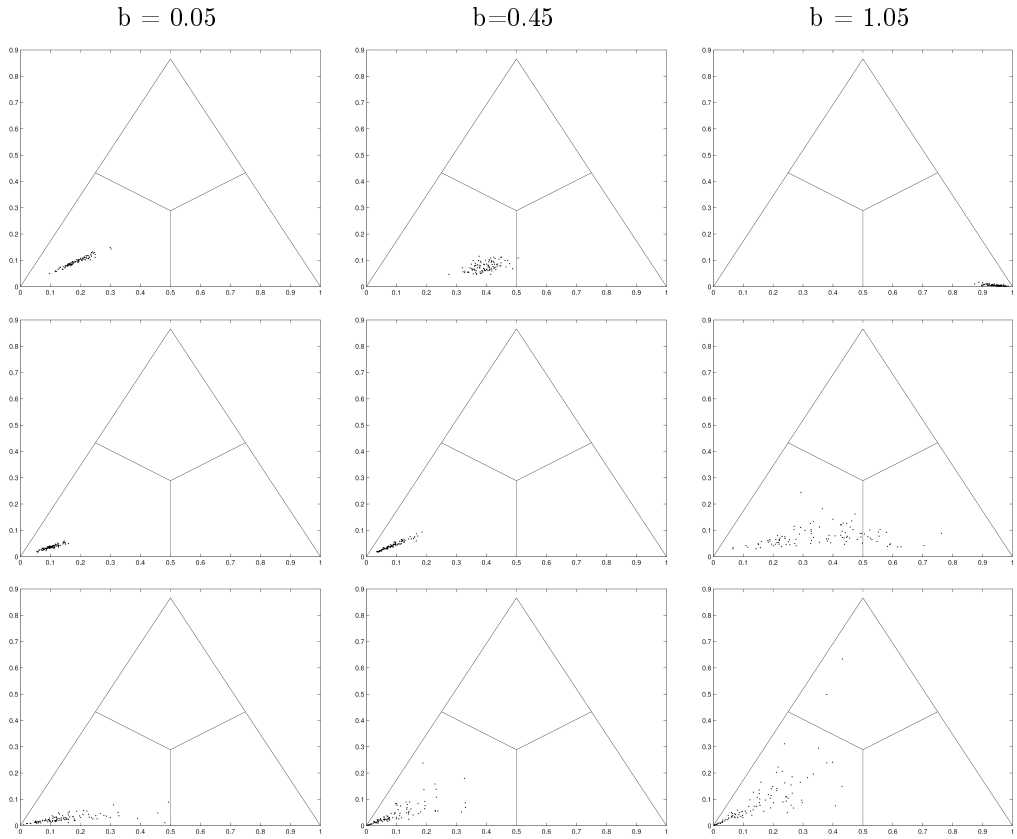

Table 4.2: On the top: $a = 0.05$. In the middle: $a = 0.45$. At the bottom: $a = 1.05$. Compare these graphics with Table 4.1.

For a reliable reconstruction method, the points represented in the triangle should be pictured over one of the angle bisectors of the triangle, near the vertex, and away from the perpendicular bisectors. In this way the method would distinguish well one topology from the other two. In our case, this fact occurs when a is smaller than b. On the other cases the dots are more dispersed on the triangle or near some perpendicular bisection, for that reason the method fail in some cases. Notice the right upper graphic ($a = 0.05$, $b = 1.45$), where all the points are over the right lower vertex of the triangle. As we have seen in the last two types of graphics, the method fails in this region of the parameters.

In conclusion, despite the good results obtained by the M_5-method for values $a \geq b$ the direct application of the conditions of theorem 3.2.6 does not seem to provide a good reconstruction method, since it does not work in the Felsenstein zone. Recall that, Theorem 3.2.6 is stated assuming a tree topology, and gives some conditions for having stochastic parameters. For that reason our first idea was to use the scores of M_5 as a complementary information for a reconstruction method as *Erik+2*. However the results of M_5 do not give new information that allow us to improve the already existing inference methods.

