

Combining Prior Information for the Prediction of Transcription Factor Binding Sites

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)

im Fachgebiet

Informatik

Vorgelegt

von MSc ETH Philipp F. Benner
geboren am 09.04.1984 in Göttingen.

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Jürgen Jost (MPI-MiS, Leipzig)
2. Prof. Dr. Ivo Große (Martin-Luther-Universität Halle-Wittenberg)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der
Verteidigung am 23. Februar 2016 mit dem Gesamtprädikat summa
cum laude.

Contents

Chapter 1. Motivation	1
Chapter 2. The space of phylogenetic trees	5
2.1. Phylogenetic trees	5
2.2. Basic tree space structure	7
2.3. Nonpositively curved spaces	10
2.4. Combinatorics and geometry of the tree space	17
2.5. Computing means and medians in tree space	19
2.6. Computing geodesics in tree space	30
2.7. Phylogenetic networks	35
2.8. Converting tree representations	38
Chapter 3. Substitution processes	41
3.1. Substitution models on phylogenetic trees	41
3.2. The likelihood of an alignment	46
3.3. Computing the likelihood	49
3.4. Likelihood decomposition	51
3.5. Detecting conserved sites	56
Chapter 4. Point estimates in phylogenetic reconstructions	59
4.1. Consensus vs. posterior mean	61
4.2. Star-tree problem	62
4.3. Statistical model	63
4.4. Approximation of the posterior distribution	69
4.5. MCMC with Hamiltonian dynamics	70
4.6. Estimation results	72
Chapter 5. Prediction of transcription factor binding sites	77
5.1. Statistical model	78
5.2. Process priors	80
5.3. The label switching problem	84
5.4. Predictions	85
5.5. Phylogenetic likelihood approximation	89
5.6. Discovery of transcription factor binding sites in <i>Drosophila</i>	91

Chapter 6. Conclusion	99
Software	102
Appendix A. Introduction to probability and statistics	103
A.1. Essentials of probability theory	103
A.2. Conditional probability distributions	109
A.3. Summarizing a distribution	112
A.4. Markov chain Monte Carlo methods	114
A.5. Variational Bayes	117
A.6. Probability in statistics	121
A.7. Statistical inference	123
A.8. Statistical decision theory	124
Appendix. Bibliography	131

CHAPTER 1

Motivation

Despite the fact that each cell in an organism has the same genetic information, it is possible that cells fundamentally differ in their function. The molecular basis for the functional diversity of cells is governed by biochemical processes that regulate the expression of genes. Key to this regulatory process are proteins called *transcription factors* that recognize and bind specific DNA sequences of a few nucleotides. These *transcription factor binding sites* (TFBS), or *cis*-regulatory elements, are located in the vicinity of genes. Once a binding site is occupied by its factor, the transcription of the nearby gene is either up or down regulated through a variety of different molecular mechanisms [cf. e.g. Latchman, 1997, Ptashne and Gann, 1997].

Understanding the regulatory network might therefore reveal much of how cells function. Hence, the identification of transcription factors and their binding sites has become a major research area in recent years. However, the prediction of binding preferences from the structure of transcription factors is still an unsolved problem. For that reason, binding sites are commonly identified by searching for overrepresented sites in a given collection of nucleotide sequences. Such sequences might be known regulatory regions of genes that are assumed to be coregulated, or they are obtained through so-called *ChIP-seq* experiments that identify approximately the sites that were bound by a given transcription factor. In both cases, the provided nucleotide sequences are much longer than the actual binding sites and computational tools are required to uncover the actual binding preferences of a factor. Aggravated by the fact that transcription factors recognize not only a single nucleotide sequence, the search for overrepresented patterns in a given collection of sequences has proven to be a challenging problem.

The first computational methods merely relied on the given set of sequences, but quickly it was recognized that additional information is required in order to make reliable predictions. This information is obtained from related species where the same transcription factor is known to be present. The set of nucleotide sequences is augmented by their orthologs, i.e. sequences from other species that have evolved from a common ancestor. By constructing multiple sequence alignments of the orthologous sequences it is possible to identify functional regions that are under selective pressure

and therefore appear more conserved than others. The processing of the additional information exerted by ortholog sequences relies on a *phylogenetic tree* equipped with a *nucleotide substitution model* that not only carries information about the ancestry, but also about the expected similarity of functional sites.

During the last two decades, a myriad of software tools have been developed to tackle the problem of identifying binding sites [Wasserman and Sandelin, 2004, Tompa et al., 2005]. One of the most complex and probably the closest approach to the one presented here is PhyloGIBBS, a software package developed by Siddharthan et al. [2005]. It is based on a statistical model for enriched patterns in the set of multiple sequence alignments. The analysis requires a phylogenetic tree to identify functional regions. The interpretation of the edge lengths of the tree crucially depends on the substitution model. However, there exists a significant discrepancy between the commonly used models for inferring phylogenetic trees from sequence alignments and the models for the identification of binding sites. The model mismatch can be partly corrected by rescaling the tree to an appropriate size, which is equivalent to adjusting the substitution rate. Nevertheless, the question remains how a phylogenetic tree can be obtained that allows a consistent analysis of the data.

To circumvent model mismatches as well as having a clear interpretation of the model parameters, the phylogenetic tree should be estimated using the correct substitution model. Some estimation methods proceed by maximizing a posterior distribution or a likelihood function, and are amenable to an exact reconstruction of the optimal tree, but Bayesian phylogenetic analyses generally produce posterior distributions that are best explored by generating posterior samples. While a large enough posterior sample offers a faithful representation of the posterior knowledge, it is of little scientific interest unless summarized by some statistics [Robert, 2001]. A summary can balance contributions from the different tree topologies occurring in the sample, resulting in a legit phylogenetic tree, or combine them within a phylogenetic network. Here the focus is on the former, showing how to build a phylogenetic tree that faithfully represents the sample in its entirety, despite competing topologies occur [Benner et al., 2014].

Building upon the work published by Billera et al. [2001], who deciphered the geometric structure of the space of phylogenetic trees and first proposed a construction of the *tree space* (sometimes also called *BHV tree space*, where BHV is an acronym of Billera, Holmes, and Vogtmann), it is shown how the computation of the posterior mean of a sample of phylogenetic trees can be achieved by simply reaching out for the appropriate

geometry. The BHV space is obtained by gluing together the positive orthants of the linear space associated to each topology, so that a point in this space identifies both a tree topology and the lengths of the corresponding edges. The adjacency structure between any two orthants reflects the edges shared by the two corresponding topologies, and permits the definition of paths visiting several orthants. Any two trees are connected by at least one path and the one with minimal length is called a *geodesic*. Therefore, the length of a geodesic qualifies as a distance function between phylogenetic trees, and offers a theoretically and practically appealing alternative to existing distances (e.g NNI or the Robinson-Foulds distance). Furthermore, using implicit characterizations of the posterior mean and median as minimizers of appropriate loss functions [Benner et al., 2014, Holmes, 2005], algorithms developed by Bačák [2014a,b], Miller et al. [2012], Sturm [2003] compute an approximation of these statistics by walking along geodesics. Here, the determination of the geodesics is done in polynomial time thanks to an algorithm due to Owen and Provan [2011].

While it seems natural to estimate the phylogenetic tree with the required substitution model, it was so far hindered by the lack of appropriate algorithms. The substitution model for the identification of binding sites is heavily parameterized. Each position in an alignment is equipped with its own *stationary distribution* that can be thought to represent the prevalences for the observed nucleotides, which reflects the selective pressure acting on that site. This parameterization leads to a slow convergence rate of the Markov chain Monte Carlo (MCMC) algorithm that is used to explore the posterior distribution. Hence, an analytical integration over the stationary distributions is inevitable, for which an algorithm is presented.

As a result, a Bayesian method for the identification of TFBS is presented, which allows a clear interpretation of all model parameters. In particular, the binding preferences of the transcription factor can be identified as the stationary distributions at the binding sites. A clear interpretation of the model parameters also significantly eases the analysis of experimental data and leads to more reliable posterior estimates.

This thesis is structured as follows. In chapter 2 the geometry of the BHV tree space is studied in detail [published in a similar but less detailed version in Benner and Bačák, 2013, Bačák, 2014b]. Furthermore, all required methods for computing a proper summary of a set of phylogenetic trees are explained in detail. The definition of the statistical model depends on the substitution model that is used in combination with the phylogenetic tree. In section 3.1 the basic theory of substitution models as continuous-time Markov processes is discussed. In addition, important model assumptions and prior settings are explained. The algorithm for integrating analytically

over the stationary distributions is developed in the sequel. The methods for computing proper summaries of a set of phylogenetic trees are very recent. To show that the resulting estimates are indeed more plausible, we compare them in chapter 4 to existing methods [published as Benner et al., 2014]. In section 5.1 the model for identifying binding sites is outlined. The posterior distribution of this model is inaccessible, even to MCMC methods. A novel approach motivated by variational Bayesian methods is used to obtain an approximate posterior distribution that can be efficiently explored. The full framework presented here is then demonstrated on a data set by Zinzen et al. [2009]. Throughout this thesis, it is assumed that the reader is familiar with the philosophical and technical aspects of Bayesian inference as well as the foundations of probability theory. A brief summary of the topic can be found in Appendix A.

CHAPTER 2

The space of phylogenetic trees

Phylogenetic trees are central to the study of evolution, so much that the sketch of a tree of species by Sir Charles Darwin has become the icon of this theory. Phylogenetic trees relating units of selection (be it functional domains, genes, or species) are structures of primary interest for systematists, but also instrumental to a wealth of other studies where evolutionary correlations need to be accounted for (see for instance McCue et al. [2001]). Various statistical models pertaining to diverse types of observables can be found in the literature, as well as methods for estimating their parameters and reconstructing a phylogenetic tree [Gascuel, 2005]. The estimation of phylogenetic trees requires a more abstract perspective, where a tree is merely seen as parameter of the statistical model. This raises the question of what the parameter space of the statistical model is and how a summary of the posterior distribution can be computed. To answer this question we begin the discussion in section 2.1 with a formal description of phylogenetic trees.¹ Crucial is the observation that a tree consists of a set of edges that are assigned a length. Hence, a tree can be represented as a point in a continuous space and the basic structure of this space is explained in section 2.2. To summarize a distribution in this space, it is necessary to take a closer look at the geometry, which we do in section 2.4 after reviewing some essential theoretical aspects in section 2.3. In the sequel, we turn to more technical aspects of how means and medians of a set of phylogenetic trees can be computed. The discussion is mostly self-contained and is intended for readers who are not experts in analysis or metric spaces.

2.1. Phylogenetic trees

A phylogenetic n -tree, with $n \in \mathbb{N}$ and $n \geq 3$, is a connected graph with no circuits that has $n + 1$ terminal vertices, labeled from 0 to n , which we call *leaves*. Vertices that are not leaves are called *interior vertices*, or sometimes simply *branching points*. For the following discussion we do not assign labels to the interior vertices, the tree is therefore called semi-labeled.

¹A similar discussion can be found in Benner and Bačák [2013] and Bačák [2014b], which relies on the basic definition of phylogenetic trees as split systems [see e.g. Semple and Steel, 2003, Dress et al., 2012].

It is important that the phylogenetic tree is not required to have only binary branching points, i.e. the degree of an interior vertex is greater or equal to two. The leaves represent the species that are considered in a phylogenetic analysis, which are those that are usually still present today. The vertex connected to leaf 0 is sometimes called the *root* of the tree, in which case it is thought to represent the common ancestor of species $1, \dots, n$, whereas leaf 0 is called the *outgroup*. However, the root plays no distinguished role in our discussion and the tree should rather be seen as an *unrooted tree*, because in practice the causal direction of evolution is not identifiable, as discussed in chapter 3.

The combinatorial structure or *topology* of the tree is defined through its set of edges. An edge that is incident to a leaf is called a *leaf edge*, while all remaining edges are *interior edges*. For the later discussion of the geometry of the tree space, it is necessary to uniquely identify edges across topologies. Therefore, we define an edge e as a bipartition $(A | B)$ of the set of leaves $L = \{0, \dots, n\}$, i.e. $L = A \cup B$, which is also called a *split*. Intuitively speaking, if an edge is removed from a tree, we obtain two trees with leaf sets that are defined by the split of the edge. Every edge e of a tree t is also assigned a length $|e|_t \geq 0$. An example is given in Figure 1. For instance,

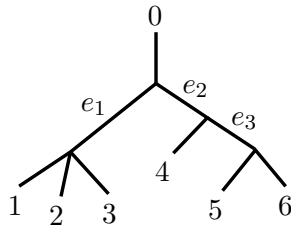


FIGURE 1. An example of a 6-tree with three interior edges.

the edge e_1 is defined as the split $(0, 4, 5, 6 | 1, 2, 3)$. For a fixed n the number of possible splits is equal to the number of bipartitions of L . The Stirling number of the second kind $S(n, k)$ gives the number of partitions of L into k non-empty subsets, which in the case of bipartitions reduces to $S(n, 2) = 2^n - 1$. In a phylogenetic tree, the leaf edges are always present, which means that the topology of the tree is defined solely by the set of interior edges. A phylogenetic tree has at most $n - 2$ interior edges, which have to be pairwise *compatible*. Two edges e_1 and e_2 with splits $A_1|B_1$ and $A_2|B_2$ are compatible if and only if one of the sets

$$A_1 \cap A_2, \quad A_1 \cap B_2, \quad A_2 \cap B_1, \quad B_1 \cap B_2$$

is empty [Buneman, 1971]. Otherwise, we call them *incompatible*. For further details, see for instance Semple and Steel [2003], Dress et al. [2012]. This leads us to the following definition of a phylogenetic tree.

DEFINITION 2.1.1 (Phylogenetic tree). A phylogenetic tree is a connected graph with no circuits, which has $n + 1$ leaves and at most $n - 2$ pairwise compatible interior edges.

2.2. Basic tree space structure

In this section, we discuss the basic structure of the phylogenetic tree space introduced by Billera, Holmes, and Vogtmann [2001], in order to gain an intuitive understanding of the space. Further details, especially about the geometry of the space, is outlined in section 2.4.

The space of all phylogenetic trees with $n + 1$ leaves is denoted \mathcal{T}_n . As already discussed, regardless of the topology the same $n + 1$ leaf edges are always present and might only have assigned different lengths, so that they can be identified as a point in Euclidean space of dimension $n + 1$. Therefore, we will focus our attention to the structure of the space limited to the interior edges, which is much more complicated. The full tree space \mathcal{T}_n is simply the product of the space of leaf edges and that of the interior edges.

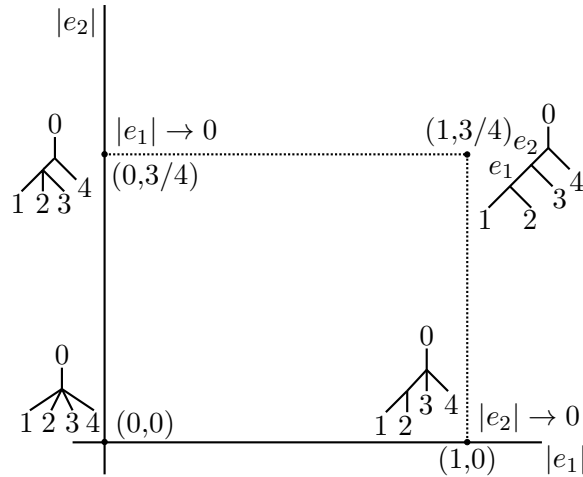


FIGURE 2. An orthant of \mathcal{T}_4 .

Consider a fixed phylogenetic tree $t \in \mathcal{T}_n$ with $r \leq n - 2$ interior edges $I = (e_1, \dots, e_r)$ of positive lengths l_1, \dots, l_r . The r -tuple $l = (l_1, \dots, l_r)$ is an element of the the open orthant $(0, \infty)^r$ of dimension r . Therefore, we may associate with each point $l' \in (0, \infty)^r$ a tree s with the same topology

as t but with interior edge lengths defined by l' . That is, t and s have the same set of interior edges as well as the same leaf edge lengths, but different interior edge lengths. The boundary $\partial(0, \infty)^r$ of $(0, \infty)^r$ is defined to consist of those trees which only have a subset of the edges of t . Equivalently, one may say that the lengths of some edges of a tree $s \in \mathcal{T}_n$ were shrunk to zero. For instance, given a fixed binary tree which has $r = n - 2$ interior edges, the faces of the corresponding orthant represent non-binary trees. An example is given in Figure 2 for trees with 5 leaves. In this example, e_1 represents the edge $(1,2 \mid 0,3,4)$ and e_2 the edge $(1,2,3 \mid 0,4)$. A point on the vertical boundary is reached by shrinking the length of edge e_1 to zero. On the other hand, the horizontal boundary ray is reached by shrinking e_2 .

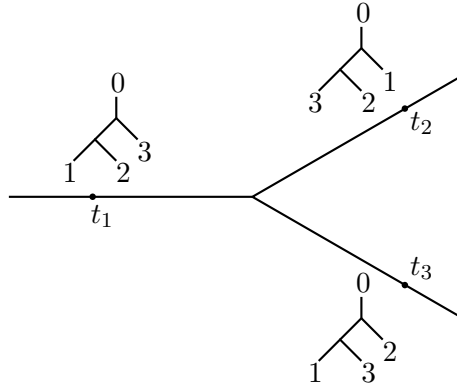
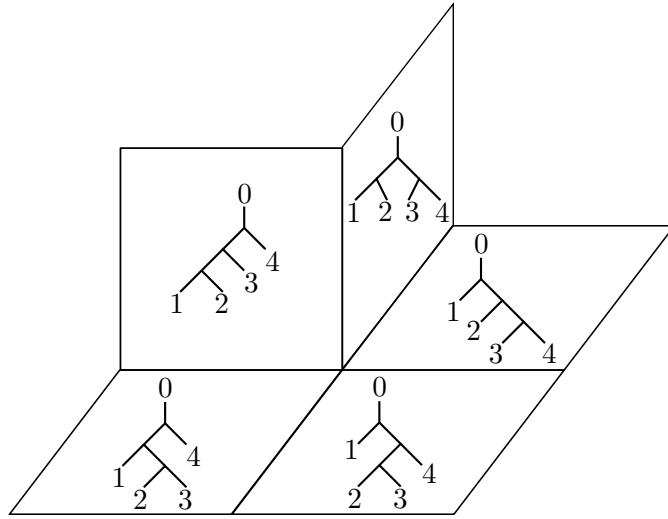


FIGURE 3. Treespace \mathcal{T}_3 , which consists of three rays glued together at the origin.

We may therefore associate with each set I of $n - 2$ pairwise compatible edges an orthant $\mathcal{O}(I)$ of the form $[0, \infty)^{n-2}$. Each point $l \in \mathcal{O}(I)$ defines a tree of topology given by the set of edges associated with this orthant and interior edge lengths l . The BHV tree space \mathcal{T}_n consists of $(2n - 3)!! = (2n - 3)(2n - 5) \cdots 5 \cdot 3$ orthants, one for each pairwise compatible set of edges, which are *glued* together at common lower-dimensional faces that correspond to non-binary trees. An exact definition of how orthants are glued together will be given in later sections. The simplest space is \mathcal{T}_3 , which comprises of only three distinct topologies identified by the single internal edge. As depicted in Figure 3, it consists of three rays $[0, \infty)$, one for each of the edges $(0 \mid 1, 2)$, $(0, 1 \mid 2)$, and $(0, 2 \mid 1)$. The rays are glued together at a single common point, the origin $\mathbf{0}$. In contrast, the space \mathcal{T}_4 is already much more complicated. Figure 4 shows five out of 15 orthant of \mathcal{T}_4 . There is no isometric embedding of \mathcal{T}_4 into three-dimensional Euclidean space, which is why only part of it can be shown. The figure suggests that there exists not only a well-defined intrinsic distance between any two

FIGURE 4. Five out of 15 orthants of \mathcal{T}_4 .

trees, but also a geodesic, i.e. a (unique) shortest path. If the two trees are within the same orthant, the geodesic simply becomes a Euclidean line segment between the two trees. Otherwise, we have to shrink edge lengths to zero and possibly expand other edges along the geodesic. An algorithm for computing geodesics was developed by Owen and Provan [2011].

The BHV tree space \mathcal{T}_n is therefore equipped with the induced *length metric* $d : \mathcal{T}_n \times \mathcal{T}_n \rightarrow [0, \infty)$, which makes it a geodesic metric space. A precise definition is given in section 2.3. Billera et al. [2001] proved that the tree space is a Hadamard space, which is a complete geodesic metric space and has nonpositive curvature (in the sense of Alexandrov). Intuitively, in such spaces triangles appear “slimmer” than a triangle in Euclidean space with the same edge lengths, see Figure 5. Already in Figure 4 one may ob-

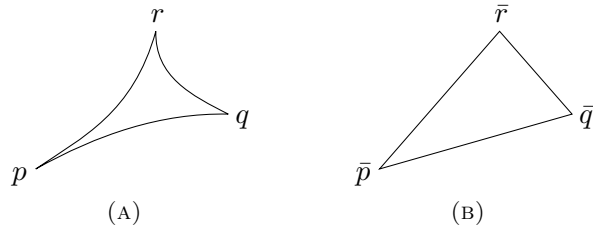


FIGURE 5. (a) Triangle in a space of nonpositive curvature.
 (b) Comparison triangle in Euclidean space.

serve that triangles might be slimmer than in Euclidean space if the vertices lie in different orthants. In a general Hadamard space many operations that

we know from Euclidean space are not defined. For instance, adding two points $t, s \in \mathcal{T}_n$ is not defined, however, the convex combination

$$t_\tau = (1 - \tau)s + \tau t,$$

for $\tau \in [0, 1]$ is a possible operation, which represents a unique tree $t_\tau \in \mathcal{T}_n$ lying on the geodesic from s to t satisfying $d(s, t_\tau) = \tau d(s, t)$. Convex combinations are important for computing medians and means, as will be outlined later. The BHV tree space however has much more structure than a general Hadamard space. The space is piecewise Euclidean and allows for instance to scale a point by a positive real constant, which is not defined in a Hadamard space, and we also have a well-defined origin. To gain a better understanding of the geometry of the space we first need to take a look at general metric spaces of nonpositive curvature.

2.3. Nonpositively curved spaces

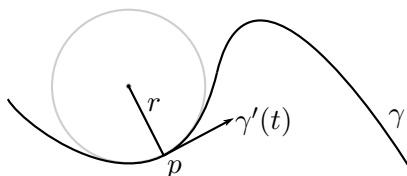
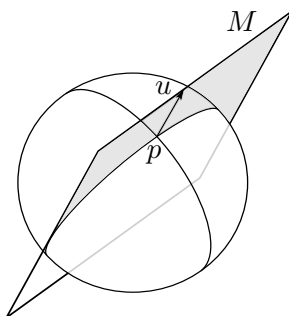
In this section we summarize some basic definitions and properties of nonpositively curved metric spaces, which can also be found in Bačák [2014b], Jost [1997], Bridson and Haefliger [1999], Davis [1994]. Many important results were also established by Jost [1997]. However, before discussing more abstract metric spaces, we have to develop some basic understanding of curvature (see also the discussion in Lee [1997]).

Given a curve $\gamma : I \rightarrow \mathbb{R}^2$, $I \subset \mathbb{R}$, in the Euclidean plane, the curvature at a point p on γ is a measure of how much γ locally deviates from a straight line. We assume that γ has unit speed parametrization, i.e. the arc length of γ from $\gamma(0)$ to $\gamma(t)$ is t . Formally, the curvature of γ is then defined as

$$\kappa^*(t) = \left\| \frac{d^2}{dt^2} \gamma(t) \right\|.$$

If γ is a straight line, the curvature is zero because the velocity is assumed to be constant. Otherwise, the velocity vector changes along γ , leading to a strictly positive curvature. Many applications require an extended definition of curvature, which takes both positive and negative values indicating to which side the curve is turning. We refer to this definition as the *signed curvature* $\kappa(t)$, so that $\kappa^*(t) = |\kappa(t)|$.

A more intuitive approach to understanding curvature is to consider tangent circles at a point $p = \gamma(t)$. The curve touches the circle at a point where the velocity vector on the circle coincides with $d/dt \gamma(t)$ (assuming unit velocity on the circle as well). The radius r of the circle is chosen such that the acceleration vector at p equals the acceleration on γ , see Figure 6, so that $\kappa^*(t) = r^{-1}$.