5
A new method for phylogenetic reconstruction

In this chapter we propose a new reconstruction method based on the ideas discussed in Chapter 4 but with a new implementation. Basically we will use again the distance of matrices to the variety of symmetric positive definite matrices and some additional information based in Corollary 3.2.9. We keep the notation of Chapter 3 and Chapter 4.

By Corollary 3.2.9, we have $Flatt_{13|24}(\tilde{P}) = Flatt_{14|23}(\tilde{P})$ if $\tilde{P}$ arises from a tree with tree topology $T_{12|34}$ and therefore

$$\delta_F\big(Flatt_{13|24}(\tilde{P})\big) = \delta_F\big(Flatt_{14|23}(\tilde{P})\big).$$

Moreover in Proposition 4.1.4 we have seen that this equality is not satisfied if we compute the same but with the assumption that $\tilde{P}$ arises from some other topology.

On the other hand since distances of matrices of (4.1) are usually small numbers and in many cases we can not compare them, we have decided to work with $-ln\left(\delta_F\big(Flatt_{13|24}(\tilde{P})\big)\right)$ instead of $\delta_F\big(Flatt_{13|24}(\tilde{P})\big)$, where $ln(x)$ is the natural logarithm. Now, with these new scores we will look for the highest one, which implies the minimum distance to the set of symmetric positive semidefinite matrices. Therefore, for any tree topology T (w.l.o.g. suppose $T = T_{12|34}$) we compute the following scores:

(i) $-ln\left(\delta_F\big(Flatt_{13|24}(\tilde{P})\big)\right)$, $-ln\left(\delta_F\big(Flatt_{14|23}(\tilde{P})\big)\right)$ with $\tilde{P}$ as in (4.1). We compute the mean m_1 of these 32 values. This gives information on how likely is $\tilde{P}$ to come from stochastic parameters.

(ii) For any $\tilde{P}$ (4.1) we compute $\mid -ln\left(\delta_F\big(Flatt_{13|24}(\tilde{P})\big)\right)+ln\left(\delta_F\big(Flatt_{14|23}(\tilde{P})\big)\right) \mid$ which gives us 16 values. Here we also compute the mean m_2 of these 16 scores. This value gives information on the tree topology.

The output of this method are three scores m_1 and three m_2, that is, a pair (m_1, m_2) for each topology. As we have said we want the topology of maximum m_1 and since

theoretically $|-ln\left(\delta_F\left(Flatt_{13|24}(\tilde{P})\right)\right)+ln\left(\delta_F\left(Flatt_{14|23}(\tilde{P})\right)\right)|=0$, we also ask for the minimum m_2.

In general the highest m_1 and the minimum m_2 do not coincide in the same topology. For that reason we will follow the next procedure.

Phylogenetic reconstruction method proposed:

1) Discard the tree topology with minimum m_1.
2) Discard the tree topology with maximum m_2.
3) If the two discarded topologies are different the output topology will be the only one that has not been rejected. Otherwise, if we discard just one topology we will output the one with minimum m_2, since this mean is a score for the topology of the tree whereas m_1 measures the stochasticity of the parameters.