FIGURE 6. Osculating circle at a point $p = \gamma(t)$.FIGURE 7. A plane M that contains a point p on the sphere as well as the normal vector u at p .

The curvature of a surface S at a point p can in principle be obtained from the set of curves on S that go through p . However, we have to restrict the set of curves to those that are as straight as possible. Let u be the normal vector at p and M a plane that contains both p and u , see Figure 7. The set of curves that we consider are the intersections of all such planes M with the surface S . We then compute the signed curvature for each curve and take only the minimum and maximum, which we call the *principal curvatures* κ_1 and κ_2 at p . This leads to the definition of the *Gauss curvature*

$$\kappa_g = \kappa_1 \kappa_2.$$

It is easy to see that the curvature of a 2-sphere must be $1/r^2$. Negative curvature is obtained if the surface bends opposite in the principal directions. When studying more abstract metric spaces where curves are not differentiable, a notion of curvature is obtained by looking at the shape of triangles. As shown in Figure 8, triangles in negatively curved spaces appear slimmer than in Euclidean space. On the other hand, in positively curved spaces, the sides of triangles bend outwards (see also Toponogov's theorem).

A *metric space* (X, d) is a set of points X equipped with a metric (distance function) $d : X \times X \rightarrow [0, \infty)$. A *path* in this space is a continuous mapping $\gamma : [0, 1] \rightarrow X$ that connects two points $a, b \in X$, i.e. $\gamma(0) = a$ and $\gamma(1) = b$. A *geodesic* $\gamma : [0, 1] \rightarrow X$ is the shortest path between two

points such that $d(\gamma(s), \gamma(t)) = |s - t|d(a, b)$, for every $s, t \in [0, 1]$.² We may sometimes also use the shorthand notation $[a, b] : \tau \mapsto (1 - \tau)a + \tau b$ to refer to a geodesic from a to b . A metric space is called *geodesic* if every two points are the endpoints of a geodesic.

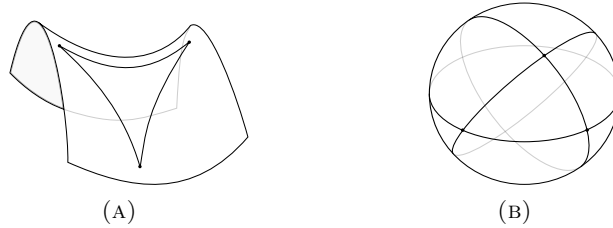


FIGURE 8. Triangles in curved surfaces. (A) Hyperbolic saddle with negative curvature. (B) Sphere with positive curvature.

Geodesic metric spaces are studied by comparing them to a model space M_κ^n , i.e. a smooth Riemannian manifold of dimension n with constant sectional curvature $\kappa \in \mathbb{R}$, i.e. the generalization of Gauss curvature to n -dimensional manifolds. We have to distinguish between the following three cases (see Figure 8):

- $\kappa < 0$: M_κ^n is the hyperbolic n -space,
- $\kappa = 0$: M_κ^n is the n -dimensional Euclidean space, and
- $\kappa > 0$: M_κ^n is the n -dimensional sphere with radius $1/\sqrt{\kappa}$.

Let (X, d) be a geodesic metric space and $\Delta(p, q, r)$ a geodesic triangle in this space. The comparison triangle $\Delta(\bar{p}, \bar{q}, \bar{r})$ is defined as a triangle in the model space M_κ^2 such that its edge lengths are equal to the edge lengths of $\Delta(p, q, r)$. The CAT(κ) inequality, named after Élie Cartan, Aleksandr Aleksandrov and Victor Toponogov, states that a triangle in X is “slimmer” than the comparison triangle in M_κ^2 . That is, for any vertex of $\Delta(p, q, r)$, say p , and any point $x_\tau = (1 - \tau)q + \tau r$ on the opposite edge we have

$$(1) \quad d(p, x_\tau) \leq d^*(\bar{p}, \bar{x}_\tau),$$

where \bar{x}_τ is the corresponding point on $\Delta(\bar{p}, \bar{q}, \bar{r})$ and d^* the metric on M_κ^2 .

DEFINITION 2.3.1 (CAT(κ) space). Let (X, d) be a geodesic metric space. X is said to be CAT(κ) if Equation 1 holds for any triangle.

Note also that in a CAT(κ) space X the curvature is bounded from above, which is why for any $\kappa' > \kappa$ the space X is also a CAT(κ') space.

²In this definition of the geodesic, the speed is given by the distance $d(\gamma(0), \gamma(1))$.

The above definition leads us to a bound on the *curvature* of a space. We say that a metric space has *curvature* $\leq \kappa$ if the $\text{CAT}(\kappa)$ inequality is satisfied locally. For smooth Riemannian manifolds it is known that this definition coincides with the sectional curvature, i.e. a smooth Riemannian manifold has curvature $\leq \kappa$ if and only if it has sectional curvature $\leq \kappa$.

Our interest here is mostly restricted to $\text{CAT}(0)$ spaces where the model space is the n -dimensional Euclidean space. The $\text{CAT}(0)$ inequality can be written as

$$d(p, x_\tau)^2 \leq (1 - \tau)d(p, q)^2 + \tau d(p, r)^2 - \tau(1 - \tau)d(q, r)^2.$$

From the $\text{CAT}(0)$ inequality it can also be concluded that the distance function d is convex, i.e. for any two geodesics $\sigma_1, \sigma_2 : [0, 1] \rightarrow X$ we have

$$d(\sigma_1(\tau), \sigma_2(\tau)) \leq (1 - \tau)d(\sigma_1(0), \sigma_2(0)) + \tau d(\sigma_1(1), \sigma_2(1)).$$

As a consequence, $\text{CAT}(0)$ spaces are also *uniquely geodesic*, meaning that any two points in X are connected by a unique geodesic. We also restrict our attention to *complete* spaces where every Cauchy sequence converges to some point *in* the space.

DEFINITION 2.3.2 (Hadamard space). A Hadamard space is a complete $\text{CAT}(0)$ space.

At first, Hadamard spaces seem very abstract and one may not expect to find instances in real applications. There is however a class of spaces with curvature bounded from above, which is constructed by “gluing” together pieces of Euclidean space. We will discuss this class in the context of simplicial complexes, but first we need to understand what is meant by the term “gluing” and how a metric is obtained on such spaces.

2.3.1. Gluing together metric spaces. For the following discussion, consider a family of metric spaces $(B_\lambda, d_\lambda)_\lambda$ with $\lambda \in \Lambda$ and let

$$\bar{B} = \coprod_{\lambda \in \Lambda} B_\lambda$$

denote the disjoint union with metric

$$\bar{d}((x, \lambda), (y, \lambda')) = \begin{cases} d_\lambda(x, y) & \text{if } \lambda = \lambda', \\ \infty & \text{otherwise.} \end{cases}$$

Furthermore, for some equivalence relation \sim let $B = \bar{B}/\sim$ denote the set of equivalence classes. The *quotient pseudometric* d on B is defined as

$$d(x, y) = \inf_S \sum_{i=1}^n \bar{d}(\bar{x}_i, \bar{y}_i)$$

where $S = (\bar{x}_1, \bar{y}_1, \dots, \bar{x}_n, \bar{y}_n)$ is a sequence of points such that $\bar{x}_1 \in x$, $\bar{y}_n \in y$ and $\bar{y}_i \sim \bar{x}_{i+1}$.

To define an equivalence relation we consider a simple way to combine metric spaces. Let $i_\lambda : A \rightarrow A_\lambda \subset B_\lambda$ be an isometry for all $\lambda \in \Lambda$. The equivalence relation \sim is defined through i_λ by

$$i_\lambda(a) \sim i_{\lambda'}(a) \quad \text{for all } a \in A, \lambda, \lambda' \in \Lambda,$$

and the resulting quotient space B is called a *gluing*. A simple example is illustrated in Figure 9, where B_1 and B_2 show two pieces of Euclidean space glued together along isometric subspaces. For a gluing B the quotient

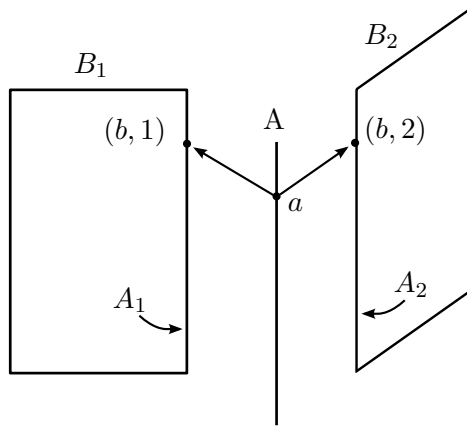


FIGURE 9. Two pieces of Euclidean space B_1 and B_2 glued together at a common face through an isometry on A .

pseudometric for some $x \in B_\lambda$ and $y \in B_{\lambda'}$ is given by

$$d(x, y) = \begin{cases} d_\lambda(x, y) & \text{if } \lambda = \lambda' \\ \inf_{a \in A} \{d_\lambda(x, i_\lambda(a)) + d_{\lambda'}(i_{\lambda'}(a), y)\} & \text{otherwise,} \end{cases}$$

which is a metric on B . When parts of Euclidean spaces are glued together, we may call d also the induced *length metric* of B .

2.3.2. Cubical and simplicial complexes. The BHV tree space can be understood as a collection of orthants glued together at isometric faces. To study the global curvature of this space, we first reduce it to a *cubical complex*. The global geometric properties of such a space are captured by the so called *link of the origin*, which is a *simplicial complex* and consists of all points at a fixed distance to the origin. This complex is then instrumentalized to show that the BHV tree space is CAT(0).

Let us first introduce some basic notions. A *simplex* $S \subset M_\kappa^n$ is defined as the convex hull of a set V of $n + 1$ points that are not contained in any

$(n - 1)$ -dimensional hyperplane (affinely independent). The elements of V are called the vertices of S and $T \subseteq S$ is called a *face* of S if it is the convex hull of a non-empty subset of V . A collection K of simplices is called an M_κ *simplicial complex* if it satisfies two conditions: A face of any simplex $S \in K$ is in K , and furthermore, every two simplices of K are allowed to intersect only at their faces (they are glued together at isometric faces). Note that the simplices of an M_κ simplicial complex are not required to be of the same dimension. A *cubical complex* C is a collection of cubes $I = [0, 1]^{n_I}$, $n_I \in \mathbb{N}^+$, where any two cubes may be glued together at isometric faces. The vertices of a cubical complex are the corners of the cubes in the collection. In the following we assume that both cubical and simplicial complexes consist of finitely many cubes or simplices respectively.

While curvature is a local characterization of a space, global statements, i.e. whether the space is $\text{CAT}(\kappa)$, are obtained using the Cartan-Hadamard theorem. This requires further assumptions, such as simple connectedness. Intuitively speaking, for global statements we need to assume that the space has no holes. An important tool for the study of global curvature of a cubical complex C is the *link* of a vertex v in C .

DEFINITION 2.3.3 (Geometric link). Let I be a cube $[0, 1]^{n_I}$. The *geometric link* (or simply *link*) $\text{Lk}_\epsilon(v, I)$ of a vertex v in I is defined as the set of points

$$\{i \in I \mid d(v, i) = \epsilon\},$$

for some $\epsilon \in (0, 1]$. Let C be cubical complex and C_v the collection of cubes that share a given vertex v due to the gluing. The link complex of a vertex v is defined as

$$\text{Lk}_\epsilon(v, C) = \coprod_{I \in C_v} \text{Lk}_\epsilon(v, I).$$

Take for instance the cube $I = [0, 1]^3$ and let v denote one of the vertices. The link of v in this cube is a subset of the sphere with radius ϵ (see Figure 10a). On the other hand, take a cubical complex C that consists of three cubes $[0, 1]^2$ glued together as shown in Figure 10b. The link complex of the vertex v consists of three quarters of a circle.

In general, the link complex $K_v = \text{Lk}_\epsilon(v, C)$ at a vertex v of a cubical complex C is an M_1 simplicial complex. A cubical complex C satisfies the *link condition* if the link complex K_v is a $\text{CAT}(1)$ space for every vertex v of C . An important result by Gromov [1987, p. 120] states the relation of link complexes to the global curvature of cubical complexes.

THEOREM 2.3.4. *A simply connected cubical complex is a $\text{CAT}(0)$ space if and only if it satisfies the link condition.*

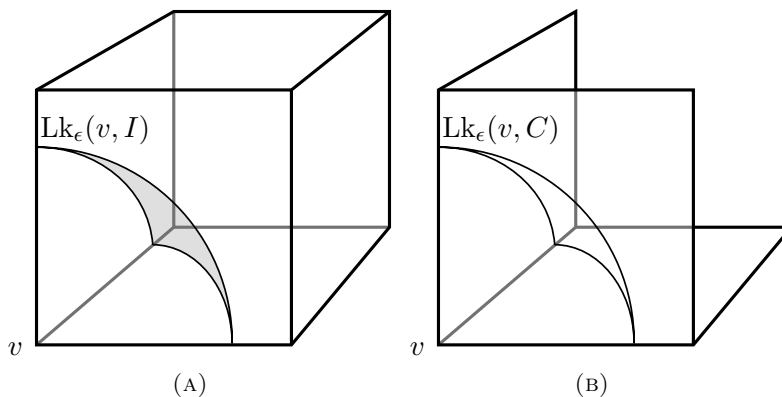


FIGURE 10. (A) Geometric link of a vertex in a three-dimensional cube. (B) Link complex of a vertex in a complex of three two-dimensional cubes.

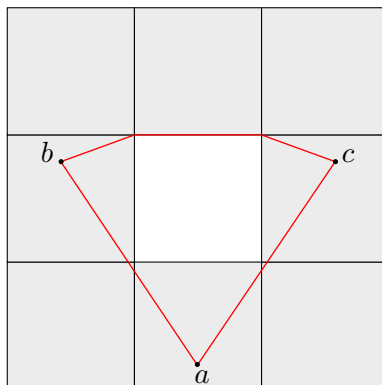


FIGURE 11. Triangle $\Delta(a, b, c)$ in a cubical complex, which is not simply connected.

In fact, the theorem is much more general than presented here and applies to polyhedral complexes. A space that is not simply connected contains “holes” and it is easy to verify that a triangle around such a hole does not satisfy the CAT(0) inequality (see Figure 11).

For instance, take again the cube $I = [0, 1]^3$. Since the link of any vertex is a subset of the unit sphere, we know that the link complex is CAT(1) and therefore the cube is a CAT(0) space. However, this is not true anymore if we only consider the surface of a cube. The link of a vertex is then isometric to a circle with circumference $3/2\pi$ and radius $3/4$ meaning that it is a CAT(16/9) space and therefore the surface of a cube is not CAT(0) space.

The link condition provides us with a powerful tool to check whether a space is CAT(0). What remains to be shown is that the link complex is a CAT(1) space. W.l.o.g. we will in the following assume that $\epsilon = 1$ for any link complex and use the shorthand notation $\text{Lk}(v, C) = \text{Lk}_1(v, C)$ instead.

DEFINITION 2.3.5 (All-right simplicial complex). An M_1 simplicial complex is called *all-right* if every edge has length $\pi/2$.

It is easy to verify that for instance the link complex of any cubical complex is an all-right simplicial complex. We have already observed that a cube is a CAT(0) space, however the surface of a cube is not. This is the basic observation that leads to the definition of a *flag complex*, which Gromov called the *no (empty) triangles condition*.

DEFINITION 2.3.6 (Flag complex). Let K be a simplicial complex. We call a set of vertices fully connected if the vertices are pairwise joined by an edge. K is called a flag complex if every fully connected set of vertices is also a simplex in K .

Hence, if for instance all the edges of a triangle are a simplex in K , then so is the full triangle. The same holds for tetrahedrons and so on. This leads us to the following theorem due to Gromov [1987].

THEOREM 2.3.7. *An all-right simplicial complex is CAT(1) if and only if it is a flag complex.*

This theorem is the most important result of this section, which we will use later to show that the BHV tree space is indeed a Hadamard space.

2.4. Combinatorics and geometry of the tree space

We have reviewed all basic facts that are required to gain a deeper understanding of the construction and geometry of the BHV tree space. As already discussed, the tree space \mathcal{T}_n consists of pieces of Euclidean space $[0, \infty)^{n-2}$, which we call orthants, glued together at isometric faces. We also saw, that for such a gluing an intrinsic length metric d exists. As was shown by Billera et al. [2001], an easy combinatorial argument exists which shows that the BHV tree space has nonpositive curvature.

To understand the curvature of the space we need to consider the link complex of every vertex. We restrict our attention to the link of the origin, which is the most important complex to study. The link is shown for a piece of \mathcal{T}_4 in Figure 12, which consists of subsets of the unit circle. The figure shows that $\{e_1, e_4\}$, $\{e_2, e_4\}$, and $\{e_3, e_4\}$ are pairwise compatible edge sets. The link is an M_1 simplicial complex where each edge has length $\pi/2$ and therefore the complex is all-right. One also observes that for instance e_1 and

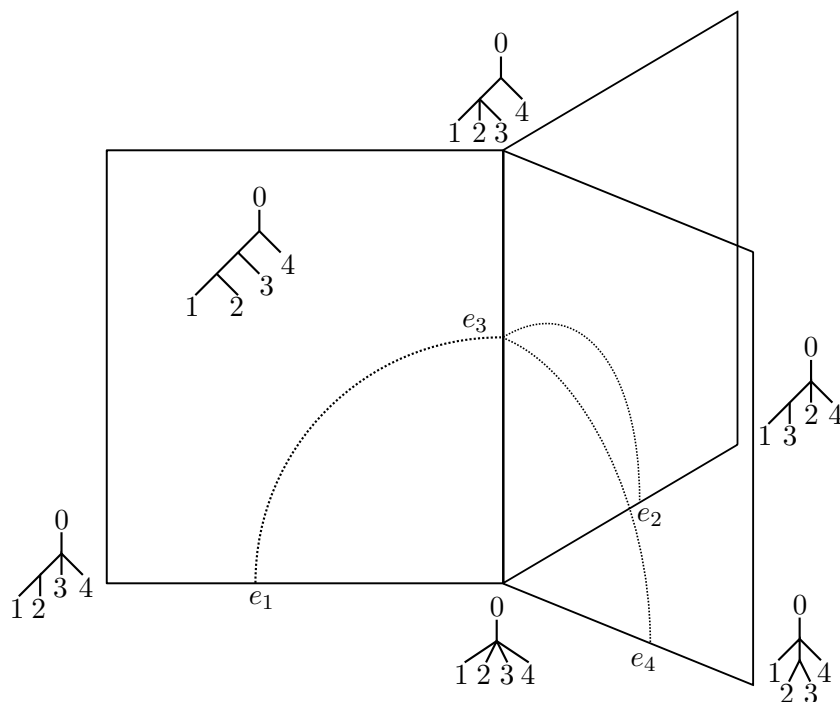


FIGURE 12. A piece of \mathcal{T}_4 with the link of the origin.

e_2 are not compatible and the space has no corresponding orthant. This also indicates that the link is a flag complex.

The following important theorem from Billera et al. [2001] states that the tree space has globally nonpositive curvature.

THEOREM 2.4.1. *The space \mathcal{T}_n is a Hadamard space.*

PROOF. The tree space is a gluing of orthants where the link of the origin is an all-right simplicial complex. To show that the space has nonpositive curvature we therefore only need to show that the link complex is a flag complex. A phylogenetic tree with $n + 1$ leaves consists of at most $n - 2$ pairwise compatible edges. Therefore, any set of pairwise compatible edges is represented by an orthant of the space, which proves the theorem. \square

We have established that the space has nonpositive curvature without an actual understanding of how to compute geodesics. Two trees lie in the same face of a simplex in tree space if they share the same topology. In this case, the geodesic is trivial to compute. Otherwise, computing geodesics turns out to be more difficult. A first glimpse is gained by understanding which orthants are glued together. A tree within an orthant has all $n - 2$ interior edges. Computing a path to a neighboring orthant would mean to

shrink one of the interior edge lengths to zero and extending one of the two other possible edges. In biology, a similar operation exists and is known as the *nearest neighbor interchange* (NNI), which is illustrated in Figure 13.

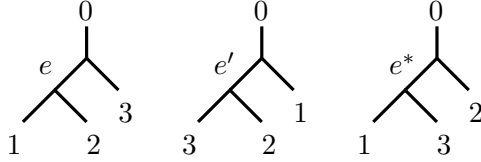


FIGURE 13. Nearest neighbor interchange (NNI). The edge e can be replaced by either e' or e^* . The leaves might represent more complex subtrees.

2.5. Computing means and medians in tree space

Given a set of trees $\bar{t} = (t_1, \dots, t_K)$ that are elements of the space \mathcal{T}_n , we would like to summarize this set with only one tree. A common approach is to compute for instance the mean or median. The tree space requires more general definitions of the mean and median that do not rely on any Euclidean properties. One commonly used possibility is given by the *Fréchet mean*

$$\Psi(\bar{t}) = \arg \min_{s \in \mathcal{T}_n} \Psi(\bar{t}; s) = \arg \min_{s \in \mathcal{T}_n} \frac{1}{K} \sum_{t \in \bar{t}} d(s, t)^2$$

and the *geometric median*

$$\Xi(\bar{t}) = \arg \min_{s \in \mathcal{T}_n} \Xi(\bar{t}; s) = \arg \min_{s \in \mathcal{T}_n} \frac{1}{K} \sum_{t \in \bar{t}} d(s, t).$$

It is easy to see that the Fréchet mean in Euclidean space is equivalent to the standard definition of the mean. Unlike the usual definition of the median, we do not require that the geometric median itself is an element of the set \bar{t} .

The most eminent question is how the minimizers of $\Psi(\bar{t})$ and $\Xi(\bar{t})$ can actually be computed. Take for instance three points $p, q, r \in \mathcal{T}_n$. The mean is the *barycenter* of the corresponding triangle $\Delta(p, q, r)$. In analogy to Euclidean space, we might be tempted to compute it as

$$\mu_1 = \frac{2}{3} \left(\frac{1}{2}p + \frac{1}{2}q \right) + \frac{1}{3}r.$$

However, we could also compute the mean as

$$\mu_2 = \frac{2}{3} \left(\frac{1}{2}p + \frac{1}{2}r \right) + \frac{1}{3}q,$$

and it is easy to verify that in general $\mu_1 \neq \mu_2$. An example is given in Figure 14.

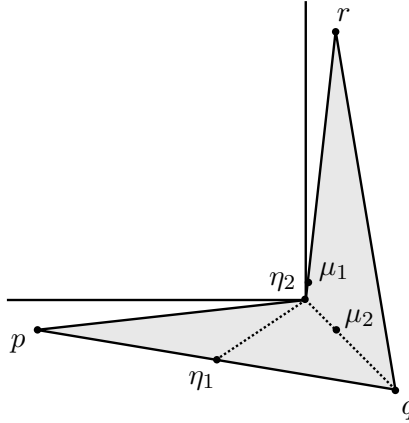


FIGURE 14. Triangle $\Delta(p, q, r)$ in tree space, with $\eta_1 = \frac{1}{2}p + \frac{1}{2}q$, $\mu_1 = \frac{2}{3}\eta_1 + \frac{1}{3}r$, $\eta_2 = \frac{1}{2}p + \frac{1}{2}r$, and $\mu_2 = \frac{2}{3}\eta_2 + \frac{1}{3}q$.

Both the mean and the median are convex functions and since tree space has much more structure than a general Hadamard space, we may use the gradient within orthants to find a minimizer. This approach seems much more promising but there are several practical difficulties. Both the Fréchet mean and geometric median in tree space have a property which is called *stickiness* [see e.g. Miller et al., 2012]. We will discuss this property in more detail later. As a result, the respective minimizers will often lie at the boundaries of orthants, where many directions of possible descent might exist. As an example take the space \mathcal{T}_3 with three trees $\bar{t} = (t_1, t_2, t_3)$ one on each orthant at an equal distance to the origin. To verify that the origin is the Fréchet mean we may formalize the conditions simply as the directional derivative

$$d_t \Psi(\bar{t}; \mathbf{0}) = \lim_{\tau \searrow 0} \frac{\Psi(\bar{t}; (1 - \tau)\mathbf{0} + \tau t) - \Psi(\bar{t}; \mathbf{0})}{\tau}$$

and check that it is nonnegative for each $t \in \bar{t}$. It should also be noted that the directional derivative is discontinuous at the origin. In this simple example we only have to check three conditions. It was pointed out by Miller et al. [2012] that in general we have to check all neighboring orthants for possible directions of descent, which in higher dimensions can easily become impractical. For the geometric median also the non-differentiability of the function $\Xi(\bar{t}, \cdot)$ at locations $t \in \bar{t}$ can cause numerical issues if \bar{t} contains many closely located trees.