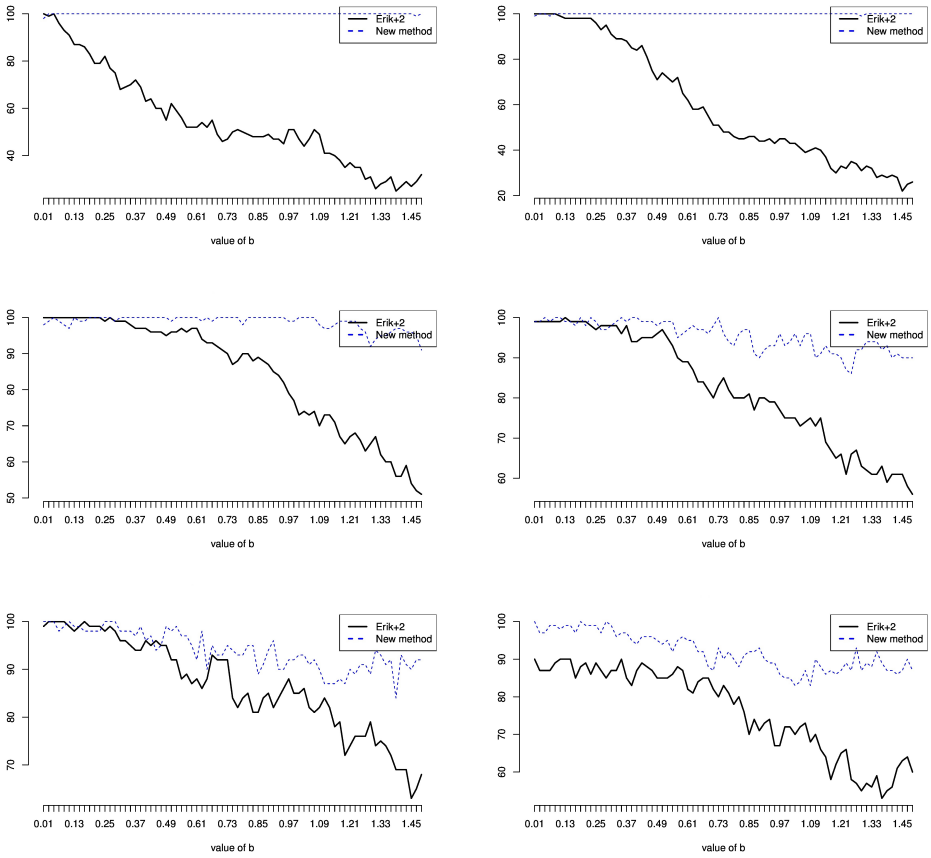

Figure 5.1: On the top: From left to right. graphics with $a = 0.01, 0.05$. In the middle: from left to right. $a = 0.45, 0.85$. At the bottom: from left to right. $a = 1.05, 1.45$.

This new method gives us really good results, as presented below. We have tested this method with trees of the treespace as in Chapter 4. For any pair of branch lengths

a, b we have used the same 100 simulated alignments of length 1000 as in Chapter 4. The following graphics show, for fixed a and every b, in how many of these 100 alignments *Erik+2* or this new method recover the right tree topology.

From these graphics we observe that this new method recovers the correct topology in more than 80 percent of the simulated alignments. Moreover we have really good results in Felsestein zone, which is the most difficult part of the treespace.

In order to justify the method we also present graphics that show the values m_1 for the three tree topologies $T_{12|34}$, $T_{13|24}$ and $T_{14|23}$ for fixed a and every b.

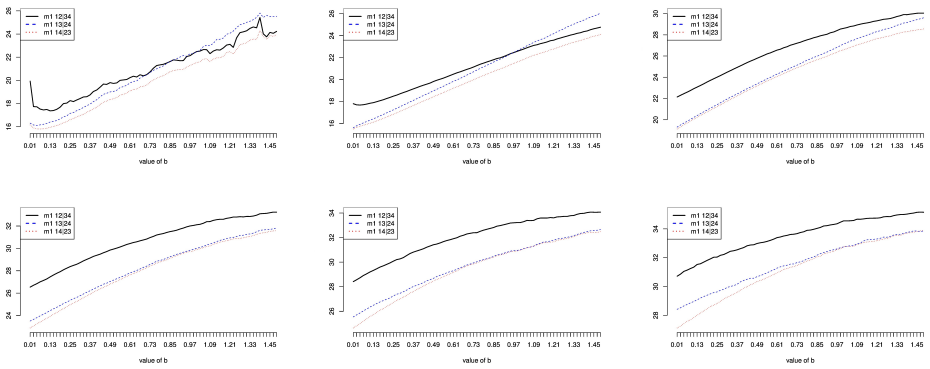

Figure 5.2: Average value m_1. On the top: from left to right, graphics with $a = 0.01$, 0.05 and $a = 0.45$. At the bottom: from left to right, 0.85 ,$a = 1.05$ and 1.45..