2.5.1. The proximal point algorithm in Hadamard spaces. We utilize a modified version of the *proximal point algorithm* (PPA) to solve this type of minimization problem. Assume for the moment that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a convex differentiable function. Finding a minimizer of f can be interpreted as solving the differential equation

$$\frac{d}{dt}x(\tau) = -\nabla f(x(\tau)),$$

for $x : \mathbb{R}_+ \rightarrow \mathbb{R}^n$, which is called the *gradient flow* of f . A numerical solution may be found by using the *backward Euler method* to discretize the equation so that

$$\frac{x_{i+1} - x_i}{\lambda} = -\nabla f(x_{i+1}),$$

where $\lambda > 0$ is a step-size parameter. By rewriting the equation we obtain

$$\nabla f(x_{i+1}) + \frac{1}{\lambda}(x_{i+1} - x_i) = 0,$$

which can be expressed as

$$x_{i+1} = \arg \min_{y \in \mathbb{R}^n} \left(f(y) + \frac{1}{2\lambda} \|y - x_i\|_2^2 \right).$$

The second term ensures that x_{i+1} will be close to the previous solution x_i . It is similar to *Tikhonov regularization*, a widely used method in convex optimization [Boyd and Vandenberghe, 2004]. We call

$$J_\lambda(f; x) = \arg \min_{y \in \mathbb{R}^n} \left(f(y) + \frac{1}{2\lambda} \|y - x\|_2^2 \right)$$

the *resolvent* of f at x , which was introduced by Jost [1995, 1998] under the name of *Moreau-Yosida regularization*. Iterating the equation

$$x_{i+1} = J_{\lambda_i}(f; x_i)$$

with step size λ_i is known as the PPA [Martinet, 1970, Rockafellar, 1976]. For the algorithm to properly converge it is required that $\lambda_i > 0$ and $\sum \lambda_i = \infty$, which ensures that a minimum at a point arbitrarily far away from the starting point can be reached. The relation to the backward Euler discretization of the gradient flow is well known and for instance described in Parikh and Boyd [2013]. An important modification of the PPA was introduced by Bertsekas [2011]. Assume that f is of the form

$$f(x) = \sum_{k=1}^K f_k(x),$$

where $f_k : \mathbb{R}^n \rightarrow \mathbb{R}$, $k = 1, \dots, K$, are convex functions. Instead of applying the resolvent to f , it is possible to apply it to each component separately,

leading to the proximal iteration

$$x_{i+1} = J_{\lambda_i}(f_{k_i}; x_i),$$

where $k_i = (i \bmod K) + 1$. It is assumed that λ_i is constant within each cycle, i.e. $\lambda_i = \lambda_{i+1} = \dots = \lambda_{i+K-1}$ for all i with $k_i = 1$. We will refer to this algorithm in the following as the *cyclic incremental proximal point algorithm* (cyclic IPPA). The advantage of this method is that the resolvent of a single component f_k is often much easier to compute and in tree space it is possible to derive a closed form solution. Since in each iteration i the resolvent is applied to only one component f_{k_i} of f , it is necessary to introduce a further constraint on the step size λ_i that assures a global convergence of the cyclic IPPA. As will be discussed later, we will require λ_i to decrease with i such that $\sum \lambda_i^2$ converges and therefore λ_i converges to zero.

In a stochastic version of the algorithm, k_i is chosen at random from a uniform distribution. The convergence can be shown using the *supermartingale convergence theorem* [Bertsekas and Tsitsiklis, 1995]. Indeed, this algorithm is in its essence very similar to a variant of *stochastic gradient descent*, where at each iteration the gradient of a function f is approximated by the gradient of a randomly chosen component f_k [cf. e.g. Bottou, 1998]. Although in practice the stochastic variant often converges faster, in this discussion, we will focus our attention to the sequential version, since the convergence can be established with a less technical proof.

It is straightforward to generalize the resolvent to Hadamard spaces. Let (\mathcal{H}, d) be a Hadamard space and $f : \mathcal{H} \rightarrow \mathbb{R}$ a convex function. The resolvent of f with parameter $\lambda > 0$ is defined as

$$J_\lambda(f; x) = \arg \min_{y \in \mathcal{H}} \left(f(y) + \frac{1}{2\lambda} d(x, y)^2 \right).$$

The convergence of the respective PPA on locally compact Hadamard spaces was established in Bačák [2013] and the IPPA version in Bačák [2014a]. Obviously, the tree space \mathcal{T}_n is locally compact since it is composed of finitely many pieces of Euclidean space and we can therefore use the IPPA to compute medians and means.

We now state the formal definition of the cyclic and stochastic IPPA in locally compact Hadamard spaces.

DEFINITION 2.5.1 (cyclic IPPA). Let (\mathcal{H}, d) be a locally compact Hadamard space and let $f : \mathcal{H} \rightarrow \mathbb{R}$ with $f = \sum_{k=1}^K f_k$ be composed of K convex functions f_k . Given a starting point $x_0 \in \mathcal{H}$, in each iteration i a point

$$x_{i+1} = J_{\lambda_i}(f_{k_i}; x_i)$$

is computed, where $k_i = (i \bmod K) + 1$ and the step-size parameters λ_i are nonnegative reals such that

$$\sum_{i=1}^{\infty} \lambda_i = \infty, \quad \text{and} \quad \sum_{i=1}^{\infty} \lambda_i^2 < \infty.$$

Furthermore, we require a constant step-size within each cycle, i.e. $\lambda_i = \lambda_{i+1} = \dots = \lambda_{i+K-1}$ for all i with $k_i = (i \bmod K) + 1 = 1$.

DEFINITION 2.5.2 (stochastic IPPA). Let (\mathcal{H}, d) be a locally compact Hadamard space and let $f : \mathcal{H} \rightarrow \mathbb{R}$ with $f = \sum_{k=1}^K f_k$ be composed of K convex functions f_k . Given a starting point $x_0 \in \mathcal{H}$, in each iteration i a point

$$x_{i+1} = J_{\lambda_i}(f_{k_i}; x_i)$$

is computed, where k_i is drawn from a uniform distribution over $\{1, \dots, K\}$. The step-size parameters λ_i are nonnegative reals such that

$$\sum_{i=1}^{\infty} \lambda_i = \infty, \quad \text{and} \quad \sum_{i=1}^{\infty} \lambda_i^2 < \infty.$$

The following theorem due to Bačák [2014a] establishes the convergence of the cyclic IPPA. The preliminaries of the proof will be explained in detail afterwards.

THEOREM 2.5.3 (cyclic IPPA convergence). *Let (\mathcal{H}, d) be a locally compact Hadamard space and $f_k : \mathcal{H} \rightarrow \mathbb{R}$, $k = 1, \dots, K$, convex functions. Let $f = \sum_{k=1}^K f_k$ and assume that (x_i) is generated by the cyclic IPPA (see Definition 2.5.1). If $\min(f) \neq \emptyset$ and there exists an L such that*

$$(2) \quad f_k(x_{jK}) - f_k(x_{jK+k}) \leq Ld(x_{jK}, x_{jK+k})$$

$$(3) \quad f_k(x_{jK+k-1}) - f_k(x_{jK+k}) \leq Ld(x_{jK+k-1}, x_{jK+k})$$

for every $j \in \mathbb{N}$ and $k = 1, \dots, K$, then the sequence (x_i) converges to a minimizer of f .

PROOF. Let y be a minimizer of f . By Proposition 2.5.6 we have

$$d(x_{jK+K}, y)^2 \leq d(x_{jK}, y)^2 - 2\lambda_j [f(x_j) - f(y)] + 2\lambda_j^2 L^2 K(K+1),$$

for any $j \in \mathbb{N}$. According to Lemma 2.5.7 we know that the sequence $(d(x_{jK}, y))$ converges as $j \rightarrow \infty$. Furthermore, also by Lemma 2.5.7 the series

$$\sum_{j=0}^{\infty} \lambda_{jK} [f(x_{jK}) - f(y)]$$

converges. By Lemma 2.5.8 and because $\sum \lambda_i = \infty$ we know that there exists a subsequence $(x_{j_l K})$ of (x_{jK}) for which $f(x_{j_l K})$ converges to $f(y)$ as

$l \rightarrow \infty$. Since y might not be a unique minimizer of f , we may exploit that the space is locally compact and find that because $(x_{j_i K})$ is bounded, it has a subsequence that converges to a minimizer \hat{y} of f . Although Lemma 2.5.5 states that successive points of the sequence (x_i) get arbitrarily close as $i \rightarrow \infty$, it cannot be directly concluded that the whole sequence is Cauchy (see e.g. Figure 15). We have already established that $(d(x_{j_K}, \hat{y}))$ converges as $j \rightarrow \infty$ and since a subsequence of (x_{j_K}) converges to \hat{y} , we can conclude that

$$d(x_{j_K}, \hat{y}) \rightarrow 0.$$

In addition, again by virtue of Lemma 2.5.5 we know that the distance between successive points of (x_i) converges to zero and therefore we can conclude that

$$\lim_{j \rightarrow \infty} x_{jK+k} = \hat{y}$$

for every $k = 1, \dots, K$, from which it follows that the whole sequence (x_i) converges to \hat{y} as $i \rightarrow \infty$. \square

Conditions 2 and 3 are satisfied for instance if the functions f_k are Lipschitz continuous. While this is true for the distance function, it does not hold for the squared distance. When considering the mean of a set of trees $\bar{t} = \{t_1, \dots, t_K\}$, we know that the minimizer has to be contained in the convex hull spanned by \bar{t} . The sequence (x_i) is fully contained in the closed convex hull of $\bar{t} \cup \{x_0\}$, which is a bounded subset of \mathcal{H} . Within this region the squared distance is indeed also Lipschitz and therefore the conditions are met. In particular, the conditions are satisfied for functions of the form

$$f(y) = \frac{1}{K} \sum_{k=1}^K d(t_k, y)^p, \quad \text{where } p \in [1, \infty).$$

The convergence of the cyclic IPPA relies on the following properties of the resolvent.

LEMMA 2.5.4. *Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be a convex function and $x_{i+1} = J_{\lambda_i}(f; x_i)$. For any $y \in \mathcal{H}$*

$$d(x_{i+1}, y)^2 - d(x_i, y)^2 \leq 2\lambda_i [f(y) - f(x_{i+1})].$$

PROOF. By the definition of the resolvent we have

$$x_{i+1} = J_{\lambda_i}(f; x_i) = \arg \min_{y \in \mathcal{H}} \left(f(y) + \frac{1}{2\lambda_i} d(x_i, y)^2 \right)$$

and therefore

$$f(x_{i+1}) + \frac{1}{2\lambda_i} d(x_i, x_{i+1})^2 \leq f(y_p) + \frac{1}{2\lambda_i} d(x_i, y_p)^2$$

for any $y_p \in \mathcal{H}$, which we rewrite as

$$d(x_i, x_{i+1})^2 - d(x_i, y_p)^2 \leq 2\lambda_i [f(y_p) - f(x_{i+1})] .$$

For some $y \in \mathcal{H}$ and $y_\tau = (1 - \tau)y + \tau x_{i+1}$, $\tau \in [0, 1)$, we apply the CAT(0) inequality

$$d(x_i, y_\tau)^2 \leq (1 - \tau)d(x_i, y)^2 + \tau d(x_i, x_{i+1})^2 - \tau(1 - \tau)d(x_{i+1}, y)^2$$

and obtain

$$\begin{aligned} (1 - \tau)d(x_i, x_{i+1})^2 - (1 - \tau)d(x_i, y)^2 + \tau(1 - \tau)d(x_{i+1}, y)^2 \\ \leq 2\lambda_i [f(y_\tau) - f(x_{i+1})] . \end{aligned}$$

Since f is convex, we have

$$f(y_\tau) - f(x_{i+1}) \leq (1 - \tau)f(y) - f(x_{i+1})$$

and therefore

$$d(x_i, x_{i+1})^2 - d(x_i, y)^2 + \tau d(x_{i+1}, y)^2 \leq 2\lambda_i [f(y) - f(x_{i+1})] .$$

Setting $\tau = 1$ and dropping $d(x_i, x_{i+1})^2$ finishes the proof. \square

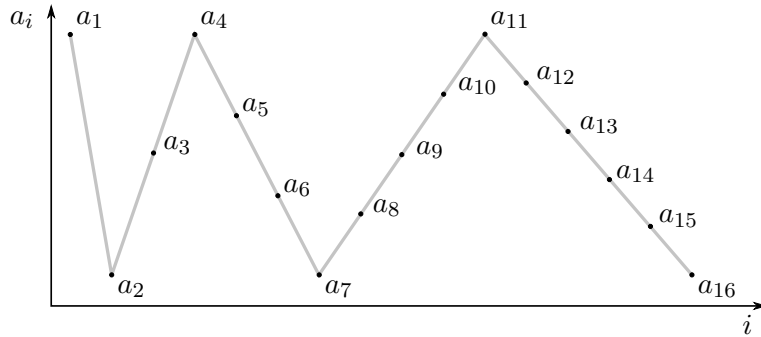


FIGURE 15. An example of a real sequence (a_i) , which is not Cauchy but where $|a_i - a_{i+1}| \rightarrow 0$. The subsequences a_1, a_4, a_{11}, \dots and a_2, a_7, a_{16}, \dots are constant and therefore converge to different limits.

The following result is important since it establishes that the distance of successive elements of (x_i) converges to zero if $\lambda_i \rightarrow 0$. However, this does not imply that the sequence is Cauchy (see e.g. Figure 15).

LEMMA 2.5.5. *Let $f : \mathcal{H} \rightarrow \mathbb{R}$ be a convex function and $x_{i+1} = J_{\lambda_i}(f; x_i)$. If there exists an L such that*

$$f(x_i) - f(x_{i+1}) \leq Ld(x_i, x_{i+1})$$

then

$$d(x_i, x_{i+1}) \leq 2\lambda_i L.$$

PROOF. By the definition of the algorithm or by applying Lemma 2.5.4 with $y = x_i$ we get

$$d(x_i, x_{i+1})^2 \leq 2\lambda_i [f(x_i) - f(x_{i+1})]$$

which can be written as

$$d(x_i, x_{i+1}) \leq 2\lambda_i \frac{f(x_i) - f(x_{i+1})}{d(x_i, x_{i+1})}$$

and it follows that

$$d(x_i, x_{i+1}) \leq 2\lambda_i L.$$

□

PROPOSITION 2.5.6. *Let $f_k : \mathcal{H} \rightarrow \mathbb{R}$, $k = 1, \dots, K$, be convex functions and let $f = \sum_{k=1}^K f_k$. Assume that (x_i) is generated by the cyclic IPPA (see Definition 2.5.1). If there exists an L such that*

$$(4) \quad f_k(x_{jK}) - f_k(x_{jK+k}) \leq Ld(x_{jK}, x_{jK+k})$$

$$(5) \quad f_k(x_{jK+k-1}) - f_k(x_{jK+k}) \leq Ld(x_{jK+k-1}, x_{jK+k})$$

for every $j \in \mathbb{N}$ and $k = 1, \dots, K$, then

$$d(x_{jK+K}, y)^2 \leq d(x_{jK}, y)^2 - 2\lambda_i [f(x_{jK}) - f(y)] + 2\lambda_{jK}^2 L^2 K(K+1).$$

PROOF. For some $j \in \mathbb{N}$ let $i = jK$. We have

$$x_{i+k} = J_{\lambda_i}(f_k; x_{i+k-1})$$

and by applying Lemma 2.5.4 we obtain

$$d(x_{i+k}, y)^2 \leq d(x_{i+k-1}, y)^2 - 2\lambda_i [f_k(x_{i+k}) - f_k(y)].$$

By adding $2\lambda_i [f_k(x_i) - f_k(x_i)]$ to the right side, the inequality can be written as

$$d(x_{i+k}, y)^2 \leq d(x_{i+k-1}, y)^2 - 2\lambda_i [f_k(x_i) - f_k(y)] + 2\lambda_i [f_k(x_i) - f_k(x_{i+k})].$$

By assumption 4 we know that

$$f_k(x_i) - f_k(x_{i+k}) \leq Ld(x_i, x_{i+k}).$$

Utilizing assumption 5 with Lemma 2.5.5 results in

$$Ld(x_i, x_{i+k}) \leq L[d(x_i, x_{i+1}) + \dots + d(x_{i+k-1}, x_{i+k})] \leq 2\lambda_i L^2 k$$

and it follows that

$$d(x_{i+k}, y)^2 \leq d(x_{i+k-1}, y)^2 - 2\lambda_i [f_k(x_i) - f_k(y)] + 4\lambda_i^2 L^2 k.$$

Summing on both sides over k leads to

$$\sum_{k=1}^K d(x_{i+k}, y)^2 \leq \sum_{k=1}^K d(x_{i+k-1}, y)^2 - 2\lambda_i [f(x_i) - f(y)] + 2\lambda_i^2 L^2 K(K+1),$$

which is equivalent to

$$d(x_{i+K}, y)^2 \leq d(x_i, y)^2 - 2\lambda_i [f(x_i) - f(y)] + 2\lambda_i^2 L^2 K(K+1).$$

□

The proof of convergence for the IPPA relies on the following lemma from Bertsekas and Tsitsiklis [1995].

LEMMA 2.5.7. *Let (a_i) , (b_i) , and (c_i) be sequences of nonnegative real numbers. If*

$$a_{i+1} \leq a_i - b_i + c_i$$

for all $i \in \mathbb{N}$ and

$$\sum_{i=1}^{\infty} c_i < \infty,$$

then the sequence (a_i) converges and $\sum_{i=1}^{\infty} b_i < \infty$.

PROOF. Summing the equations

$$\begin{aligned} a_{i+1} &\leq a_i - b_i + c_i \\ a_{i+2} &\leq a_{i+1} - b_{i+1} + c_{i+1} \\ &\vdots \\ a_{l+1} &\leq a_i - b_l + c_l \end{aligned}$$

for some $l > i$ leads to

$$a_{l+1} \leq a_i - \sum_{j=i}^l b_j + \sum_{j=i}^l c_j$$

and taking the $\limsup_{l \rightarrow \infty}$ results in

$$\begin{aligned} \limsup_{l \rightarrow \infty} a_{i+1} &\leq a_i + \limsup_{l \rightarrow \infty} \left[- \sum_{j=i}^l b_j + \sum_{j=i}^l c_j \right] \\ &\leq a_i + \limsup_{l \rightarrow \infty} \left[- \sum_{j=i}^l b_j \right] + \sum_{j=i}^{\infty} c_j \\ &\leq a_i + \sum_{j=i}^{\infty} c_j \end{aligned}$$

because the sequence (c_i) converges. Taking the $\liminf_{i \rightarrow \infty}$ shows that (a_i) converges, i.e.

$$\limsup_{n \rightarrow \infty} a_{i+1} \leq \liminf_{i \rightarrow \infty} a_i.$$

Since

$$\sum_{j=1}^l b_j \leq a_1 - a_{l+1} + \sum_{j=1}^l c_j$$

for any $l \in \mathbb{N}$, the sequence (b_i) also converges. \square

The following lemma is very simple, but we include it for the readers convenience.

LEMMA 2.5.8. *Let (λ_i) and (a_i) be two nonnegative real sequences. Assume that $\sum_i \lambda_i = \infty$ and let $s_l = \sum_{i=1}^l \lambda_i a_i$. If*

$$\lim_{l \rightarrow \infty} s_l < \infty$$

then there exists a subsequence (a_{i_j}) of (a_i) that converges to zero.

PROOF. Assume that $\liminf a_i = a \geq 0$, then $a_i \geq a$ for all sufficiently large i . The series $\sum_i \lambda_i a$ diverges if $a > 0$. \square

It is important to see that only a subsequence of (a_i) converges to zero. For instance, assume that $\lambda_i = 1/i$ and

$$a_i = \begin{cases} 1 & \text{if } i \text{ is a perfect square,} \\ 0 & \text{otherwise.} \end{cases}$$

Hence, $\sum_i \lambda_i a_i = \sum_i 1/i^2 < \infty$, but (a_i) does not converge to zero.

REMARK 2.5.1. For the sake of simplicity, we have only considered convex functions with codomain \mathbb{R} . It should be noted that the convergence results of the PPA and IPPA also hold for convex lower semicontinuous functions with codomain $(-\infty, \infty]$.

2.5.2. Computing means and median in tree space. So far we have determined that the cyclic IPPA is an appropriate algorithm for computing means and medians in tree space. We have also shown that the algorithm converges to a minimizer in locally compact Hadamard spaces. For the actual computation, the explicit form of the resolvent in each step has to be determined. In each step of the cyclic IPPA a new point

$$x_{i+1} = J_{\lambda_i}(f_{k_i}; x_i)$$

is determined. Let $\bar{t} = (t_1, \dots, t_K)$ denote a set of trees in \mathcal{T}_n . For the mean we get

$$x_{i+1} = \arg \min_{s \in \mathcal{T}_n} \left[d(t_{k_i}, s)^2 + \frac{1}{2\lambda_i} d(x_i, s)^2 \right].$$

The new point x_{i+1} has to lie somewhere on the geodesic between the old point x_i and t_{k_i} . Therefore, we may assume that

$$x_{i+1} = (1 - \tau_i)x_i + \tau_i t_{k_i},$$

so that the parameter τ_i has to be determined. Computing the derivative at s in the direction of t_{k_i} gives the condition

$$\begin{aligned} 0 &= -2d(t_{k_i}, s) + \frac{1}{\lambda_i}d(x_i, s) \\ &= -2d(t_{k_i}, x_i)(1 - \tau_i) + \frac{1}{\lambda_i}d(x_i, t_{k_i})\tau_i \\ &= -2(1 - \tau_i) + \frac{1}{\lambda_i}\tau_i, \end{aligned}$$

from which it follows that

$$\tau_i = \frac{2\lambda_i}{1 + 2\lambda_i}.$$

An illustration of the resulting cyclic IPPA is shown in Figure 16.

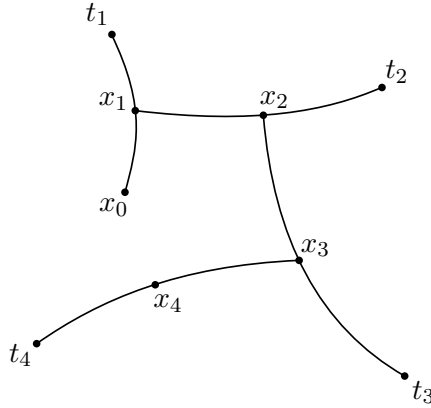


FIGURE 16. Illustration of the cyclic IPPA for computing the mean of a set of trees $\bar{t} = (t_1, t_2, t_3, t_4)$.

Similarly, for the median we have

$$x_{i+1} = \arg \min_{s \in \mathcal{T}_n} \left[d(t_{k_i}, s) + \frac{1}{2\lambda_i}d(x_i, s)^2 \right],$$

and therefore

$$\tau_i = \min \left\{ 1, \frac{\lambda_i}{d(x_i, t_{k_i})} \right\}.$$

To evaluate the mean and the median require to compute the geodesic between pairs of trees in each iteration. An algorithm for doing so will be discussed in the following.

2.6. Computing geodesics in tree space

The problem of computing geodesics was first addressed by Billera et al. [2001], but an algorithm was not found until much later by Owen and Provan [2011]. This algorithm is essential for the computation of means and medians of a set of points in BHV tree space.

Let us first recall some basic properties of geodesics in tree space that were already observed by Billera et al. [2001]. We know that the BHV tree space is $\text{CAT}(0)$ and therefore a unique geodesic exists between any two points. For the actual computation, we have to distinguish between several cases.

The first case we need to consider is when two trees $t, s \in \mathcal{T}_n$ contain the same edge e . Both trees might however assign a different length to that edge. The geodesic in this dimension would then simply be the Euclidean line segment while the remaining parts of the trees can be treated independently. In fact, we may cut the trees somewhere at the edge e in half and obtain a set of trees $\{t_1, t_2\}$ from t and $\{s_1, s_2\}$ from s . The geodesic between t and s is given by the combination of geodesics $[t_1, s_1]$ and $[t_2, s_2]$. In the following discussion, we will therefore restrict our attention to the case where the two trees do not share a common edge. We call such trees *disjoint*.

An extreme case is given when no edge of a tree t is compatible with any of the edges of a tree s . Assume for instance that $t, s \in \mathcal{T}_4$ and t has edges

$$e_5 : (0, 1 \mid 2, 3, 4), \quad e_6 : (0, 1, 2 \mid 3, 4),$$

whereas s contains a single edge $e_3 : (0, 4 \mid 1, 2, 3)$. As can be easily checked, e_3 is not compatible with any of the two edges of t . In this case it is easy to see that the geodesic between t and s consists of first shrinking the interior edges of t to length zero and afterwards extending the single interior edge of s (see also Figure 4). We will call a path that goes through the origin a *cone path*.

The last and most important case is when two trees $t, s \in \mathcal{T}_n$ are disjoint, but when there exist edges e of t and f of s which are compatible. In this case, the geodesic might not be a cone path and would go through an orthant of a tree that contains both e and f .

EXAMPLE 2.6.1. Let $t, t' \in \mathcal{T}_4$ have edges $e_3 : (0, 4 \mid 1, 2, 3)$ and $e_4 : (0, 1, 4 \mid 2, 3)$, and let $s, s' \in \mathcal{T}_4$ have edges $e_5 : (0, 1 \mid 2, 3, 4)$, and $e_6 : (0, 1, 2 \mid 3, 4)$. The locations of the four trees in \mathcal{T}_4 are depicted in Figure 17. The two trees t and s are located such that the geodesic $[t, s]$ goes through the origin, i.e. it is a cone path. However, the geodesic $[t', s']$ is not a cone path and goes through the interior of the orthant $\mathcal{O}(\{e_3, e_5\})$.

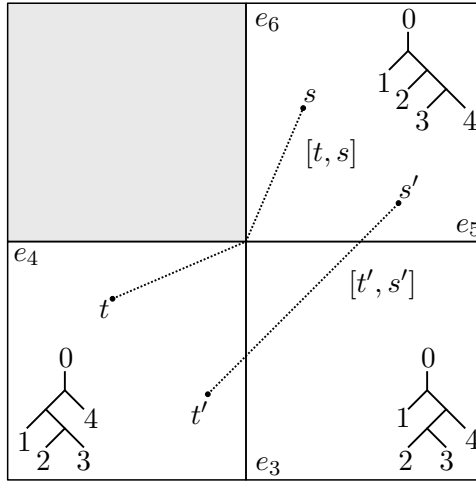


FIGURE 17. A piece of \mathcal{T}_4 with two geodesics $[t, s]$ and $[t', s']$. The upper left quadrant is not accessible since e_4 and e_6 are not compatible.

Figure 17 also suggests that an easy criterion exists for whether the cone path is a geodesic. For some $t \in \mathcal{T}_n$ let t_{Lk} denote the projection of t onto the link of the origin $\text{Lk}(\mathbf{0}, \mathcal{T}_n)$. Since every tree lies on a unique ray from the origin, the projection is simple and well-defined. The angle $\angle(t, s)$ between two trees $t, s \in \mathcal{T}_n$ is given as

$$\angle(t, s) = d_{\text{Lk}}(t_{\text{Lk}}, s_{\text{Lk}}),$$

where $d_{\text{Lk}}(\cdot, \cdot)$ is the spherical metric on $\text{Lk}(\mathbf{0}, \mathcal{T}_n)$. In Example 2.6.1 the geodesic between t and s is the cone path, because $\angle(t, s) \geq \pi$. To compute the angle, we identify the orthants $\mathcal{O}(t)$ and $\mathcal{O}(s)$ respectively as the totally negative and totally positive orthants of the 2-dimensional Euclidean space. Furthermore, let $\angle_E(\cdot, \cdot)$ denote the Euclidean angle between two vectors. The cone path between trees t and s is not a geodesic if

$$\angle(t, s) = \angle_E((|e_5|_s, |e_6|_s), (1, 0)) + \angle_E((-|e_4|_t, -|e_3|_t), (1, 0)) < \pi.$$

Using basic geometry, this condition may be rewritten to obtain the simple criterion

$$\frac{|e_4|_t}{|e_5|_s} < \frac{|e_3|_t}{|e_6|_s}.$$

The intuitive interpretation is that if this condition is met, then on the geodesic from t to s we would first replace the edge e_4 by e_5 and afterwards e_3 by e_6 .

Billera et al. [2001, Proposition 4.3] generalized this condition to higher dimensions. For some $t \in \mathcal{T}_n$ let A be a subset of the edges of t . We denote

by

$$\|A\|_t = \sqrt{\sum_{e \in A} |e|_t^2}$$

the Euclidean norm of the edge set A .

PROPOSITION 2.6.1. *Let $t, s \in \mathcal{T}_n$ be two disjoint trees and assume that $A = \{A_1, A_2\}$ and $B = \{B_1, B_2\}$ respectively are bipartitions of the edge sets of t and s such that $B_1 \cup A_2$ is a compatible set of edges. The cone path is not a geodesic if*

$$\frac{\|A_1\|_t}{\|B_1\|_s} < \frac{\|A_2\|_t}{\|B_2\|_s}.$$

Let $t(A)$ denote the tree t with all edges removed except for those in A . If the condition of Proposition 2.6.1 is satisfied for two trees $t, s \in \mathcal{T}_n$, then the angle $\angle(t(A_2), s(B_1))$ is less than π . Informally, this proposition provides a condition for when a path through an orthant $\mathcal{O}(B_1 \cup A_2)$ is shorter than the cone path. Also in this case, on the geodesic from t to s we would first replace all edges in A_1 by those in B_1 and afterwards the edges A_2 by B_2 . If one considers again Example 2.6.1 it can be easily verified that this condition is satisfied for the trees t' and s' .

Billera et al. [2001, Proposition 4.1] also showed the following important result:

PROPOSITION 2.6.2. *Let $t, s \in \mathcal{T}_n$ be two disjoint trees. If the cone path between t and s is not a geodesic, then there exist non-trivial partitions of the edge sets of t and s , respectively denoted $A = (A_1, \dots, A_k)$ and $B = (B_1, \dots, B_k)$, such that*

- $C_i(A, B) = B_1 \cup \dots \cup B_i \cup A_{i+1} \cup \dots \cup A_k$ is a compatible set of edges for all $i < k$, and
- the geodesic $[t, s]$ traverses each orthant $\mathcal{O}(C_i(A, B))$ sequentially for $i = 1, \dots, k$.

A geodesic $[t, s]$ is therefore fully contained in the union of orthants $\bigcup_{i=1}^k \mathcal{O}(C_i(A, B)) \subset \mathcal{T}_n$. The proposition however merely states the existence of such a sequence of orthants, but does not provide a way to find it. Owen [2011] developed further necessary and sufficient conditions that such a sequence must satisfy, which lead to the development of an efficient iterative algorithm [cf. Owen and Provan, 2011]. We summarize the results in the following theorem.

THEOREM 2.6.3. *Let $t, s \in \mathcal{T}_n$ be two disjoint trees. Let $A = (A_1, \dots, A_k)$ respectively $B = (B_1, \dots, B_k)$ be nontrivial partitions of the edge sets of t and s such that*

- (1) $C_i(A, B) = B_1 \cup \dots \cup B_i \cup A_{i+1} \cup \dots \cup A_k$ is a compatible set of edges for all $i < k$,
- (2) the partitions satisfy

$$\frac{\|A_1\|_t}{\|B_1\|_s} < \frac{\|A_2\|_t}{\|B_2\|_s} < \dots < \frac{\|A_k\|_t}{\|B_k\|_s}, \quad \text{and}$$

- (3) for every $i \in 1, \dots, k$ there exists no nontrivial bipartition $C_1 \cup C_2$ of A_i and $D_1 \cup D_2$ of B_i such that $D_1 \cup C_2$ is a compatible set of edges and

$$\frac{\|C_1\|_t}{\|D_1\|_s} < \frac{\|C_2\|_t}{\|D_2\|_s}.$$

The geodesic $\gamma = [t, s]$ traverses each orthant $\mathcal{O}(C_i(A, B))$ for $i = 1, \dots, k$ in the given order. The pair (A, B) is called the support of γ .

Informally, this theorem provides us with a sequential order of how edges are replaced on the geodesic from t to s and there exists a simple condition for when the union of orthants $\mathcal{O}(C_i(A, B))$ in fact contains the geodesic. The theorem can be easily turned into an iterative algorithm. We begin with the cone path, therefore the partitions A and B both consist of one set, which is the edge sets of t and s . In each iteration, we refine the partitions A and B by searching for bipartitions that match condition (iii) until no such bipartitions can be found. The geodesic is then given as follows:

LEMMA 2.6.4. For a geodesic $\gamma = [t, s]$ let the partitions $A = (A_1, \dots, A_k)$ and $B = (B_1, \dots, B_k)$ satisfy the conditions of Theorem 2.6.3. The geodesic γ is given by a sequence of legs (γ_i) , $i = 0, \dots, k$, such that

$$\gamma(\tau) = \gamma_i(\tau) \quad \text{if} \quad \frac{\|A_i\|_t}{\|B_i\|_s} < \frac{\tau}{1-\tau} \leq \frac{\|A_{i+1}\|_t}{\|B_{i+1}\|_s},$$

where γ_i is associated with a tree $t' \in \mathcal{O}(C_i(A, B))$. The interior edge lengths of t' are given by

$$|e|_{t'} = \begin{cases} \frac{(1-\tau)\|A_j\|_t - \tau\|B_j\|_s}{\|A_j\|_t} & \text{if } e \in A_j \text{ and } j > i, \\ \frac{\tau\|B_j\|_s - (1-\tau)\|A_j\|_t}{\|B_j\|_t} & \text{if } e \in B_j \text{ and } j \leq i. \end{cases}$$

The length of γ is given by

$$\text{length}(\gamma) = \sqrt{\sum_{i=1}^k (\|A_i\|_t + \|B_i\|_s)^2}.$$

2.6.1. Computing bipartitions. We now turn our attention to the problem of finding partitions of the edge sets that satisfy all three conditions of Theorem 2.6.3. This problem can be mapped to finding a *minimum weighted vertex cover* of a bipartite graph, for which efficient algorithms already exist.

Assume that $A = (A_1, \dots, A_k)$ and $B = (B_1, \dots, B_k)$ are partitions that satisfy conditions (i) and (ii) of Theorem 2.6.3 and describe the current path from t to s . For each $i \in \{1, \dots, k\}$ we try to find a nontrivial bipartition $C_1 \cup C_2$ of A_i and $D_1 \cup D_2$ of B_i , such that $D_1 \cup C_2$ are pairwise compatible and

$$(6) \quad \frac{\|C_1\|_t}{\|D_1\|_s} < \frac{\|C_2\|_t}{\|D_2\|_s}.$$

To find a compatible set of edges $D_1 \cup C_2$ we construct an *incompatibility graph* $G(A_i, B_i)$ as follows: For each edge in A_i and B_i there is a vertex in the graph. If $e \in A_i$ and $f \in B_i$ are incompatible, then the respective vertices in $G(A_i, B_i)$ are joined by an edge. It is easy to see that the resulting graph is bipartite (see e.g. Figure 18). A *vertex cover* is a set of vertices V such that every edge in $G(A_i, B_i)$ is incident to at least one vertex in V . The *minimum vertex cover* is a vertex set V with minimal cardinality. The complement of a minimum vertex cover forms an independent set of vertices in the graph and therefore is a set of pairwise compatible edges.

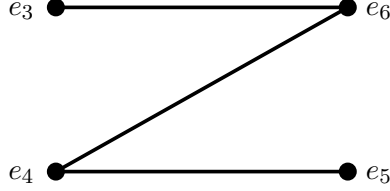


FIGURE 18. Incompatibility graph $G(\{e_3, e_4\}, \{e_5, e_6\})$ for Example 2.6.1. The minimum vertex cover is given by $\{e_4, e_6\}$ and its complement is a set of pairwise compatible edges.

The problem is not to find any independent set, but one that also satisfies Equation 6. W.l.o.g. we may assume that $\|A_i\|_t = \|B_i\|_s = 1$, such that

$$\frac{\|C_1\|_t}{\|D_1\|_s} < \frac{\|C_2\|_t}{\|D_2\|_s} \Leftrightarrow \frac{1 - \|C_2\|_t^2}{\|D_1\|_s^2} < \frac{\|C_2\|_t^2}{1 - \|D_1\|_s^2},$$

which may also be written as

$$\|C_2\|_t^2 + \|D_1\|_s^2 = \sum_{e \in C_2} |e|_t^2 + \sum_{f \in D_1} |f|_s^2 > 1.$$

Therefore, each edge length might be interpreted as a weight for the corresponding vertex in $G(A_i, B_i)$ and our objective is to find a *minimum weighted vertex cover* $V = C_1 \cup D_2$ that satisfies

$$\|C_1\|_t^2 + \|D_2\|_s^2 < 1.$$

If such a cover is found, the complement is used to refine the partitions A_i and B_i . Thereby, the current path from t to s will be shortened by also visiting the orthant $\mathcal{O}(C_2 \cup D_1)$.

For general graphs the problem of finding a minimum vertex cover is known to be NP-hard. However, since the incompatibility graph is bipartite, a solution can be found in polynomial time. A common approach to solve the problem is to transform it into a *flow network*. Utilizing the *max-flow min-cut theorem* [Lawler, 1976], we observe that a solution to the maximum flow problem indeed solves the vertex cover problem. As suggested by Owen and Provan [2011], the problem can therefore be solved with the *push relabel algorithm* due to Goldberg and Tarjan [1988]. A description of a slightly more general version of the algorithm with cubic complexity can be found in Ahuja et al. [1993, Section 7.7].

The problem can also be stated as an *integer linear program* (ILP). For each vertex v of the graph let x_v denote a variable that takes values in $\{0, 1\}$ and indicates whether v is part of the vertex cover. The ILP is given by

$$\begin{aligned} \text{minimize} \quad & \sum_{e \in A_i} x_e |e|_t + \sum_{f \in B_i} x_f |f|_s \\ \text{subject to} \quad & x_e + x_f \geq 1 \text{ if } e \text{ and } f \text{ are incompatible.} \end{aligned}$$

2.7. Phylogenetic networks

A special type of phylogenetic network, also called *split network*, arises when splits are added to a phylogenetic tree that are incompatible with the already present splits [cf. e.g. Semple and Steel, 2003, Dress et al., 2012]. As an example, take a tree from \mathcal{T}_3 with topology defined by the split $s_1 : (0, 1 \mid 2, 3)$. By adding the incompatible split $s_2 : (0, 2 \mid 1, 3)$ we obtain the network shown in Figure 19a. The two incompatible splits are visualized by a two-dimensional cube, which shows that the splits of a phylogenetic network may not be identified with a single edge. In general, a set of k pairwise incompatible splits results in a hypercube of dimension k . Hence, by adding the split $s_3 : (0, 3 \mid 1, 2)$ to the network, we get the network shown in Figure 19b. In this example, not every vertex of the hypercube is incident to a leaf edge.

In the literature, two basic definitions of incompatible split sets can be found.



FIGURE 19. Phylogenetic networks. (A) Two incompatible splits $s_1 = (0, 1 | 2, 3)$ and $s_2 = (0, 2 | 1, 3)$. (B) All three incompatible splits of \mathcal{T}_3 .

DEFINITION 2.7.1 (cf. Bandelt and Dress [1992]). A set of splits Σ is *weakly compatible* if for any three splits $s_1 = (A_1 | B_1)$, $s_2 = (A_2 | B_2)$, and $s_3 = (A_3 | B_3)$ in Σ one of the intersections

$$A_1 \cap A_2 \cap A_3, B_1 \cap B_2 \cap B_3, B_1 \cap A_2 \cap B_3, A_1 \cap B_2 \cap B_3,$$

is empty.

DEFINITION 2.7.2 (cf. Dress et al. [2012]). A set of splits Σ is *k-compatible*, if the largest subset $\sigma \subseteq \Sigma$ of pairwise incompatible splits has cardinality $|\sigma| = k$.

Any 2-compatible set of splits is also *weakly compatible*, however, the two conditions are not equivalent. For our purposes, *k-compatible* split systems are more natural and allow a simple geometric interpretation when lengths are assigned to the splits. For instance, consider the tree space \mathcal{T}_3 which consists of three rays, one for each interior edge, glued together at the origin $\mathbf{0}$. The space can be visualized by mapping the space to the 1-dimensional faces of \mathbb{R}_+^3 (see Figure 20). Phylogenetic networks with 2-compatible split sets are identified as points in one of the 2-dimensional faces of \mathbb{R}_+^3 . Points in the interior of \mathbb{R}_+^3 correspond to 3-compatible networks. It can be easily seen that the space of 2-compatible networks is not CAT(0), see subsection 2.3.2, and the methods for computing means and medians do not apply here.

Key to the formal construction of phylogenetic networks from a set of splits is the identification of interior vertices. For instance, consider the network in Figure 21a with splits

$$\begin{aligned} s_1 &= (A_1 | B_1) = (0, 1 | 2, 3, 4), \\ s_2 &= (A_2 | B_2) = (0, 3, 4 | 1, 2), \\ s_3 &= (A_3 | B_3) = (0, 1, 2 | 3, 4). \end{aligned}$$

Each split $s = (A | B)$ defines two subgraphs obtained by removing the edges associated with s . Equivalently, the subgraph with leaf set A contains

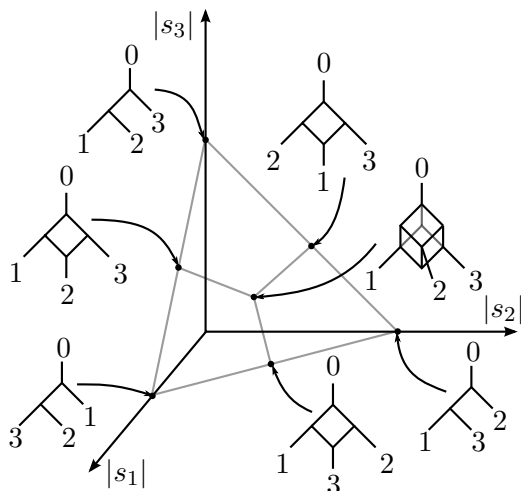


FIGURE 20. Space of phylogenetic networks with four leaves mapped into \mathbb{R}_+^3 . The coordinate axes show the length $|s_i|$, $i = 1, 2, 3$, of each nontrivial split. \mathcal{T}_3 appears as the 1-dimensional faces of \mathbb{R}_+^3 .

all shortest paths between every pair of leaves (a_1, a_2) , where $a_1, a_2 \in A$. An interior vertex may be uniquely identified by the intersection of a collection of subgraphs, one from each split. For instance, the vertex v_1 (see Figure 21b) is defined by the intersection of subgraphs associated with A_1 , A_2 , and A_3 , whereas v_2 is identified by B_1 , A_2 , and A_3 . Clearly, the intersection of subgraphs is non-empty if and only if every pair of leaf sets in the collection has a non-empty intersection.



FIGURE 21. Phylogenetic network with three splits $s_1 = (0, 1 \mid 2, 3, 4)$, $s_2 = (0, 3, 4 \mid 1, 2)$, and $s_3 = (0, 1, 2 \mid 3, 4)$. (A) Subgraphs defined by the splits. (B) Two interior vertices of the network.

Given the set of vertices, it is easy to obtain the edge set. Intuitively, any two vertices are incident to an edge if and only if they are separated by a single split. More formally, let v_1 and v_2 be two vertices of the network.

If for any split $s = (A \mid B)$ either A or B is an element of both v_1 and v_2 , we say that s *agrees* with v_1 and v_2 , otherwise they *disagree*. Vertices that agree with a split appear on the same subgraph induced by the split. If two vertices disagree with only a single split, they are joined by an edge.

DEFINITION 2.7.3 (Buneman graph, cf. Barthelemy [1989], Semple and Steel [2003]). A Buneman graph $G = (V, E)$ with vertices V and edges E is constructed from a set of splits Σ as follows: A vertex $v \in V$ is a collection of sets with pairwise non-empty intersection, such that for every split $s = (A \mid B) \in \Sigma$ either $A \in v$ or $B \in v$. An edge is a pair of vertices that disagree with a single split in Σ .

2.8. Converting tree representations

The computation of geodesics in tree space, as well as the computations of means and medians require the trees to be represented as sets of splits. Other applications, such as the computation of probabilistic models, are much more convenient when a native tree representation is used. Therefore, it is often necessary to convert a tree between representations. While extracting splits from a native tree is simple, the opposite direction is more difficult to implement.

Meacham [1981] proposed the *tree popping algorithm* for converting a set of splits into a native tree [see also Meacham, 1983, Bandelt and Dress, 1986]. The algorithm is initialized with a star-tree where all leaves are directly connected to the root. In each iteration, a split s is picked from Σ . An edge is added to a vertex v of the tree if s further partitions the set of leaves that are directly connected to v . The algorithm stops after iterating once through Σ .

While this algorithm is easy to describe, it is nevertheless somewhat involved to implement. Therefore we use a different approach. First, a leaf a is selected which will be connected to the root node. For all splits $s = A \mid B$ we use the convention that the leaf a is always an element of A , i.e. the first set of the bipartition. A tree with splits as its vertices is constructed using the partial ordering

$$s_1 \succ s_2 \quad \text{if} \quad B_1 \not\subseteq A_2 \wedge B_1 \not\subseteq B_2$$

for any two splits $s_1 = A_1 \mid B_1$ and $s_2 = A_2 \mid B_2$ whenever s_1 is an ancestor of s_2 . This tree of splits can then be easily converted to a phylogenetic tree.

EXAMPLE 2.8.1. Consider the set of splits

$$\begin{aligned} s_1 &= (0,3,4,5 \mid 1,2), & s_2 &= (0,4,5 \mid 1,2,3), \\ s_3 &= (0,1,2,3 \mid 4,5) \end{aligned}$$

where $s_2 \succ s_1$. The tree of splits and the phylogenetic tree are shown in Figure 22.

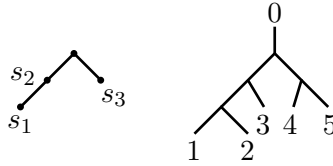


FIGURE 22. Tree of splits and the corresponding phylogenetic tree.

CHAPTER 3

Substitution processes

3.1. Substitution models on phylogenetic trees

Substitution models (sometimes also called mutation models or evolutionary models) are used to describe the process of substitutions in DNA, RNA, and amino acid sequences over time. In their pioneering studies, Zuckerkandl and Pauling [1962, 1965] observed that the genetic distance of ortholog DNA sequences that code for the same protein increases linearly with divergence time [cf. e.g. Salemi and Vandamme, 2003]. This fundamental observation lead to the hypothesis that the number of mutations are governed by a Poisson process. Zuckerkandl and Pauling [1965] interpreted this process as a *molecular clock* progressing at random points in time with a certain rate, and where a tick corresponds to a single substitution event [cf. e.g. Gillespie, 1991]. The molecular clock hypothesis agrees with the *neutral theory of evolution* developed by Kimura et al. [1968] [see also Kimura, 1985]. In fact, we will always assume that mutations are neutral so that the substitution process is *time reversible* and *stationary*. There exists a large variety of different models, which cannot all be covered in this discussion. For a more in depth introduction see for instance Gascuel [2005] or Isaev [2006].

Substitution models rely on a fixed phylogenetic tree and their interpretation requires a given direction of evolution. The trees that were considered so far are unrooted (or can at least be interpreted as such) and therefore have no intrinsic direction of evolution. In the following, we consider trees where leaf 0 and the edge incident to it are removed. The inner vertex that was incident to this edge is called the *root* of the tree, which is interpreted as the *common ancestor* of the remaining vertices (see Figure 1). Such trees will be called *rooted phylogenetic trees*. It should be noted that this is not the only possible definition of a rooted tree. For instance, one may simply select one of the leaf vertices to represent the common ancestor of the remaining species. Once a rooted tree is obtained, we define the causal direction of evolution from the root down to the leaf vertices. In combination with the substitution process, the resulting model can be seen as a Bayesian network. We will see that the *time reversibility* of the substitution models causes the

resulting probability assignments to be independent of the position of the root.