The only region where the value m_1 for the tree topology $T_{12|34}$ is not the greatest is in the Felsenstein zone, i.e. when b is much larger than a. It seems that in the other areas of the treespace we could recover the tree topology with only this score m_1.

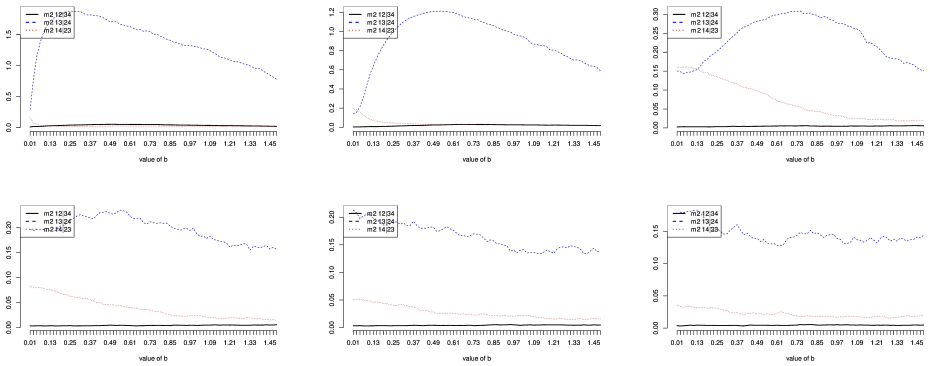

Figure 5.3: Average value m_2. On the top: from left to right, graphics with $a = 0.01$, 0.05 and $a = 0.45$. At the bottom: from left to right, 0.85 ,$a = 1.05$ and 1.45.

In this case, for $a \geq 0.45$ the least score of m_2 is taken by the topology $T_{12|34}$. Almost everywhere the value of $T_{13|24}$ is the highest, so we discard this topology and we avoid the problems that would have only with m_1 in the Felsenstein zone.

Finally we compare the method with *Erik+2* in the following table, computing the percentage of times that these two methods coincide in a correct or wrong topology and the times that they recover different topologies.

	a=0.01	a=0.05	a=0.45	a=0.85	a=1.05	a=1.45
Erik+2 and the new method provide the correct topology	72.13	76.17	92.47	85.35	87.4	80.68
Erik+2 and the new method provide the same topology, but an incorrect one	0	0	0.15	1.41	0.8	2.23
Erik+2 and the new method provide different incorrect topologies	0	0	0.05	0.01	0.01	0.08
Only *Erik+2* recovers the correct topology	0.04	0.04	1.08	3.07	5.04	5.63
Only the new method recovers the correct topology	27.83	23.79	6.25	10.16	6.75	11.38

Table 5.1: Values of the table are percentages.

If we consider both methods the percentage of success when both topologies coincide is really high. Nevertheless, when both methods do not coincide, this new method recovers the right topology in more cases.

In the following graphics we compare the success of *Erik+2* and *New method* with the success of the traditional reconstruction methods *Maximum likelihood* and *Neighbor joining*.

The *Maximum Likelihood* method is a model dependent method, which means that it needs an evolutionary model and the output of the algorithm can be different for different models. This method provides a score of a particular tree, that is the likelihood of the observed alignment having evolved according to the chosen tree model. The optimal tree is the tree with the highest likelihood score. In other words, *ML* finds the tree and the parameters of the model that produce the observed data with the highest probability. However finding the maximum likelihood tree is a hard problem requiring numerical methods, and limiting the size of the trees which can be constructed. The maximum likelihood method was invented by Fisher [Fis22] and later applied to phylogenetic inference by Felsenstein [Fel81].