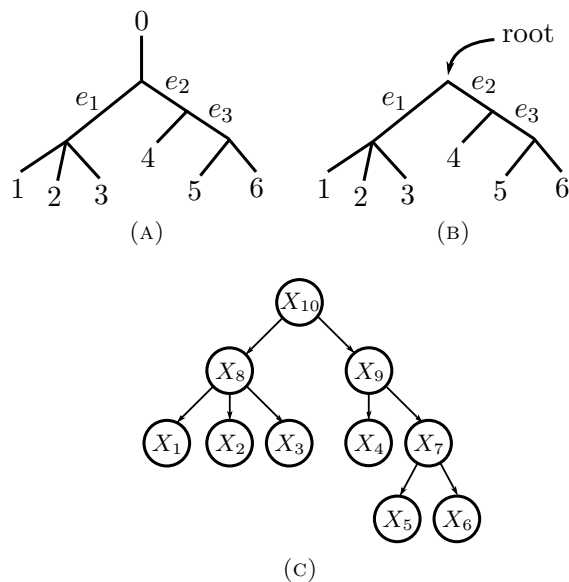


FIGURE 1. A phylogenetic tree with three inner edges (A), from which leaf 0 and the edge incident to it are removed (B). Combined with a mutation model, the phylogenetic tree can be interpreted as a Bayesian network (C), where each vertex i is assigned a random variable X_i .

For the sake of simplicity, we focus on trees with binary branching points and the $n - 1$ internal vertices of the tree (including the root) are labeled from $n + 1$ to $2n - 1$. Non-binary trees can be expressed by allowing edge lengths to be zero. Each of the $2n - 1$ vertices of the tree are associated with random variables $X^{(i)}$, $i = 1, \dots, 2n - 1$ that take values in a finite alphabet \mathbb{A} . The alphabet describes the set of possible observations. For instance, it may represent nucleotides or amino acids. Substitutions occur at the edges of phylogenetic trees and the tree structure is assumed to fully determine the dependency structure between random variables. In particular, we have the conditional independence

$$X^{(i)} \perp\!\!\!\perp X^{(j)} \mid X^{(k)},$$

for any vertex k with children i and j (assuming that the parameters of the substitution process are known and fixed). Hence, the phylogenetic tree in combination with the substitution process is a Bayesian network (see

Figure 1). A good review on Bayesian networks can be found in Wainwright and Jordan [2008].

The substitution process is assumed to be the same for each edge of the tree and it is merely parameterized by the edge lengths. Therefore, we may first have a more general look at the process itself. Let the family of random variables $\{X(\tau) \mid \tau \geq 0\}$ describe the substitution process along an edge of a phylogenetic tree. For instance, if we look at the process between vertices k and i , where vertex i is a child of vertex k , we have $X(0) = X_k$ and $X(\tau^*) = X_i$ for some $\tau^* \geq 0$. The value of τ^* depends on the edge length between vertices k and i and the exact relation will be explained later. The process is assumed to be a *continuous-time finite Markov process*, where the state space is given by the alphabet \mathbb{A} . A stochastic process is usually specified through the joint distributions of every finite family $\{X(\tau_i) \mid \tau_i \in [0, \infty)\}$ of random variables. The *Kolmogorov extension theorem* guarantees the existence of a process under certain conditions on the joint distributions [see e.g. Karlin and Taylor, 1975]. Markov processes are commonly specified in a different form, namely by their conditional distributions. First of all, we will restrict our attention to *stationary processes* (also called *time homogeneous*) that satisfy

$$(X(\tau_1 + \nu), \dots, X(\tau_n + \nu)) \stackrel{d}{=} (X(\tau_1), \dots, X(\tau_n))$$

for all $\nu \geq 0$ and any choice of $0 < \tau_1 < \dots < \tau_n < \infty$. For Markov processes of this kind, the time evolution may be specified through its *transition matrix* $P(\tau) = (p_\tau(x, y))_{x, y \in \mathbb{A}}$, where $\{p_\tau(x, y) \mid \tau \geq 0\}$ is a family of transition functions, such that

$$\text{Pr}_{X(\tau_0 + \tau) \mid X(\tau_0)}(y \mid x) = p_\tau(x, y)$$

for all $\tau_0, \tau \geq 0$ and $x, y \in \mathbb{A}$.

DEFINITION 3.1.1 (Markov transition function). The transition functions $p_\tau : \mathbb{A} \times \mathbb{A} \rightarrow [0, 1]$, $\tau \geq 0$, of a continuous-time Markov process $\{X(\tau)\}$ are differentiable and satisfy

$$\sum_{y \in \mathbb{A}} p_\tau(x, y) = 1, \quad \lim_{\tau \searrow 0} p_\tau(x, x) = p_0(x, x) = 1,$$

as well as the *Chapman-Kolmogorov* equation

$$p_{\nu + \tau}(x, y) = \sum_{z \in \mathbb{A}} p_\nu(x, z) p_\tau(z, y),$$

for all $x, y \in \mathbb{A}$.

Using matrix notation, the Chapman-Kolmogorov equation can also be stated as

$$P(\nu + \tau) = P(\nu)P(\tau),$$

which is also called the *semi-group property*. The Markov process $\{X(\tau)\}$ is uniquely determined by the transition matrix and an initial probability distribution $\phi = (\phi_x)_{x \in \mathbb{A}}$ over the alphabet \mathbb{A} with $X(0) \sim \phi$. In many applications it is preferred to specify the process through its *rates of state transitions* for an infinitesimally small time step. We define

$$q(x, y) = \left. \frac{d}{d\tau} p_\tau(x, y) \right|_{\tau=0}$$

and call $Q = (q(x, y))_{x, y \in \mathbb{A}}$ the *Q-matrix* of the Markov process. This matrix is also referred to as the *instantaneous rate matrix*. It follows that $q(x, y) \geq 0$ for $x \neq y$ and $\sum_{y \in \mathbb{A}} q(x, y) = 0$. By differentiating the Chapman-Kolmogorov equation with respect to ν and setting $\nu = 0$, i.e.

$$\left. \frac{d}{d\nu} p_{\nu+\tau}(x, y) \right|_{\nu=0} = \left. \frac{d}{d\nu} \sum_{z \in \mathbb{A}} p_\nu(x, z) p_\tau(z, y) \right|_{\nu=0},$$

we obtain the equation

$$\frac{d}{d\tau} p_\tau(x, y) = \sum_{z \in \mathbb{A}} q(x, z) p_\tau(z, y),$$

which in matrix notation can be written as

$$\frac{d}{d\tau} P(\tau) = QP(\tau).$$

The last equation is called the *Kolmogorov backward equation* and the solution is given by

$$P(\tau) = \exp(Q\tau) = \mathcal{L}_\nu^{-1} \left[\frac{1}{\nu I - Q} \right] (\tau),$$

where \mathcal{L}_ν^{-1} denotes the inverse Laplace transform and $(\nu I - Q)^{-1}$ is called the *resolvent* of Q . Similarly, if the Chapman-Kolmogorov equation is differentiated with respect to τ , the resulting equation is called the *Kolmogorov forward equation*.

There exists a tight connection between the Markov process $\{X(\tau)\}$ and Poisson processes, which we explore in the following. Let $R = (r(x, y))_{x, y \in \mathbb{A}}$ be a real matrix such that the Q-matrix of the Markov process $\{X(\tau)\}$ takes the form $Q = (R - I)\alpha$, where $\alpha > 0$. Using the solution of the Kolmogorov backward equation, the transition matrix of the Markov process may be written as

$$\begin{aligned} P(\tau) &= e^{-\alpha\tau} e^{R\alpha\tau} = e^{-\alpha\tau} \sum_{m=0}^{\infty} \frac{R^m (\alpha\tau)^m}{m!} = \sum_{m=0}^{\infty} \frac{R^m (\alpha\tau)^m e^{-\alpha\tau}}{m!} \\ &= \sum_{m=0}^{\infty} R^m p_{\alpha\tau}(k), \end{aligned}$$

where $p_{\alpha\tau}(k) = (\alpha\tau)^m e^{-\alpha\tau}/m!$ is the probability function of the Poisson distribution with parameter $\alpha\tau$. The matrix element $r(x, y)$ can be interpreted as the probability of a transition from x to y whenever there is a mutation, so that R^m contains all m -step transition probabilities. Hence, a process $\{Y(\tau)\}$ that counts the number of transitions of $\{X(\tau)\}$ in the time interval $[0, \tau]$ is a Poisson process with rate parameter α .

An important concept of Markov processes is the *equilibrium distribution*, which describes the probability of each state when $\tau \rightarrow \infty$. It is defined as

$$\vartheta = \lim_{\tau \rightarrow \infty} \phi P(\tau),$$

if the limit exists, and where ϕ is some initial probability distribution. The equilibrium distribution is invariant under the application of $P(\tau)$, which leads to the concept of a *stationary distribution* of a Markov process.

DEFINITION 3.1.2 (Stationary distribution). A distribution $\vartheta = (\vartheta_x)_{x \in \mathbb{A}}$ on \mathbb{A} is called *stationary* for a Markov process $\{X(\tau)\}$ if it satisfies

$$(7) \quad \vartheta_y = \sum_{x \in \mathbb{A}} \vartheta_x p_\tau(x, y)$$

for all $\tau \geq 0$ and $y \in \mathbb{A}$.

Obviously, a distribution is stationary, if and only if $\sum_{x \in \mathbb{A}} q(x, y)\vartheta_x$ equals zero for all $y \in \mathbb{A}$. A related concept is the *reversibility* of a Markov process. The Markov process reversed in time has transition matrix $P^*(\tau) = (p_\tau^*(x, y))_{x, y \in \mathbb{A}}$ with

$$p_\tau^*(y, x) = \frac{\phi_x p_\tau(x, y)}{\sum_{z \in \mathbb{A}} \phi_z p_\tau(z, y)}.$$

Assuming that ϕ is the stationary distribution ϑ and that $\vartheta_x > 0$ for all $x \in \mathbb{A}$, then

$$p_\tau^*(y, x) = \frac{\vartheta_x}{\vartheta_y} p_\tau(x, y).$$

By putting $P^* = P$ we obtain the definition of a reversible Markov process.

DEFINITION 3.1.3 (Reversible Markov process). Let $M = \{X(\tau)\}$ be a Markov process with transition function p_τ and stationary distribution ϑ . M is called reversible, if it satisfies the *detailed balance condition*

$$\vartheta_y p_\tau(y, x) = \vartheta_x p_\tau(x, y)$$

for all $\tau \geq 0$ and $y \in \mathbb{A}$.

We have seen so far that the Markov process can be specified through its Q-matrix and in simple cases, the solution of the Kolmogorov backward

equation provides us with a method to compute the corresponding transition probabilities $P(\tau)$. We will now focus on particular choices of substitution processes. Further details on continuous-time Markov processes can be found in Liggett [2010] or Karlin and Taylor [1975].

Among the earliest models is the *Jukes-Cantor model*, which was introduced in Jukes and Cantor [1969]. A generalization of the model was published by Felsenstein [1981].

DEFINITION 3.1.4. The *Felsenstein F81* model is defined through its Q-matrix $Q = (q_{xy})_{x,y \in \mathbb{A}}$ with

$$q_{xy} = \begin{cases} -\alpha(1 - \vartheta_y) & \text{if } x = y, \\ \alpha\vartheta_y & \text{if } x \neq y, \end{cases}$$

for some substitution rate $\alpha > 0$ and a parameter vector $\vartheta = (\vartheta_x)_{x \in \mathbb{A}}$ that satisfies $\vartheta_x \geq 0$ for all $x \in \mathbb{A}$ and $\sum_{x \in \mathbb{A}} \vartheta_x = 1$.

Solving the backward equation shows that the corresponding transition matrix $P(\tau) = (p_\tau(x, y))_{x,y \in \mathbb{A}}$ is given by

$$p_\tau(x, y) = \begin{cases} (1 - \exp(-\alpha\tau))\vartheta_y + \exp(-\alpha\tau) & \text{if } x = y, \\ (1 - \exp(-\alpha\tau))\vartheta_y & \text{if } x \neq y. \end{cases}$$

It is easy to verify that the parameter vector ϑ coincides with the stationary distribution of the model, by computing

$$\begin{aligned} \vartheta_y &= \sum_{x \in \mathbb{A}} p_\tau(x, y)\vartheta_x \\ &= (1 - \exp(-\alpha\tau))\vartheta_y + \exp(-\alpha\tau)\vartheta_y. \end{aligned}$$

It is also easy to show that the process is reversible. The model is not widely used nowadays, since it does not consider different substitution rates for *transitions* and *transversions*. There exist various models that generalize the F81 model. The most commonly used one is the generalized time reversible (GTR) model introduced by Tavaré [1986], although it is not without dispute [cf. e.g. Sumner et al., 2012a,b]. The transition matrix for this model can be computed by diagonalizing the Q-matrix [cf. e.g. Bryant et al., 2005]. However, we do not consider the GTR model, because for our applications we need to analytically integrate over the model parameters. How this can be achieved for the F81 model is explained in section 3.3.

3.2. The likelihood of an alignment

The substitution model is used to describe the process of substitutions in DNA, RNA or amino acid sequences. In the context of a statistical analysis, it serves to define the likelihood, which assigns probabilities to

observations. In our case, those observations are alignments of n ortholog DNA sequences. A multiple alignment has a fixed length L and because of its matrix form, we say that it has n rows and L columns¹. For each $\iota = 1, \dots, L$, the n observations at column ι are assumed to have evolved from the site of a common ancestor according to the given phylogenetic tree and the substitution model. Hence, each column in the alignment is assumed to be independent but not necessarily identically distributed, since the model parameters might be chosen differently for each column.

There are a number of consequences that follow from the choice of the substitution model and we also need to decide on a precise interpretation of the model parameters. For simplicity we consider the Felsenstein F81 model, although most of the following discussion also applies to other models. Consider again a vertex k which has vertex i as a child and within a specific column of the alignment let $X^{(k)}$ and $X^{(i)}$ denote the random variables associated with these vertices. The variable $X^{(i)}$ is conditionally independent of the remaining vertices given the event $\{X^{(k)} = x\}$ at the parent vertex. The conditional distribution of $X^{(i)}$ is defined as a mixture of a categorical distribution and a delta peak at x , i.e.

$$X^{(i)} | X^{(k)} = x \sim M_i \text{Categorical}(\vartheta) + \bar{M}_i \delta_x$$

where M_i denotes the probability of a mutation from vertex k to i . M_i depends on the time τ_i that separates the two vertices as well as the mutation rate α_i , i.e. $M_i = 1 - \exp(-\alpha_i \tau_i)$. Furthermore, \bar{M}_i denotes the probability of no mutation, given as $\bar{M}_i = 1 - M_i$. In its most general form, the mutation rate is defined as

$$\alpha_i = c(\vartheta) \gamma \rho_i,$$

where we call $c(\vartheta)$ the *rate normalization constant*, γ the *rate scaling factor*, and ρ_i the *branch rate factor* for the process between vertices k and i . All three factors will be explained in the following.

If the branch rate factor ρ_i is equal for all branches in the phylogenetic tree, the model is said to have a *strict molecular clock*. This assumption would constrain the phylogenetic tree to a very limited subset of the BHV tree space [cf. e.g. Gavruskin and Drummond, 2014]. It is known that a strict molecular clock is in general too unrealistic [cf. e.g. Gascuel, 2005, p. 83], therefore we will not assume it for our applications. However, since the model depends on the product of time τ_i and the branch rate factor ρ_i , it is not possible to infer both separately. Instead, we will define the branch length l_i between vertices k and i as $l_i = \rho_i \tau_i$, which can be thought of as the

¹In the literature, it is common to refer to a column simply as a site. This terminology seems quite confusing and we will not make use of it, but rather say that a column of the alignment consists of n ortholog sites that evolved from a site of the common ancestor.

expected number of mutations between the two vertices, and infer l_i directly. The branch lengths of the phylogenetic tree therefore do not represent time but the tree can be interpreted as a similarity measure, or simply distance, between ortholog sites.

If one considers genetic coding regions, it is well known that the level of conservation within a codon is highly heterogeneous due to the structure of the genetic code [Li et al., 1985, Yang, 1996]. When learning phylogenetic trees from such regions, it is common to introduce a column specific factor γ that scales the length of every branch in the tree. This parameter significantly hampers the inferential task, which is why it is usually restricted to a discrete set of values, called *rate categories* in the literature [Yang, 1993]. We do not introduce such a parameter in the model, although it would certainly be reasonable to also consider heterogeneous levels of conservation.

Whenever there is a mutation between two vertices, a new nucleotide is generated. In the standard substitution models, there is no mechanism that prevents generating the same nucleotide as already present at the parent vertex. Such spurious mutations are often seen as a mere mathematical convenience and the mutation rate is corrected to represent the rate of actual mutations [e.g. Bryant et al., 2005]. The probability of non-spurious mutations is

$$1 - \sum_{x \in \mathbb{A}} \vartheta_x r(x, x) = - \sum_{x \in \mathbb{A}} \vartheta_x q(x, x) / \alpha_i,$$

hence, we would expect

$$-\tau \sum_{x \in \mathbb{A}} \vartheta_x q(x, x)$$

such events in the time interval $[0, \tau]$. The mutation rate is normalized such that the expected number of non-spurious mutations in a unit time interval equals $\gamma \rho_i$. For the F81 model, we obtain

$$c(\vartheta) = \left[\sum_{x \in \mathbb{A}} \vartheta_x (1 - \vartheta_x) \right]^{-1}.$$

Most models assume that the stationary distribution is the same for all columns in the alignment, although there are a few exceptions [e.g. Lartillot and Philippe, 2004]. In our model, each column of the alignment has its own stationary distribution. We will also not use the normalization constant $c(\vartheta)$ as described above, but simply set it to one, since we want to make explicit use of spurious mutations. The stationary distribution ϑ therefore plays a crucial role. If ϑ has low entropy, i.e. only one character of the alphabet \mathbb{A} has high probability, then this character is repeatedly drawn whenever there is a mutation and the column of the alignment will most likely be highly conserved. On the other hand, a high entropic stationary distribution

causes a column to be poorly conserved with high probability (assuming the distance from the root to the leaves is large enough). In this model, the stationary distribution can be interpreted as the selective pressure that acts on a column of the alignment (see also section 4.3). It was used in this way already by Siddharthan et al. [2005] to analyze motifs of transcription factor binding sites.

The reversibility of the substitution model causes the likelihood to be indiscriminate to the position of the root. Take for instance a tree with two leaves that are represented by the random variables $X^{(1)}$ and $X^{(2)}$ and which are connected to the root $X^{(3)}$. Furthermore, let $\{X^{(1)} = x_1, X^{(2)} = x_2, X^{(3)} = x_3\}$ be the observations. By the reversibility of the Markov process we have

$$\vartheta_{x_3} \text{pr}_{X^{(1)} | X^{(3)}}(x_1 | x_3) \text{pr}_{X^{(2)} | X^{(3)}}(x_2 | x_3) = \vartheta_{x_2} \text{pr}_{X^{(1)} | X^{(3)}}(x_1 | x_3) \text{pr}_{X^{(3)} | X^{(2)}}(x_3 | x_2),$$

where the time at the edge between the root and the second vertex is reversed. Hence, the root can be neglected when computing the likelihood of the observations $\{X^{(1)} = x_1, X^{(2)} = x_2\}$ and we may directly connect the leaves by joining the two branches (see Figure 2). Therefore, when learning a phylogenetic tree from data, we cannot reconstruct the position of the root and inference is hence restricted to unrooted trees.

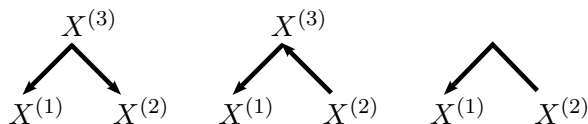


FIGURE 2. Time reversibility causes the likelihood function to be indiscriminate to the position on the root. The tree on the left has two leaves connected to the root of the tree. The arrows of the branches show the direction of time. By reversing time on the right branch, the tree in the middle is obtained. The root can then be removed by joining the two branches, which leads to the tree on the right.

3.3. Computing the likelihood

Let $\bar{X} = \{X^{(1)}, \dots, X^{(n)}\}$ be the set of random variables associated with the leaves of a phylogenetic tree and $X^{(n+1)}, \dots, X^{(2n-1)}$ the random variables assigned to the internal vertices. The values of the internal random variables are usually unobserved. Hence, evaluating the likelihood

$\text{pr}_X(x)$ requires to sum over all possible states of the random variables $X^{(n+1)}, \dots, X^{(2n-1)}$. The tree pruning algorithm proposed by Felsenstein [1981], which is an instance of the sum-product algorithm [Pearl, 1982], allows an efficient numerical evaluation of the likelihood if the parameters of the substitution model are given. However, we consider the stationary distribution as unknown and need to integrate it out. A numerical approximation is computationally too expensive, since in our model each column of the alignment is assigned its own stationary distribution. Therefore, we need to analytically integrate it out, which requires a symbolic representation of the likelihood. Expanding all terms of the likelihood results in a polynomial of the form

$$\text{pr}_{\bar{X}}(\bar{x}) = \sum_i \nu_i \prod_{y \in \mathbb{A}} \vartheta_y^{c_i(y)}$$

with coefficients ν_i and exponents $c_i(y)$, which both depend on the observations $\{\bar{X} = \bar{x}\}$ and the phylogenetic tree. For large phylogenetic trees, the polynomial computed with the tree pruning algorithm has many terms and it is necessary to simplify the polynomial while expanding the likelihood. For demonstration purposes, consider a rooted binary tree with only two leaves and let $\{\bar{X} = \bar{x}\} = \{X^{(1)} = a, X^{(2)} = c\}$. The likelihood is given by

$$\begin{aligned} \text{pr}_{\bar{X}}(\bar{x}) &= \sum_{y \in \mathbb{A}} \vartheta_y \text{pr}_{\bar{X} | X^{(3)}}(\bar{x} | y) \\ &= \sum_{y \in \mathbb{A}} \vartheta_y [M_1 M_2 \vartheta_a \vartheta_c + \delta_{y,a} \bar{M}_1 M_2 \vartheta_c + \delta_{y,c} M_1 \bar{M}_2 \vartheta_a] , \end{aligned}$$

which simplifies to

$$= (1 - \bar{M}_1 \bar{M}_2) \vartheta_a \vartheta_c ,$$

where δ is the Kronecker delta function. An algorithm that computes a symbolic expansion of the likelihood should identify terms that do not depend on the nucleotides at parent vertices and take advantage of the fact that the stationary distribution sums up to one. This means that for the above example the sum

$$\sum_{y \in \mathbb{A}} \vartheta_y M_1 M_2 \vartheta_a \vartheta_c$$

should not be fully expanded to

$$\vartheta_a M_1 M_2 \vartheta_a \vartheta_c + \vartheta_c M_1 M_2 \vartheta_a \vartheta_c + \vartheta_g M_1 M_2 \vartheta_a \vartheta_c + \vartheta_t M_1 M_2 \vartheta_a \vartheta_c ,$$

but rather be evaluated as $M_1 M_2 \vartheta_a \vartheta_c$. As it turns out, it is possible to compute such a simplified polynomial for an arbitrary tree in a simple recursive way.

3.4. Likelihood decomposition

The objective of the following discussion is to develop a recursive method that allows to compute the polynomial of the likelihood. The common definition of a phylogenetic tree in terms of splits is not well suited for the description of such a method. A more convenient description should consider the recursive structure of rooted phylogenetic trees. For instance, consider a tree t with internal vertex k that has two children i and k . We may refer to the subtree below vertex k as t_k . In the same way, the subtrees below vertices i and j may be denoted t_i and t_j respectively. To express that vertex k is the parent of i and j we simply write $t_k = (t_i, t_j)$. Note that t_i and t_j may refer to leaves or subtrees.

For the discussion, it is instructive to first look at the likelihood of star-trees with an arbitrary number of leaves. We will then show that the likelihood of trees of any topology can be expressed in terms of simple star-trees, which was also observed by Siddharthan et al. [2005].

3.4.1. Star-trees. To analyze the likelihood we first consider the case of star-trees, i.e. trees with no internal edges and an arbitrary number of leaves that are directly connected to the root vertex. Consider the tree $t_3 = (t_1, t_2)$, where t_1 and t_2 are leaves. The likelihood for an observation $\{X^{(1)} = x_1, X^{(2)} = x_2\}$ is given by

$$\text{pr}_{X^{(1)}, X^{(2)}}(x_1, x_2) = (1 - \bar{M}_1 \bar{M}_2) \vartheta_{x_1} \vartheta_{x_2} + \bar{M}_1 \bar{M}_2 \vartheta_{x_1} \delta_{x_1, x_2},$$

where δ is the Kronecker delta function. This example shows two important aspects. First, $X^{(3)}$ is marginalized out and does not appear in the final likelihood. Second, the likelihood consists of two cases. Either there is a mutation at any of the two edges or there is no mutation at both. The first case shows that x_1 and x_2 are generated independently. For instance, x_1 could be generated at the root which is then inherited by leaf 1, in which case there is a mutation between the root and leaf 2 that generates x_2 . In the second case the nucleotide is generated at the root and then inherited by both leaves. Of course, this case occurs only if the same nucleotide is observed at both leaves.

Let $t_{n+1} = (t_1, \dots, t_n)$ be a star-tree with n leaves and observations $\{\bar{X} = \bar{x}\} = \{X^{(1)} = x_1, \dots, X^{(n)} = x_n\}$. To compute the likelihood

$$\text{pr}_{\bar{X}}(\bar{x}) = \sum_{y \in \mathcal{A}} \vartheta_y \prod_{i=1}^n \text{pr}_{X^{(i)} | X^{(n+1)}}(x_i | y),$$

we introduce the following recursive

DEFINITION 3.4.1. Consider a star-tree $t_{n+1} = (t_1, \dots, t_n)$ with observations $\{\bar{X} = \bar{x}\} = \{X^{(1)} = x_1, \dots, X^{(n)} = x_n\}$. For a positive integer $i \leq n$

let $t_{1:i}$ denote the reduced tree with i leaves. Define

$$\varphi(t_{1:i}) = \sigma(t_{1:i}) + \sum_{y \in \mathcal{A}} \vartheta_y \varphi_y(t_{1:i})$$

with

$$\begin{aligned} \sigma(t_{1:i}) &= \prod_{j=1}^i M_j \vartheta_{x_j} \\ \varphi_y(t_{1:i}) &= \delta_{x_i, y} \bar{M}_i (\sigma(t_{1:i-1}) + \varphi_y(t_{1:i-1})) + M_i \vartheta_{x_i} \varphi_y(t_{1:i-1}) , \end{aligned}$$

where

$$\varphi_y(t_{1:1}) = \begin{cases} \bar{M}_1 & \text{if } x_1 = y , \\ 0 & \text{otherwise .} \end{cases}$$

To understand the rationale behind this definition consider the following more intuitive description:

- $\sigma(t_{1:i})$: The tree $t_{1:i}$ is independent of the observation at the root and therefore all observations at the leaves need to be generated.
- $\varphi_y(t_{1:i})$: The tree $t_{1:i}$ depends on the observation y at the root, but there are two possible cases depending on whether or not there is a mutation at the edge between vertex i and the root. If there is a mutation, x_i is generated and the remaining tree $t_{1:i-1}$ has to depend on y . However, φ_y vanishes if y is not observed in the remaining tree. If there is no mutation, which is possible only if $x_i = y$, then leaf i inherits the observation y from the root. However, the remaining tree $t_{1:i-1}$ might be fully independent of the root or it could depend on y as well.

The conditional likelihood can be expressed as

$$\begin{aligned} \text{pr}_{\bar{X}} | X^{(n+1)}(\bar{x} | y) &= \sigma(t_{n+1}) + \varphi_y(t_{n+1}) \\ &= (M_i \vartheta_{x_i} + \delta_{x_i, y} \bar{M}_i) (\sigma(t_{1:i-1}) + \varphi_y(t_{1:i-1})) , \end{aligned}$$

and therefore the likelihood becomes

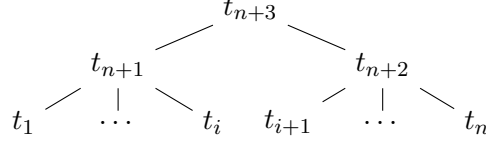
$$\text{pr}_{\bar{X}}(\bar{x}) = \varphi(t_{n+1}) = \sigma(t_{n+1}) + \sum_{y \in \mathcal{A}} \vartheta_y \varphi_y(t_{n+1}) ,$$

which allows to compute a symbolic representation in a recursive way with a minimal number of terms. Here we have exploited the simple fact that σ does not depend on the nucleotide at the root of the tree and therefore

$$\sum_{y \in \mathcal{A}} \vartheta_y \sigma(t_{n+1}) = \sigma(t_{n+1}) ,$$

and $\varphi_y(t_{1:i})$ vanishes if no nucleotide y is observed in the tree $t_{1:i}$.

3.4.2. Tree reduction. To extend the given results to trees of arbitrary topology consider first the example of two star-trees connected by a common ancestor, i.e.



That is, let $t_{n+1} = (t_1, \dots, t_i)$ and $t_{n+2} = (t_{i+1}, \dots, t_n)$ be star-trees with observations $\{\bar{X} = \bar{x}\} = \{X^{(1:i)} = x_{1:i}, X^{(i+1:n)} = x_{i+1:n}\}$ and let $t_{n+3} = (t_{n+1}, t_{n+2})$. The likelihood of the left subtree, i.e. of observations $\{X^{(1:i)} = x_{1:i}\}$, conditional on the root vertex is given by

$$\begin{aligned} \text{pr}_{X^{(1:i)} | X^{(n+3)}}(x_{1:i} | x) &= \sum_{y \in \mathcal{A}} \text{pr}_{X^{(n+1)} | X^{(n+3)}}(y | x) [\sigma(t_{1:i}) + \varphi_y(t_{1:i})] \\ &= M_{n+1} \varphi(t_{1:i}) + \bar{M}_{n+1} [\sigma(t_{1:i}) + \varphi_x(t_{1:i})] , \end{aligned}$$

where $X^{(n+1)}$ is marginalized out. Also in this case, the expression is independent of the observation $\{X^{(n+3)} = x\}$ at the root vertex if a mutation occurs at the edge between t_{n+1} and t_{n+3} . If there is no mutation, the term partly depends on $\{X^{(n+3)} = x\}$. Given this result, it is easy to obtain the likelihood of the full set of observations

$$p_{\bar{X}}(\bar{x}) = [1 - \bar{M}_{n+1} \bar{M}_{n+2}] \varphi(t_{1:i}) \varphi(t_{i+1:n}) + \bar{M}_{n+1} \bar{M}_{n+2} \varphi(t_{1:i} \cup t_{i+1:n}) ,$$

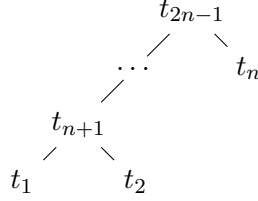
where

$$\varphi(t_{1:i} \cup t_{i+1:n}) = \sum_y \vartheta_y [\sigma(t_{1:i}) + \varphi_y(t_{1:i})] [\sigma(t_{i+1:n}) + \varphi_y(t_{i+1:n})]$$

can be interpreted as the likelihood of a star-tree that joins all n leaves. It is easy to see that in this fashion the likelihood of trees of any topology can be expressed in terms of star-trees, which was also observed by Siddharthan et al. [2005]. The likelihood takes the form

$$\text{pr}_{\bar{X}}(\bar{x}) = \sum_{\pi \in \Pi} \nu_{\pi} \prod_{s \in \pi} \varphi(t_s) ,$$

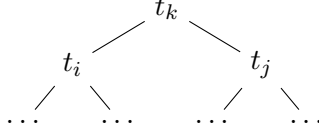
for some coefficients $\nu_{\pi} \in \mathbb{R}$ and a set of partitions of the leaves Π . However, the cardinality of Π grows too quickly with the size of the tree, which makes this approach inappropriate for larger trees. For instance, if we assume only trees of the form



with n leaves, then $|\Pi| = 2^{n-1}$, although trees of this topology constitute the worst-case scenario. It may therefore be worthwhile to expand the likelihood directly when traversing the tree.

3.4.3. General likelihood decomposition. There exists a likelihood decomposition for general binary trees, as observed by Bourguignon [2012]. We will now formalize this observation and prove the result by extending Definition 3.4.1. The form of the decomposition can be obtained from our previous results by computing the conditional likelihood and sorting the terms by their dependence on parent vertices, which leads to

DEFINITION 3.4.2. For a binary tree $t_k = (t_i, t_j)$ of the form



let

$$\varphi(t_k) = \sigma(t_k) + \sum_{y \in \mathbb{A}} \vartheta_y \varphi_y(t_k) ,$$

with

$$\begin{aligned} \varphi_y(t_k) &= \bar{M}_i \varphi_y(t_i) (\bar{M}_j \sigma(t_j) + M_j \varphi(t_j)) + \bar{M}_j \varphi_y(t_j) (\bar{M}_i \sigma(t_i) + M_i \varphi(t_i)) \\ &\quad + \bar{M}_i \bar{M}_j \varphi_y(t_i) \varphi_y(t_j) , \\ \sigma(t_k) &= (\bar{M}_i \sigma(t_i) + M_i \varphi(t_i)) (\bar{M}_j \sigma(t_j) + M_j \varphi(t_j)) . \end{aligned}$$

If for instance vertex i is a leaf, then $\sigma(t_i) = 0$ and $\varphi_y(t_i) = \delta_{x_i, y}$.

LEMMA 3.4.3. *The conditional likelihood is given by*

$$\text{pr}_{\bar{X} | X^{(k)}}(\bar{x} | y) = \sigma(t_k) + \varphi_y(t_k) ,$$

so that

$$\text{pr}_{\bar{X}}(\bar{x}) = \varphi(t_k) = \sigma(t_k) + \sum_{y \in \mathbb{A}} \vartheta_y \varphi_y(t_k) .$$

PROOF. We have already shown that the likelihood of a star-tree with two leaves is decomposable in the above sense. For the general case, assume

the above tree and let $\{X^{(i)} = x_i\}$ denote the observations of subtree t_i and $\{X^{(j)} = x_j\}$ the observations of t_j . From Lemma 3.4.3 we know that

$$\begin{aligned}\mathrm{pr}_{X^{(i)}|X^{(i)}}(x_i|y) &= \sigma(t_i) + \varphi_y(t_i), \\ \mathrm{pr}_{X^{(j)}|X^{(j)}}(x_j|y) &= \sigma(t_j) + \varphi_y(t_j).\end{aligned}$$

By computing

$$\mathrm{pr}_{\bar{X}|X^{(k)}}(\bar{x}|y) = \mathrm{pr}_{X^{(i)}|X^{(k)}}(x_i|y)\mathrm{pr}_{X^{(j)}|X^{(k)}}(x_j|y)$$

we obtain the equations of Definition 3.4.2 which proves the result. \square

The last lemma shows how to efficiently obtain a symbolic representation of the likelihood. The same intuitive interpretation of the functions σ and φ_y also applies to the case of general phylogenetic trees.

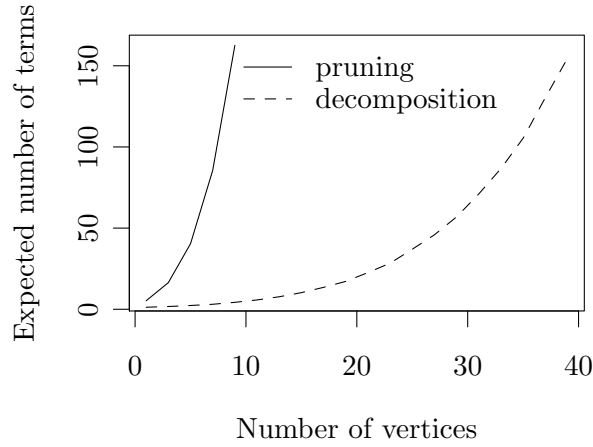
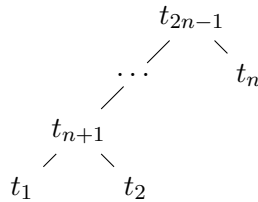


FIGURE 3. Comparison of the number of terms generated by Felsenstein’s pruning algorithm and the likelihood decomposition algorithm.

To test the performance of the algorithm we compare it to Felsenstein’s pruning algorithm by considering again trees of the form



with $2n - 1$ vertices. The number of terms in the resulting polynomial depends not only on the tree structure, but also on the observations at the leaves. Therefore, we approximate the expected number of terms by randomly drawing observations from a uniform distribution. The results are

depicted in Figure 3, which shows that both algorithms generate polynomials with an exponential growth in the number of terms. However, for the decomposition algorithm the growth rate is much slower and our applications do not require trees of more than 40 vertices.

3.5. Detecting conserved sites

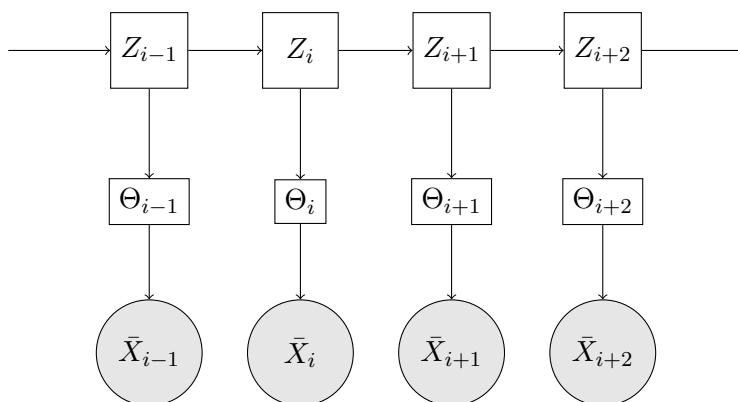


FIGURE 4. Illustration of an HMM for detecting conserved sites. The illustration shows a segment of four states. Observations are depicted as \bar{X}_i and hidden states as Z_i . The Θ_i are the direct emissions of the HMM.

In this section, we show a simple application of the alignment model that we introduced earlier. Siepel and Haussler [2004, 2005] developed a method for the detection of conserved sites, where a hidden variable Z_i at position i of the alignment indicates whether the column is conserved or not. Depending on the state Z_i , a respective rate scaling factor is used to model the alignment column (see section 3.2). In our model, we did not introduce a rate scaling factor, but instead control the level of conservation with a column specific stationary distribution. If a Dirichlet prior is used for the stationary distribution, different levels of conservation can be modeled by either putting mass on the boundary or the center of the probability simplex.

We assume a discrete-time hidden Markov model (HMM) with states Z_i for $i = 1, \dots, n$ that take values in some discrete space Z (see Figure 4). The emissions \bar{X}_i are the columns of the alignment. We obtain the marginal likelihood

$$\text{pr}_{\bar{X}_i | Z_i}(\bar{x}_i | z_i) = \int_{\Theta_i} \text{pr}_{\bar{X}_i | \Theta_i}(\bar{x}_i | \vartheta) f_{\Theta_i | Z_i}(\vartheta | z_i) d\vartheta$$

where $f_{\Theta_i | Z_i}(\vartheta | z_i)$ is the Dirichlet density with pseudocounts depending on $\{Z_i = z_i\}$. The prior for the HMM is given by

$$\text{pr}_{Z_{0:n}}(\mathbf{z}_{0:n}) = \text{pr}_{Z_0}(z_0) \prod_{i=1}^n K(z_{i-1}, z_i),$$

where pr_{Z_0} is the distribution of the initial state Z_0 and

$$K(z_{i-1}, z_i) = \text{pr}_{Z_i | Z_{i-1}}(z_i | z_{i-1})$$

the transition probability function, which controls the inertia of the process. We are interested in obtaining the marginal posterior distribution $\text{pr}_{Z_i | \bar{X}_{1:n}}$ for each hidden state Z_i , which can be computed in a forward-backward manner [Cappé et al., 2005]. In the following, we drop the subscripts of the distribution functions for better readability. The forward computation corresponds to a recursive Bayesian estimate where we compute

$$\text{pr}(z_i | \bar{x}_{1:i}) = \frac{\text{pr}(\bar{x}_i | z_i) \text{pr}(z_i | \bar{x}_{1:i-1})}{\text{pr}(x_i | \bar{x}_{1:i-1})},$$

with

$$\text{pr}(z_i | \bar{x}_{1:i-1}) \propto \sum_{z_{i-1} \in \mathcal{Z}} K(z_{i-1}, z_i) \text{pr}(\bar{x}_{i-1} | z_{i-1}) \text{pr}(z_{i-1} | \bar{x}_{1:i-2}),$$

and

$$\text{pr}(\mathbf{z}_{0:i} | \bar{x}_{1:i}) = \text{pr}(\mathbf{z}_{0:i-1} | \bar{x}_{1:i-1}) \frac{\text{pr}(\bar{x}_i | z_i) K(z_{i-1}, z_i)}{\text{pr}(\bar{x}_i | \bar{x}_{1:i-1})}.$$

The backward probabilities are computed according to

$$\text{pr}(\bar{x}_{i:n} | z_i) = \text{pr}(x_i | z_i) \sum_{z_{i+1} \in \mathcal{Z}} K(z_i, z_{i+1}) \text{pr}(\bar{x}_{i+1:n} | z_{i+1}),$$

so that we obtain the marginal posterior probabilities

$$\text{pr}(z_i | \bar{x}_{1:n}) \propto \text{pr}(z_i | \bar{x}_{1:i}) \text{pr}(\bar{x}_{i+1:n} | z_i).$$

A simple demonstration is shown in Figure 5. The hidden Markov chain has two states, one for conserved sites with Dirichlet pseudocounts set to 0.1 and one for unconserved sites with pseudocounts set to 10 for all characters in the alphabet. The transition kernel is defined as

$$K(x, y) = \begin{cases} 0.95 & \text{if } x = y, \\ 0.05 & \text{otherwise,} \end{cases}$$

which controls the inertia (or smoothness) of the resulting estimate. For applications it is important that the prior parameters are chosen carefully to match *your* a priori expectations. Since the level of conservation not only depends on the Dirichlet pseudocounts, but also on the phylogenetic tree, it is crucial to have a properly estimated tree for this application. The estimation of trees is discussed in detail in chapter 4.

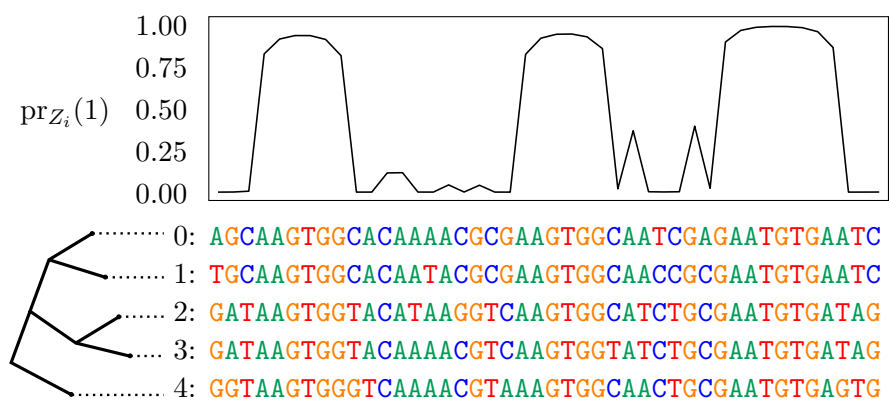


FIGURE 5. HMM for detecting conserved regions in multiple sequence alignments. The upper plot shows the marginal probability $\text{pr}_{Z_i}(1)$ for the Markov chain to be in state one, which represents conserved sites.

CHAPTER 4

Point estimates in phylogenetic reconstructions

Given a generative model and a prior distribution over its parameter space, a Bayesian analysis of observations carried across species related by evolution produces a posterior distribution over the space of all possible phylogenetic trees for this set of species [Gascuel, 2005, Robert and Casella, 1999]. The size of this space grows super-exponentially with the number of species, and it is often intractable to compute the normalization constant of this distribution. In such cases, sampling methods offer a way to explore the posterior distribution via an arbitrarily large sample drawn from it without requiring any further knowledge. However, although a posterior sample offers a representation of the full posterior distribution, it is of little scientific interest in absence of a method to summarize it, and especially to derive a point estimate of the (random) posterior tree. Building upon previous works by Billera et al. [2001], Owen and Provan [2011], Miller et al. [2012] and Bačák [2014a], we propose here to define and compute posterior means in a sound manner, an approach so far hindered by the poor geometrical insights into the space of phylogenetic trees [Benner et al., 2014, Holmes, 2005].

Provided a unique topology with n edges occurs in the sample, each tree including its edge lengths can be identified by a point in the positive orthant of the Euclidean space \mathbb{R}^n . Computing an mean of the sample in this linear representation is a straightforward operation, which produces a legit posterior mean tree. If more than one tree topology occurs, the trees are no longer mapped all to the same linear space, and the posterior mean is ill-defined. Selecting the a posteriori most probable tree topology may seem a sound alternative, however, with the unpleasant consequence of neglecting all the sampled trees of different topology, and therefore would not provide a satisfactory representation of the posterior. The construction of a consensus tree, using an absolute majority-rule¹ [Margush and McMorris, 1981] to decide which one among competing edges should be retained, has been widely adopted by the interested community as the method of choice to summarize

¹The majority-rule consensus tree is constructed by including only those edges which appear in more than 50% of the samples.

posterior samples of phylogenetic trees. On the theoretical side, decision-theoretic justifications of this construction have been proposed [Holder et al., 2003, Huggins et al., 2011]. However they are built upon loss functions that neglect edge lengths, focusing only on the tree topology. Besides, from the authors' point of view, it is also a rather conservative approach, as the absence of an absolute majority among edges results in the inclusion of none of them, thereby producing unresolved branching points. The extended majority-rule consensus method (also known as greedy consensus method) has been introduced to remedy this drawback by adding edges with less than 50% support [Bryant, 2003]. Despite this improvement, the consensus methods neglect much of the available information in a sample by ignoring the context in which an edge occurs (i.e. the remaining topology of the tree as well as all other edge lengths). Reporting a posterior mean that balances the contributions from each topology including edge lengths rather than isolated edges would therefore be of utmost interest.

Let us now exploit the geometric properties of the BHV tree space, as described in chapter 2, to summarize a sample of phylogenetic trees by a single point. Following the rationale of decision theory [e.g. Robert, 2001], the construction begins with the definition of a *loss function*, which measures how faithful a representation of the sample would be achieved by a given point in the tree space (see also section A.8). A loss function is defined as the cost $\mathcal{L}(t, t')$ of choosing a phylogenetic tree t' instead of some other t , and the decision theory literature advocates strongly to summarize a posterior distribution by choosing \hat{t} as the minimizer of the expected loss function

$$(8) \quad \hat{t} = \arg \min_{t' \in \mathcal{T}_n} \int_{\mathcal{T}_n} \mathcal{L}(t, t') d\mu(t),$$

where μ denotes the posterior distribution over phylogenetic trees given the data X [cf. Holmes, 2005]. Approximating the latter via a posterior sample of phylogenetic trees t_1, \dots, t_K , the above formula becomes

$$\hat{t} = \arg \min_{t' \in \mathcal{T}_n} \frac{1}{K} \sum_{i=1}^K \mathcal{L}(t_i, t').$$

Two very typical choices for the loss function are the distance and the squared distance. When the parameters to be estimated lie in a Euclidean space, it is well-known that the resulting estimates coincide respectively with the median and mean of the posterior distribution. Although the tree space is not Euclidean, distances between pairs of trees are well-defined, and a minimizer of Equation 8 can be sought, respectively yielding the so-called *geometric median* and *Fréchet mean* (see also section A.3).

In contrast to other approaches that provide a decision-theoretic argument for point estimates of phylogenetic trees [e.g. Holder et al., 2003, Huggins et al., 2011], the loss function considered here derives the *intrinsic metric* of the underlying space. In particular, the loss function considers both the topology and the branch lengths of phylogenetic trees, as opposed to those supporting the consensus method, and thereby considers all available information in a sample. Unfortunately, in tree space, a simple gradient search is not a practical method to solve such optimization problems [see Miller et al., 2012]. See section 2.5 for a discussion of the algorithms.