The *Neighbor joining* method is based on the criterion of minimum evolution, in which the best tree is one that minimizes the length of the inner branches. To do this, from a star tree, the pair of nearest sequences is determined and the corresponding leaves are joined at an internal node. This process is repeated with the rest of the leaves until they are all linked by internal nodes that minimize the length of each of the interior edges. It provides a tree topology and branch lengths. The *NJ* algorithm is really fast and if the distance matrix (which entries are the distances among the sequences of the alignment) is an exact description of the true tree, then neighbor

joining is guaranteed to reconstruct the correct tree. However if this distance matrix is close to be a tree metric then this method stills reconstruct the correct tree. This method was originally developed by Naruya Saitou and Masatoshi Nei in 1987 [SN87].

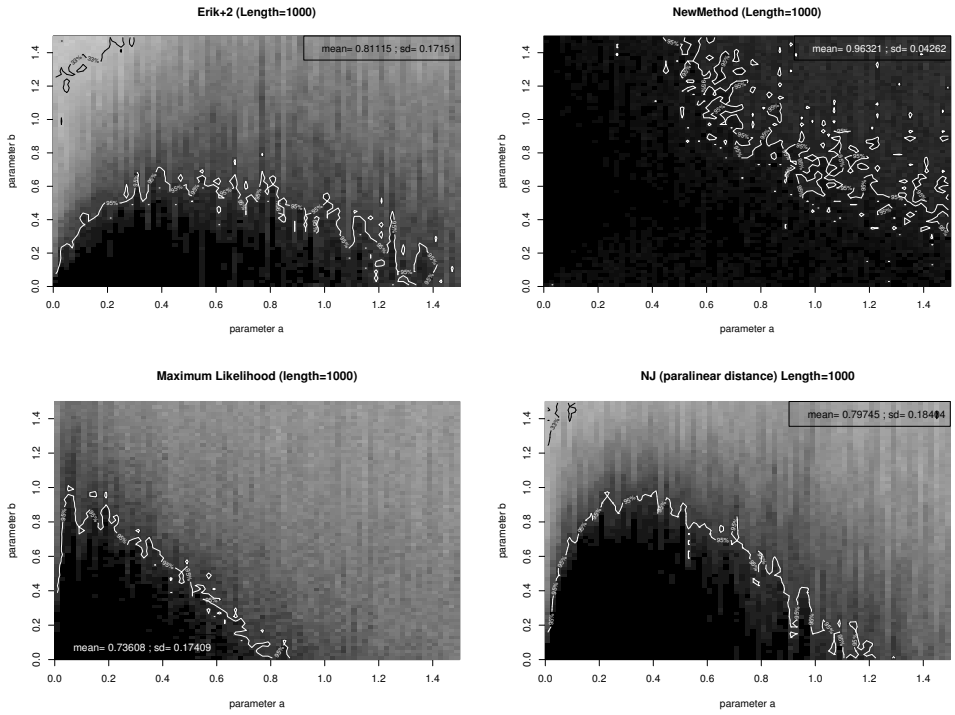

Figure 5.4: Percentage of exits of different methods on simulated alignments on on trees of Figure 4.3. On the top: from left to right *Erik+2* and *New method*. At the bottom: from left to right *Maximum likelihood* (ML) and *Neighbor joining* (NJ) on trees of Figure 4.3.

In Figure 5.4 black is used to represent 100% of exits, white to represents 0% and different tones of gray the intermediate percentages. We can see that the graphic of the *New Method* has a quite differently from the rest. Moreover the dark zone covers much part of the graph, and then this method recovers the correct topology in most cases in simulated alignments of this tree space.

Time of execution:
Given a pair $a, b \in [0, 1.5]$ the time required to compute scores of *Erik+2*, m_1 and m_2 for 100 alignments is around $12s$. For instance:

- $a = 0.01$ and $b = 0.01$: $11.661s$,
- $a = 0.01$ and $b = 1.01$: $12.821s$,
- $a = 0.51$ and $b = 0.51$: $12.119s$,
- $a = 1.45$ and $b = 1.45$: $13.129s$.