A side benefit of the method presented here is the sound definition of the sample variance, also called the Fréchet variance, which is simply given as the value of the minimization problem with the squared distance. In complement to the point estimate, this quantity provides the modeler with insight onto the reliability of the point estimate. It is noteworthy that existing phylogenetic reconstruction methods are not tied to a notion of variance, and often retort to bootstrapping methods for reporting a similar information.

4.1. Consensus vs. posterior mean

The majority-rule consensus method is a reference method to summarize samples from a posterior distribution. There, the consensus tree consists of those splits that occur in more than 50% of the samples. The mean length of a retained edge is computed using the subsample where the corresponding split does occur, thereby neglecting a fraction of the posterior mass, but also the context in which the split occurs. In contrast, the edge lengths of the Fréchet mean and geometric median account for the full posterior, and are expected to provide a more meaningful summary. However, both estimates have a property called stickiness [see e.g. Miller et al., 2012]: if there is high posterior uncertainty about the topology, this property will cause the Fréchet mean and geometric median to result in non-binary trees, a behavior that parallels the multiple branching points reconstructed by the consensus tree when no absolute majority occurs.

Take for instance the space \mathcal{T}_3 that consists of three orthants $[0, \infty)$ glued together at $\mathbf{0}$ and place a tree on each orthant. If all three trees are equally far apart, say at a distance r to the origin, then obviously the Fréchet mean lies at the origin. The term stickiness refers to the fact that the mean stays at the origin if one of the trees is moved further away. In fact, one tree may be located at a distance anywhere between r and $2r$ away from the origin without affecting the mean. Instead of moving one tree further away from the origin, one may similarly add another tree somewhere between the three trees, and the Fréchet mean would again stay at the origin.

A probabilistic counterpart of this phenomenon can be observed in the same setting. Equip \mathcal{T}_3 with a distribution whose trace in each orthant is a normalized Gaussian distribution, centered at identical distance from the origin, and truncated at $\mathbf{0}$. By symmetry the Fréchet mean is at the origin, and one can ask how far the location parameter m of one component can be perturbed without affecting the mean. In \mathcal{T}_3 , m is just a scalar, and one can study the distance of the Fréchet mean \hat{t} to the origin $\mathbf{0}$ as a function of m . An analytic but complicated solution of the distance $d(\mathbf{0}, \hat{t})$ exists, however, a fairly good answer is provided by the following approximation:

$$d(\mathbf{0}, \hat{t}) \approx \max \left\{ 0, \frac{m - \sqrt{2/(e\pi)} - 2\Phi(1)}{1 + 2\Phi(1)} \right\}, \quad m \geq 1$$

where Φ is the standard normal cumulative distribution function. The Fréchet mean stays at the origin until m reaches approximately 2.16, which approximately matches the case of only three trees. Also in this case one may similarly increase the probability mass on one orthant and the Fréchet mean would stay at the origin until a certain threshold is reached.

4.2. Star-tree problem

Suzuki et al. [2002] observed that posterior edge probabilities in Bayesian phylogenetic analysis might be excessively high. In this context, an interesting problem was discussed by Lewis et al. [2005], Yang and Rannala [2005], Yang [2007] who called it the star-tree problem. We do not attempt to provide a solution to the problem, but discuss the issue in light of the Fréchet mean and geometric median in tree space. The problem is often outlined in analogy to the following coin tossing example.

Consider a coin which is tossed n times and we are interested in whether the coin is biased towards heads or tails. The usual Bayesian model is given by

$$\begin{aligned} \Theta &\sim \text{Beta}(\alpha_1, \alpha_2) \\ X | \Theta = \theta &\sim \text{Binomial}(\theta, n) \end{aligned}$$

where θ is the rate parameter of observing heads and (α_1, α_2) the pseudocounts of the Beta distribution. Given the observation $\{X = x\}$ the posterior is again a beta distribution with pseudocounts $(x + \alpha_1, n - x + \alpha_2)$ and it can be shown that the posterior expectation converges to $1/2$ as $n \rightarrow \infty$ if the coin is fair. Estimating the parameter Θ can be seen to be similar to the inference of phylogenetic trees. Say that H_1 is the hypothesis that the coin

is biased towards heads. Its probability is given by

$$\text{pr}(H_1 | x) = \int_0^{1/2} f_{\Theta|X}(\theta | x) d\theta,$$

where $f_{\Theta|X}$ denotes the density function of $\Theta | X$ (the notation is introduced in section A.1). For a fixed n one may look at the distribution of $Y \sim \text{pr}(H_1 | x)$ by repeatedly drawing observations $\{X = x\}$ from a Binomial($1/2, n$). For $\alpha_1 = \alpha_2$ it can be shown that in the limit $n \rightarrow \infty$ the distribution of Y is uniform, which lead to the conclusion that the posterior probability is not decisive in this case. The uniform distribution of Y is explained by the small variance of the posterior distribution for large n , which causes a strong peak typically either below or above $1/2$. However, at the same time, the model is consistent since the mode of the posterior distribution moves closer to $1/2$ as n increases (see section A.8 for a definition of consistency).

Edge probabilities are similar to the probability of the coin being head or tail biased. Lewis et al. [2005] observed that if a large amount of data is generated using a star-tree, then the posterior probability might peak at one orthant of tree space, which leads to a high edge probability, although the edge is not present in the generating tree. This observation lead Lewis et al. [2005] to develop a model that puts additional mass on non-binary topologies. The tree space however has an important difference to the parameter space of the coin tossing example, namely that the boundary of an orthant is common to multiple other orthants. This fact causes the stickiness of the mean and median as discussed in section 4.1. As a result, an edge is only present in the estimated tree if there is sufficient posterior evidence, which might be interpreted as an inherent *Occam's razor* (which in a similar way holds for the majority-rule consensus tree). This shows that, at least for moderate data set sizes, additional prior mass on non-binary trees might not be required.

4.3. Statistical model

The statistical model used here consists of the alignment model introduced in section 3.2 and a prior on phylogenetic trees as well as the stationary distributions assigned to each column. The model is simple enough so that the marginal likelihood of an alignment given a phylogenetic tree can be computed analytically. In particular, it permits Bayesian model selection, or to gain some insights on how well an estimate generalizes to new data, and therefore qualifies perfectly for the purpose of comparing different downstream methods for summarizing the posterior samples.

Before discussing the details of the model, we should take a brief look at the data for which the model is intended. Several recent studies have shown

that phylogenetic trees learned on single genes (gene trees) show topological inconsistencies if compared to the corresponding species tree [Hess and Goldman, 2011, Song et al., 2012, Salichos and Rokas, 2013, Zhong et al., 2013]. Possible causes may include wrong assumptions in the substitution model [Kumar et al., 2011], or biological factors such as horizontal gene transfer, hybridization, misalignment, or incomplete lineage sorting [cf. e.g. Avise et al., 1983, Slowinski and Page, 1999, Sanderson and Shaffer, 2002, Maddison and Knowles, 2006]. It is common practice to pool data from many genes in order to fully resolve the topology of the tree. However, the proper model for this data would then be a mixture of trees. As shown by Mossel and Vigoda [2005], the resulting estimate will be highly misleading if only a single tree is assumed by the model. We will therefore restrict ourselves to the inference of single gene trees. The high posterior uncertainty due to the small data set will be addressed by computing a proper posterior estimate.²

We proceed by discussing the prior for the stationary distributions first, which is specific to each column of the alignment. The distribution will be integrated out in the full model, since we are only interested in the inference of phylogenetic trees. We introduce a random variable Θ that represents the stationary distribution and obtain the conditional probability

$$\text{pr}_{X|\Theta}(x|\boldsymbol{\vartheta}) = \vartheta_x = \vartheta(x),$$

of generating nucleotide $x \in \mathbb{A}$. We assume that Θ is a priori Dirichlet distributed with pseudocounts $\boldsymbol{\alpha} = (\alpha_x)_{x \in \mathbb{A}}$. The probability of observing $\{\bar{X} = \bar{x}\}$ becomes

$$\text{pr}_{\bar{X}}(\bar{x}) = \int_{\Delta} \text{pr}_{\bar{X}|\Theta}(\bar{x}|\boldsymbol{\vartheta}) \text{dpr}_{\Theta}(\boldsymbol{\vartheta}),$$

where the integral is defined on the $(|\mathbb{A}| - 1)$ -dimensional probability simplex Δ and can be solved analytically by first expanding the polynomial of the distribution $\text{pr}_{\mathbf{X}|\Theta}$.

It is important to select an appropriate set of parameters $\boldsymbol{\alpha}$ for the Dirichlet distribution, as they control the expected entropy

$$\mathbb{E} h(\Theta) = \int_{\Delta} h(\boldsymbol{\vartheta}) \text{dpr}_{\Theta}(\boldsymbol{\vartheta}) = \psi \left(1 + \sum_{x \in \mathbb{A}} \alpha_x \right) - \sum_{x \in \mathbb{A}} \frac{\alpha_x}{\sum_{y \in \mathbb{A}} \alpha_y} \psi(1 + \alpha_x)$$

of $\{\Theta = \boldsymbol{\vartheta}\}$, where $h(\boldsymbol{\vartheta}) = -\sum_{x \in \mathbb{A}} \vartheta_x \log \vartheta_x$ and ψ denotes the digamma function (see Figure 1) [cf. e.g. Archer et al., 2013, Nemenman et al., 2001,

²In fact, a species tree could be estimated as the mean or median of a set of gene trees.

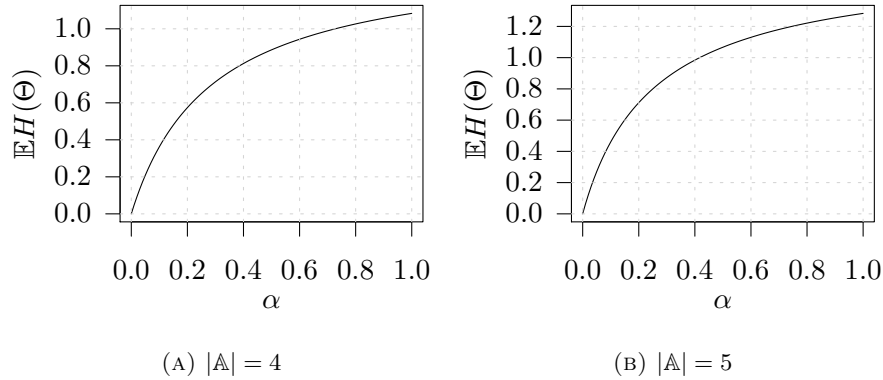


FIGURE 1. Expected entropy of the stationary distribution controlled by $\alpha = \sum_{x \in \mathbb{A}} \alpha_x$ for alphabets \mathbb{A} of cardinality four (A) and five (B).

Wolpert and Wolf, 1995]. Phylogenetic trees are commonly learned on multiple sequence alignments of genes. Such genomic regions are highly conserved, which means that selective pressure causes nucleotides in a column of the alignment to be the same with high probability. To reflect this knowledge in our prior assumption, it is important that the expected entropy is low, i.e. that only the probability of one or two nucleotides is high. This can be achieved by choosing $\alpha_x < 1$, which puts mass close to the faces of the probability simplex. The choice of α has a strong influence on inferred edge lengths. If we increase α , we observe that inferred branch lengths shorten to compensate for the increase in entropy of the stationary distribution. The choice of pseudocounts α therefore reflects our *a priori* assumption of how conserved we expect a genomic region to be. It is well known that within codons a heterogeneous selective pressure exists [Li et al., 1985, Yang, 1996], which can be modeled by introducing specific pseudocounts. For instance, consider the alignment of two codons in Figure 2. The first one shows strong variation at the third codon position, hence, the stationary distribution at this column has high entropy. The reason is that all of the observed codons ACA, ACC, and ACT code for the same amino acid (serine), which leads to a low selective pressure at that position. The second codon alignment in Figure 2 shows a fully conserved third position. In this case, we would still *a priori* assume a stationary distribution with high entropy, as depicted in the figure. Yet, there were probably not enough mutations so that we observe a poorly conserved column.

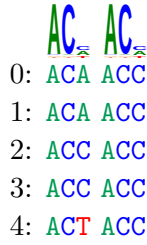


FIGURE 2. Alignments of two codons with low and high conservation. At each position, a possible estimate for the stationary distribution is shown. There is usually less selective pressure on the third codon position, which is why we expect a stationary distribution with high entropy.

REMARK 4.3.1. The pseudocounts control the expected entropy of samples from a Dirichlet distribution. However, by fixing the expected entropy, we have no control of the spread of entropy values. A possible workaround is to specify a prior distribution for the pseudocounts, but this would only allow to increase the spread. A more graceful but computationally much more demanding solution is to define a distribution directly on entropy values. Let $H \in [0, \log |\mathbb{A}|]$ be a random variable on entropy values. An event $\{H = \eta\}$ corresponds to multiple distributions ϑ in the probability simplex, because the entropy function is not invertible. Nevertheless, for statistical applications we have to be able to integrate over the stationary distribution Θ defined on the probability simplex. For that, the probability assigned to an entropy region $d\eta$ has to be uniformly distributed across all stationary distributions with entropy in $d\eta$. Let h again denote the entropy function. Formally, for a density function g on the entropy space $[0, \log |\mathbb{A}|]$ we would like to have a weight function $w : [0, \log |\mathbb{A}|] \rightarrow [0, \infty)$ such that

$$\text{pr}_H(h(A)) = \int_{h(A)} g(\eta) d\eta = \int_A g_h(\vartheta) (w \circ h)(\vartheta) d\vartheta$$

for all measurable sets A in the probability simplex $\Delta_{|\mathbb{A}|-1}$, where $g_h = (g \circ h)$. A possible solution for the weight function is given by

$$w(\eta) = \lim_{\epsilon \rightarrow 0} \frac{\epsilon}{\int_{A_\epsilon(\eta)} d\vartheta}, \quad \text{where } A_\epsilon(\eta) = \{\vartheta \mid h(\vartheta) \in [\eta, \eta + \epsilon]\}.$$

For the one-dimensional simplex (i.e. $|\mathbb{A}| = 2$), the weight function w can be computed analytically. Since the preimage of each entropy value has only two elements, we may compute the weight by using the change of variables theorem and correcting by $1/2$. Hence,

$$d\eta = (w \circ h)(\vartheta) d\vartheta = \frac{1}{2} \left| \frac{d}{d\vartheta} h(\vartheta) \right| d\vartheta = \frac{1}{2} |\log(1 - \vartheta) - \log(\vartheta)| d\vartheta.$$

In general, there exists no analytical form of the weight function, but we may use a numerical approximation instead (see Figure 3). Once an approxima-

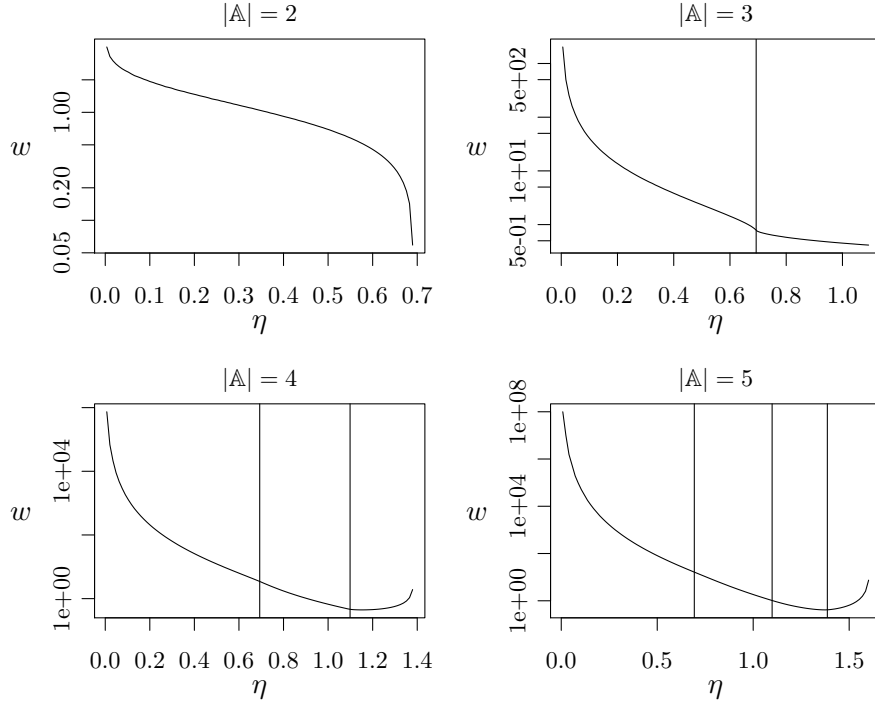


FIGURE 3. Numerical approximation (up to a constant) of the weight function w for different alphabet sizes. Vertical lines show maximal entropy values for cardinalities $k = 2, \dots, 5$.

tion is given, we may define a distribution on the entropy space $[0, \log |\mathbb{A}|]$ and map this distribution to the probability simplex. Possible choices are for instance the logitnormal or beta distribution rescaled to $[0, \log |\mathbb{A}|]$. For the following demonstration we assume the latter so that

$$g(\eta) \propto \left(\frac{\eta}{\log |\mathbb{A}|} \right)^{\beta_1 - 1} \left(1 - \frac{\eta}{\log |\mathbb{A}|} \right)^{\beta_2 - 1}$$

with pseudocounts β_1 and β_2 . The resulting density on the probability simplex is shown in Figure 4 for $|\mathbb{A}| = 3$ and two different parameter values.

The next step is to formulate a prior distribution on the edge lengths given a fixed topology. By this we obtain the posterior μ_i for a single orthant

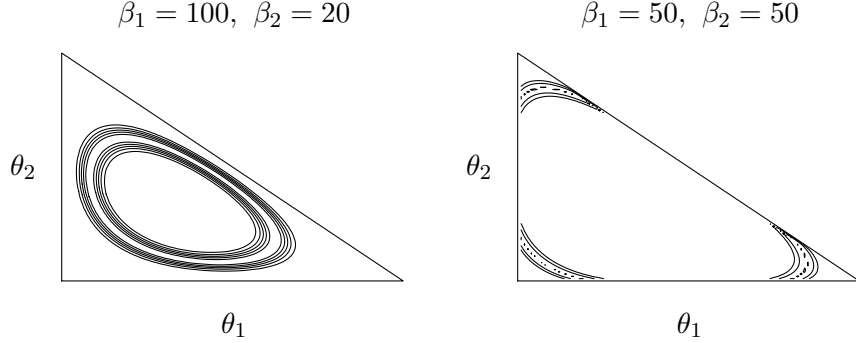


FIGURE 4. Two distributions on entropy values mapped onto the probability simplex.

\mathcal{O}_i . The same phylogenetic tree is assumed for all columns in the alignment. In fact, columns in the alignment are conditionally independent given a fixed phylogenetic tree. Since we now want to let the tree vary within one orthant of tree space, it is necessary to consider the full alignment. Let \bar{X}_n denote the random variables for the n -th column of the alignment, where $n = 1, \dots, N$. We also use the shorthand notation $\mathbf{X} = (\bar{X}_1, \dots, \bar{X}_N)$ for the full alignment. Let $\mathbf{L} = (L_k)$ denote the random variables for the edge lengths of a tree t in orthant \mathcal{O}_i . Each L_k is a priori gamma distributed with shape parameter b and scale parameter λ . The likelihood of the full alignment is given by

$$\text{pr}_{\mathbf{X}|\mathbf{L},\mathcal{O}_i}(\mathbf{x}|\mathbf{l}) = \prod_{n=1}^N \text{pr}_{\bar{X}_n|\mathbf{L},\mathcal{O}_i}(\bar{x}_n|\mathbf{l}),$$

where the stationary distribution is integrated out, and we obtain the posterior distribution μ_i restricted to orthant \mathcal{O}_i with density function

$$f_{\mathbf{L}|\mathbf{X},\mathcal{O}_i}(\mathbf{l}|\mathbf{x}) = \frac{1}{\text{pr}_{\mathbf{X}|\mathcal{O}_i}(\mathbf{x})} \text{pr}_{\mathbf{X}|\mathbf{L},\mathcal{O}_i}(\mathbf{x}|\mathbf{l}) f_{\mathbf{L}}(\mathbf{l}).$$

The full posterior distribution on \mathcal{T}_n is given by

$$\mu = \sum_{i=1}^{(2n-3)!!} w_i \mu_i,$$

where

$$w_i = \frac{\text{pr}_{\mathbf{X}|\mathcal{O}_i}(\mathbf{x})}{\sum_j \text{pr}_{\mathbf{X}|\mathcal{O}_j}(\mathbf{x})}$$

is the weight of the i -th component. We will denote the density function of μ simply as f . The weight w_i depends on the normalized partition function of μ_i , which involves computing an intractable integral. Another difficulty

is that the number of orthants grows super-exponentially with the number of leaves. It is therefore necessary to approximate μ with a Dirac mixture of posterior samples, which does not require to compute any partition functions.

To obtain samples from the posterior distribution, a Metropolis-coupled MCMC algorithm is discussed in the following section. In the sequel, these samples will serve as input to the reconstruction of the posterior median, mean, and consensus trees.

4.4. Approximation of the posterior distribution

The structure of the posterior distribution on the tree space \mathcal{T}_n is too complicated for an analytic treatment. Point estimates can nevertheless be computed from a set of posterior samples, which are obtained from Markov chain Monte Carlo (MCMC) simulations [Yang and Rannala, 1997, Li et al., 2000, Huelsenbeck and Ronquist, 2001, Drummond and Rambaut, 2007, Lartillot et al., 2009] (see also section A.4). The main difference to existing methods is that we have an analytic treatment of the stationary distribution (see section 3.3). The method discussed here should converge much faster than existing methods, since there is a strong coupling between the entropy of the stationary distribution and the branch lengths of the phylogenetic tree (see the discussion in section 4.3).

To summarize the posterior μ , we would like to compute a point estimate

$$\hat{t} = \arg \min_{s \in \mathcal{T}_n} \int_{\mathcal{T}_n} \mathcal{L}(s, t) d\mu(t),$$

for an appropriate loss function \mathcal{L} . Unfortunately, the expected loss is difficult to compute and we therefore rely on an approximation by replacing μ with the Dirac mixture

$$\pi = \frac{1}{K} \sum_{k=1}^K \delta_{t_k}$$

of K samples from μ . By the ergodic theorem, we have the convergence

$$\int_{\mathcal{T}_n} \mathcal{L}(s, t) d\pi(t) = \frac{1}{K} \sum_{k=1}^K \mathcal{L}(s, t_k) \rightarrow \int_{\mathcal{T}_n} \mathcal{L}(s, t) d\mu(t),$$

almost surely for every $s \in \mathcal{T}_n$ as $K \rightarrow \infty$ [Robert and Casella, 1999]. A set of posterior samples can be obtained with the Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] without having to evaluate the weights w_i of the single components of μ . The algorithm constructs a Markov chain with μ as the stationary distribution. Let t_k be a sample from μ with edge set \mathcal{E} . A new sample t_{k+1} is generated by the Markov chain conditional on the current sample t_k . The algorithm uses a proposal

distribution with density function $q(\cdot | t_k)$, which selects an edge $e \in \mathcal{E}$ and replaces it by another edge. We thereby obtain a new tree s that we accept as the next sample t_{k+1} with probability

$$\rho(t_k, s) = \min \left\{ 1, \frac{f(s)q(t_k | s)}{f(t_k)q(s | t_k)} \right\},$$

and otherwise $t_{k+1} = t_k$, where f still denotes the density function of μ . Note that the normalization constant of μ cancels in the ratio. The proposed tree s lies in the same orthant as t_k with probability τ . In this case, a new edge length is proposed, which is a draw from a normal distribution centered at $|e|$. However, with probability $1 - \tau$ the proposed tree lies within one of the neighboring orthants (NNI move), by replacing the edge e by one of two other possible edges of the same length (see chapter 2).

It is well known that the posterior distribution μ often shows multiple modes that are distributed in tree space. The Metropolis-Hastings algorithm might not converge well if the modes are separated by regions of low probability. In this case, it was proposed by Geyer [1991], Geyer and Thompson [1995] to augment the sampler by several heated chains that can jump more easily between modes. A heated chain uses the regular Metropolis-Hastings acceptance probability, but the density of the posterior distribution f is replaced by

$$g_\tau(t) = [f(t)]^{-\tau},$$

where $\tau \geq 1$ is the temperature of the chain. Increasing the temperature above one flattens the posterior distribution and allows the sampler to easily switch between modes.