For a a fixed and $b \in [0, 1.5]$ the program takes 15 to 20 minutes to compute 100 alignments for each pair a, b:

- $a = 0.01$: $15m59.760s$,
- $a = 0.75$: $16m10.177s$,
- $a = 1.45$: $19m46.410s$.

All these computations have been done using $c++$ in a server with processor *Intel(R) Xeon(R) CPU E5-2430 a 2.20GHz*.

6 Conclusions

In this work, we have achieved all the proposed objectives. We have seen that conditions of stochasticity of the parameters by itself did not give us much information and the implementation of these conditions gave us very bad results in the Felsentein zone. But this has allowed us to analyze the results and find a method of phylogenetic reconstruction with very good results in all the treespace. Finally we can extract the following conclusions:

- We have understood the theoretical results of stochastic conditions of the parameters and we have provided a counterexample to an error in a proof of [ART12] as well.
- We have described how to compute the transformed tensor $\tilde{P}$ out of the joint distribution P at the leaves and in how many ways it can be constructed.
- We have first used all these computations of $\tilde{P}$ to study the stochasticity of parameters and we have implemented them in $c++$. The results of this method were not quite good according to simulations, which means that just the conditions of the parameters being stochastic do not help us in the problem of phylogenetic reconstruction.
- By using the same ideas but with some modifications on the implementation we have proposed a new method of phylogenetic reconstruction. We have also implemented it in $c++$ and we have tested it in simulated data in the treespace. The results has been compared with the method *Erik+2*. We conclude that this new method has really good results in alignments of trees in the treespace.

This last item lead us to think that there are further research to do:

- Test this new method in trees of random branch length.
- Test this new method in real data.
- Check whether the new phylogenetic invariants we found are sufficient to describe the phylogenetic algebraic variety.

References

[AR03] ES Allman and JA Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144, 2003.

[AR04] ES Allman and JA Rhodes. *Mathematical models in biology, an introduction.* Cambridge University Press, January 2004. ISBN 0-521-52586-1).

[AR05] ES Allman and JA Rhodes. The mathematics of phylogenetics. University of Alaska Fairbanks, 2005.

[AR07] Elizabeth S. Allman and John A. Rhodes. Phylogenetic invariants. In *Reconstructing evolution*, pages 108–146. Oxford Univ. Press, Oxford, 2007.

[AR08] ES Allman and JA Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Advances in Applied Mathematics*, 40:127–148, 2008.

[ART12] E. S. Allman, J. A. Rhodes, and A. Taylor. A semialgebraic description of the general markov model on phylogenetic trees. *ArXiv e-prints*, dec 2012.

[Cas12] M Casanellas. Algebraic tools for evolutionary biology. *La Gaceta de la RSME*, 15:521–536, 20012.

[CFS10] M. Casanellas and J. Fernandez-Sanchez. Reconstrucción filogenética usando geometría algebraica. *Arbor. Ciencia, pensamiento, cultura*, 96:207–229, 2010.

[CFS11] M Casanellas and J Fernandez-Sanchez. Relevant phylogenetic invariants of evolutionary models. *Journal de Mathématiques Pures et Appliquées*, 96:207–229, 2011.

[CGS05] M Casanellas, LD Garcia, and S Sullivant. Catalog of small trees. In L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 15. Cambridge University Press, 2005.

[Eri05] N Eriksson. Tree construction using singular value decomposition. In L Pachter and B Sturmfels, editors, *Algebraic Statistics for computational biology*, chapter 19, pages 347–358. Cambridge University Press, 2005.

[Fel81] J Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[Fis22] R A Fisher. On the mathematical foundation of theoretical statistics. *Philosophical Trnsactions of the Royal Society of London.*, 222:309–368, 1922.

[FSC15] J Férnandez-Sánched and M Casanellas. Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages. 2015. Submitted.

[Hig88] Nicholas J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, May 1988.

[Kim81] M Kimura. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Nat. Acad. Sci. , USA*, 78:454–458, 1981.

[SN87] N Saitou and M Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.