The Metropolis-coupled Markov chain Monte Carlo (MC³) method proposed by Geyer [1991] uses one cold chain ($\tau = 1$) and several heated chains with different temperatures. By swapping states between the cold and heated chains, we improve the mixing behavior of the cold chain. Let τ_i and τ_j be the temperatures of chains i and j . A swap of states between the chains is accepted with probability

$$\min \left\{ 1, \frac{g_{\tau_i}(t_j)g_{\tau_j}(t_i)}{g_{\tau_i}(t_i)g_{\tau_j}(t_j)} \right\},$$

where t_i and t_j are the states of chain i and j . Asymptotically, the chains are independent with the respective stationary distributions [cf. Geyer, 1991].

4.5. MCMC with Hamiltonian dynamics

For the proposal distribution of the Metropolis-Hastings algorithm we have selected a normal distribution with the same variance for all branch lengths. This choice leads in most cases to good results. However, in certain

cases it is more difficult to obtain good proposals, for instance, if the alignment contains missing data, which is heterogeneously distributed among species.³ By augmenting the Metropolis-Hastings algorithm with Hamiltonian dynamics the actual shape of the posterior distribution can be taken into account [Duane et al., 1987, Neal, 1995, 1996, 2011]. However, the disadvantage of Hamiltonian Monte Carlo (HMC) is its computational cost, which is mostly caused by the computation of gradients.

For the following discussion we restrict our attention to one orthant \mathcal{O}_i of \mathcal{T}_n . Let $l = (l_1, \dots, l_m) \in \mathbb{R}^n$ represent the edge lengths of phylogenetic trees in \mathcal{O}_i and let f be the posterior density restricted to this orthant. We write the posterior density as $f(l) = \int g(l, p) dp$, where $g : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ is defined as

$$g(l, p) \propto \exp \{-U(l) - K(p)\} .$$

The density g can be understood as an extension of f to the product space $\mathbb{R}^m \times \mathbb{R}^m$ with two independent components. Inspired by Hamiltonian dynamics, l could be interpreted as the position of a particle and p as its momentum (mass times velocity). The Hamilton function is

$$H(l, p) = U(l) + K(p) ,$$

where U is the potential energy and K the kinetic energy of the particle. Furthermore, we define the dynamics

$$\frac{d}{dx} l_j(x) = \frac{\partial}{\partial p_j} K(p) , \quad \frac{d}{dx} p_j(x) = -\frac{\partial}{\partial l_j} U(l) ,$$

for $j = 1, \dots, m$, which is reversible and volume preserving. The variable x may be interpreted as time.

The basic idea behind Hamiltonian Monte Carlo is as follows. Since the position and momentum variables are independent, we may select a simple distribution for the momentum so that samples can be drawn easily. For instance, the momentum is Gaussian distributed with mean zero and variance one if $K(p) = p^2/2$. Let $l^{(k)}$ be the current position and p^* the newly drawn sample for the momentum. A new proposal (l', p') for both variables is computed by simulating the Hamiltonian dynamics for a given time period using $(l^{(k)}, p^*)$ as initial condition. The momentum variable is negated at the end of the trajectory.⁴ The proposal is accepted as the new

³The reason for this effect is the time-reversibility of the mutation model. The less data is observed in a given species, the higher the uncertainty about the actual position of the respective branch in the tree.

⁴For typical choices of K we have $K(p) = K(-p)$ and therefore the momentum need not be negated in practice.

sample $(l^{(k+1)}, p^{(k+1)})$ with probability

$$\min \left\{ 1, \exp \left(-U(l') + U(l^{(k)}) - K(p') + K(p^{(k)}) \right) \right\}.$$

The acceptance probability will always be close to one, since the value of the Hamiltonian function H is preserved by the dynamics (except for numerical errors).

In principle any type of reversible dynamics could be used. The advantage of Hamiltonian dynamics is that it is volume preserving, which significantly simplifies the computation of the acceptance probability. Otherwise, changes in volume must be accounted for. The same argument holds for the numerical integration method, which is why the *leapfrog* method is commonly used for HMC. Additionally, it is crucial that the dynamics of the discretized system are reversible as well. This is indeed the case for Hamiltonian dynamics integrated with the leapfrog method.

In practice it is often beneficial to only partially resample the momentum variables. Assume that the kinetic energy has the form $K(p) = p^\top \Sigma^{-1} p / 2$ where Σ is the covariance matrix. We update the current momentum p by

$$p^* = \alpha p + (1 - \alpha^2)^{1/2} \bar{p},$$

where $\alpha \in [-1, 1]$ and \bar{p} is a random vector from a Gaussian distribution with mean zero and covariance Σ . For α close to one, new values will be very similar to the old momentum p .

HMC is easy to implement for sampling phylogenetic trees. The gradient of the log likelihood function can be computed from the recursive solution derived in section 3.3. An example is shown in Figure 5 for a synthetic alignment.

4.6. Estimation results

Using a multiple sequence alignment from a study by Karol et al. [2001], which was slightly modified by Yang and Rannala [2005], the phylogeny of the small subunit rRNA gene (SSU rRNA) from the nuclear genome of 8 land plants and 6 charales (see Figure 6) has been reconstructed. It appears that the edge that separates *Psilotum nudum* and *Dicksonia antarctica* from the remaining tree has a very short length of approximately 0.0027. Figure 7 shows the marginal posterior distribution of this edge (e_1) and a competing one (e_2) that groups *Psilotum nudum* with *Taxus baccata* and *Arabidopsis thaliana*. There remains a high posterior uncertainty about the exact topology of the tree at this very branching. The posterior mass on e_1 is however sufficient for the majority-rule consensus tree to include this edge. However, there, it has a much longer length than in the posterior mean tree (by approximately 0.011), since this length is obtained by averaging only the

```

0: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA...AAAAAAAAA
1: TTTTTTTTTTTAAAAAAAAAAAAAAAAAAAA...AAAAAAAAA
2: GGGGGGGGGGGCCCCCCCCCCCCCCCCCCCC...CCCCCCCCC
3: CCCCCCCCCCCCCCCCCCCCCCCCCCCCCC...CCCCCCCCC
4: ACGTACGTACGTNNNNNNNNNNNNNNNNNN...NNNNNNNNN

```

(A)

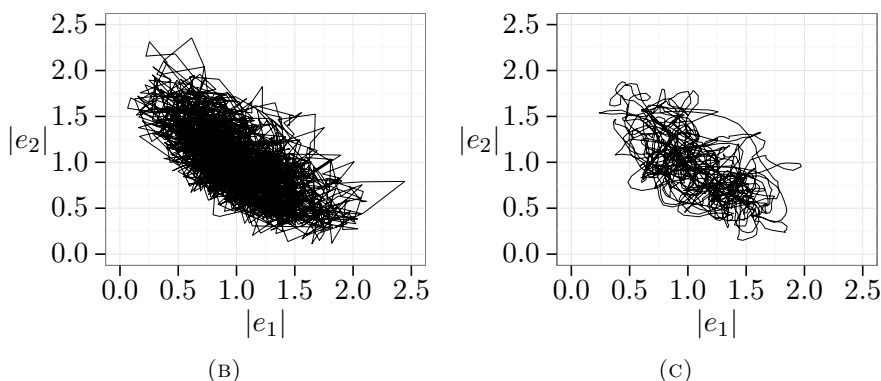


FIGURE 5. Demonstration of the Hamiltonian Monte Carlo method. (A) Data set of length 100. The character “N” is interpreted as missing data. (B) and (C) show samples generated by the HMC method for different parameters. (B) 10 leapfrog steps between samples with step size 0.01 and $\alpha = 0$. (C) 10 leapfrog steps between samples with step size 0.001 and $\alpha = 0.95$.

lengths of this edge when it occurs in the sample, neglecting the contributions of the alternative edges. Clearly, the shorter edge length born by the posterior mean tree better accounts for the uncertainty, and the consensus tree appears, in contrast, to have overestimated branch lengths.

The assessment of reconstruction methods for phylogenetic trees is notoriously hindered by the ignorance of the true evolutionary history to be uncovered, as the latter is never observed. Instead, the estimated tree of Figure 6 has been used to generate 50 alignments of length $m = 50, 100, 250,$ and 500 . For each generated data set, 210,000 posterior samples were obtained using one cold Markov chain and three heated chains. The Fréchet mean, geometric median and consensus tree of the last 200,000 samples were computed. In the whole study, tree topologies are uniformly distributed a priori, while branch lengths are distributed according to a $\text{Gamma}(1, 0.4)$.

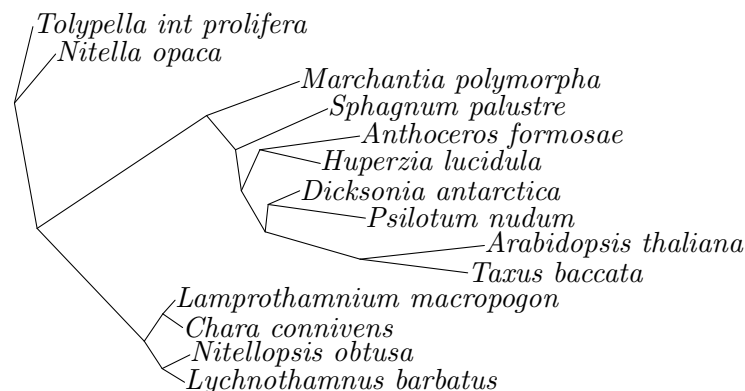


FIGURE 6. Fréchet mean estimated from the small subunit rRNA gene (SSU rRNA) from the nuclear genome of 8 land plants and 6 charales. Edge lengths are plotted in horizontal direction only.

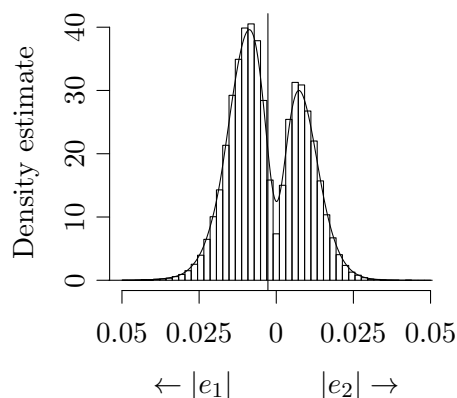


FIGURE 7. Marginal posterior density estimate of two edges e_1 and e_2 . The edge e_1 groups *Psilotum nudum* with *Dicksonia antarctica*, while e_2 groups *Psilotum nudum* with *Taxus baccata* and *Arabidopsis thaliana*. The Fréchet mean is shown as a vertical line.

Figure 8 shows the distances of the computed estimates to the generating tree. For alignments of the lengths considered, the Fréchet mean and geometric median are generally closer to the generating tree. A trend appears, from the greatest discrepancy observed for the shortest alignments, to an almost systematic agreement for the longest alignments. It should be noted that even shorter alignments generally result in so broadly distributed a posterior distribution that all three estimates coincide with the star tree. At the opposite extreme, large datasets support a clear decision about the

topology of the tree, placing most of the mass of the posterior distribution in a single orthant, and resulting in mostly agreeing estimates. One also observes that the geometric median is in most cases closer to the generating tree than the Fréchet mean. This comes at no surprise given the skewness of the gamma prior on the branch lengths.

This model permits an analytical computation of the marginal likelihood of an alignment given a phylogenetic tree, thereby offering an evaluation of how the estimated model generalizes to novel observations. Using a leave-one-out approach, the average (unnormalized) posterior value achieved by the estimators were computed on the remaining 49 data sets of the same length (see Table 1). For all alignment lengths, the Fréchet mean and geometric median show a slightly higher average posterior value compared to the majority-rule consensus tree. But the difference is too minor to make any definite statements, as also shown by the variance of the estimates. A much clearer picture is gained by considering how often the mean and median have a higher posterior value than the consensus (see Table 1). The results show that the consensus tree clearly performs worse.

Another quantity of interest is the Fréchet variance of the posterior distribution, which provides us with a measure of uncertainty. The mean variance is shown in Figure 9 separately for all four data set lengths. Similar to the case of normal distributed i.i.d. random variables, the variance decreases approximately with $1/m$. Another, maybe more intuitive statistic, is to compute a credibility region around the Fréchet mean \hat{t} that contains a given proportion c of the posterior mass. More precisely, consider the set of trees $B = \{t \in \mathcal{T}_n \mid d(\hat{t}, t) \leq d^*\}$ for some d^* such that $\int_B d\mu(t) = c$. The bound d^* may be called the credibility radius. Figure 9 shows the results for $c = 0.68$.

m	(a) posterior			(b) performance	
	mean	median	consensus	mean	median
50	-201.94 (10.99)	-201.24 (10.84)	-203.22 (10.10)	0.78	0.80
100	-407.13 (4.67)	-407.03 (4.84)	-409.39 (3.66)	0.86	0.82
250	-1035.66 (5.12)	-1035.62 (5.32)	-1036.72 (5.02)	0.90	0.90
500	-2074.11 (3.75)	-2074.01 (3.75)	-2074.96 (3.99)	0.78	0.80

TABLE 1. (a) Mean posterior values for the Fréchet mean, geometric median, and consensus tree. The variance is shown in brackets. (b) Percentage of times the mean and median show a higher posterior value on the remaining (joined) 49 data sets. Both statistics were evaluated separately on data sets of length $m = 50, 100, 250$, and 500 .

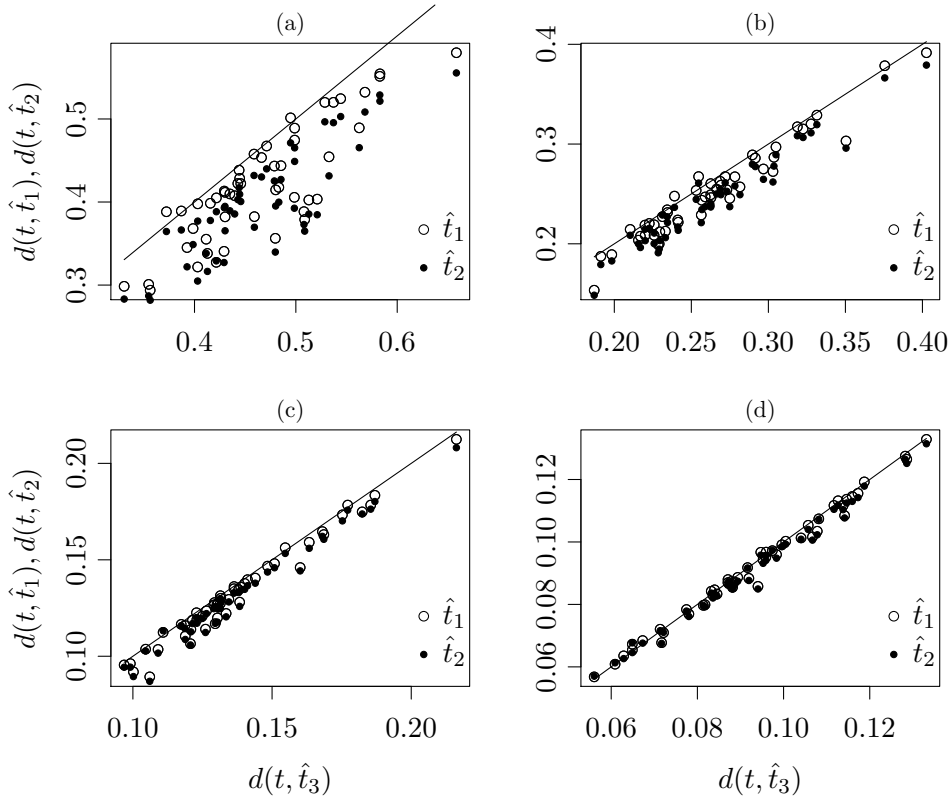


FIGURE 8. Distances $d(\cdot, \cdot)$ of the Fréchet mean \hat{t}_1 , geometric median \hat{t}_2 , and consensus tree \hat{t}_3 to the generating tree t for alignments of length 50 (a), 100 (b), 250 (c), and 500 (d). The straight line shows the main diagonal.

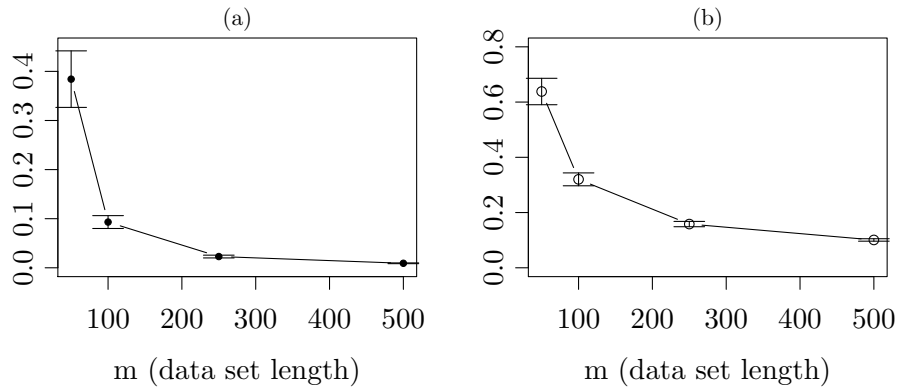


FIGURE 9. Average Fréchet variance (a) and credibility radius d^* (b) for datasets of length 50, 100, 250, and 500. The error bars show one standard deviation.

CHAPTER 5

Prediction of transcription factor binding sites

To identify transcription factor binding sites, we usually have data from either a collection of regulatory regions of genes that are assumed to be coregulated, or experimental data from ChIP-seq experiments [Park, 2009]. In the latter, transcription factors that are bound to DNA fragments are extracted using antibodies and the fragments are sequenced and mapped to the genome. The major difficulty here is that the fragments are typically much longer than the TFBS, hence the need for a statistical analysis. By the design of the experiment, we know that the data is *enriched* with binding sites. However, a transcription factor might recognize several DNA sequences and therefore we may not assume that binding sites are unique. Searching for the correct pattern within nucleotide sequences that reflects the binding preferences of a transcription factor is a very difficult problem, which requires additional information. By augmenting each sequence by a collection of orthologs¹ from related species a set of multiple sequence alignments can be constructed that reveal information about whether or not a given site is under selective pressure, a technique called *phylogenetic footprinting* [Wasserman and Sandelin, 2004]. Of course, the orthologs have to be selected with great care and only species that are likely to show the same regulatory mechanism are eligible. On the other hand, the orthologs must show a certain level of variation in order to gain additional information.

Essential to the analysis is a phylogenetic tree that can be utilized to model both functional and nonfunctional regions within the data. The model for nonfunctional regions, which we refer to as the *background model*, is important to identify columns in the alignments that are atypical in the sense that they are more conserved than most other regions. In contrast, the method developed by Siddharthan et al. [2005] uses the phylogenetic information only for functional regions, all other sequences are treated as mutually independent. Hence, the model considers regions as atypical that show different summary statistics on the individual sequence level.

¹Two sequences are orthologous if they descended from the same ancestral sequence.

5.1. Statistical model

The data consists of a set of multiple sequence alignments, where each alignment contains ortholog sequences from M related species. Sequences in an alignment consists of characters from an alphabet \mathbb{A} , which contains a character for each nucleotide and one that codes for gaps. To simplify notation we discuss the statistical model for only one multiple sequence alignment. A crucial assumption is that transcription factor binding sites are present and conserved in all M sequences of the alignment. The statistical model consists of two layers. The first is the phylogenetic dimension, represented by a tree, which models the dependencies between sites within a column of the alignment. The second layer models the dependencies between different positions of the alignment and consists of two components. The *background* component models regions of low conservation, whereas the *foreground* component consists of a set of patterns, instances of which model the highly conserved regions of the data. Due to the experimental procedure of how the data is obtained, we assume that the pattern of the transcription factor binding site has many instances in the data set. However, we do not assume that this is the only pattern and we do not a priori fix the number of patterns to be observed.

By defining a statistical model we assign a probability to every possible observation. A model can be outlined in several ways. A most natural possibility is to follow the causal direction, i.e. to describe the model from a generative perspective.

The top most part of the model is a process that assigns each site to either foreground or background. The foreground is assumed to consist of blocks of a fixed length L . For each position t we introduce a random variable Z_t that indicates whether the block of length L starting at t is assigned to the foreground. The distribution of Z_t therefore depends on the $L - 1$ preceding random variables. We assume that blocks cannot overlap and if no foreground block starts at the $L - 1$ positions before t , then

$$Z_t \sim \text{Bernoulli}(\lambda)$$

where λ is foreground rate. A sequence alignment consists of

$$N = \sum_{t=1}^T Z_t$$

foreground blocks. Blocks are grouped into clusters, which share a common pattern. We assume that some patterns have many instances in the data set, while most patterns occur only a few times. To model this, we introduce

N random variables Y_n with

$$Y_n | Y_1 = y_1, \dots, Y_{n-1} = y_{n-1} \sim \sum_{k=1}^{K_n} \frac{c_n(k) - d}{a + n - 1} \delta_k + \frac{a + K_n d}{a + n - 1} \delta_{K_n+1}$$

where $K_n = \max\{y_1, \dots, y_{n-1}\}$ is the number of clusters at sampling step n , $c_n(k) = \sum_{i=1}^{n-1} 1(Y_i = k)$ the number of instances in cluster k , and δ_k the Dirac delta function at k .² The process has two parameters, a is called the concentration parameter and controls the expected number of clusters at a finite sampling step, whereas d is called the discount parameter and controls the tail behavior. For $d = 0$ the process is known as the *Chinese restaurant process* introduced by Dubins and Pitman [cf. e.g. Aldous, 1985, Pitman, 1995]. The process is discussed in more detail in section 5.2.

For each position in the alignment we associate a random variable Θ_t , which is Dirichlet distributed and gives the probability for observing each of the nucleotides at this position. As we point out later, this is also the stationary distribution of the nucleotide substitution model used for the phylogenetic tree. If position t is associated with the background, then

$$\Theta_t \sim \text{Dirichlet}(\beta)$$

where $\beta = (\beta_1, \dots, \beta_{|\mathbb{A}|})$ are the pseudocounts of the Dirichlet distribution. The pseudocounts might be considered as unknown, in which case we use a gamma prior distribution. Instead of a column specific stationary distribution we may also assume that the stationary distribution at position t is drawn from a mixture distribution of J components, i.e.

$$\Theta_t | \Theta_1^* = \theta_1, \dots, \Theta_J^* = \theta_J \sim \sum_{j=1}^J w_j \delta_{\theta_j},$$

where w_j are the prior weights and $\Theta_j^* \sim \text{Dirichlet}(\beta)$.

The situation is more complicated for the foreground, since a pattern is supposed to be shared among several blocks. For each of the $K = K_N$ clusters we have a pattern $\tilde{\Theta}^{(k)}$ of length L with distribution

$$\tilde{\Theta}^{(k)} = (\tilde{\Theta}_1^{(k)}, \dots, \tilde{\Theta}_L^{(k)}) \sim (\text{Dirichlet}(\alpha), \dots, \text{Dirichlet}(\alpha))$$

where $\alpha = (\alpha_1, \dots, \alpha_{|\mathbb{A}|})$ are the pseudocounts of the Dirichlet distributions. The length L may vary across clusters and we assume a uniform prior distribution for all possible lengths. If positions $\{t, \dots, t+L-1\}$ belong to the foreground and are assigned to cluster k , then

$$(\Theta_t, \dots, \Theta_{t+L-1}) = \tilde{\Theta}^{(k)} \quad \text{or} \quad \text{rc}(\Theta_{t+L-1}, \dots, \Theta_t) = \tilde{\Theta}^{(k)},$$

²In contrast, Siddharthan et al. [2005] merely used an exponential prior on the *number of components*.

each with probability $1/2$ and where rc denotes the *reverse complement*.

Columns in the alignment are assumed to be independent given the stationary distributions (Θ_t) . It is therefore sufficient to discuss the lower part of the model for only one column. For each of the M ortholog sites in the column we have a random variable $X^{(i)}$ that takes values in the alphabet \mathbb{A} . A phylogenetic tree with M leaves and $M - 2$ internal vertices is used to model the dependence between sites. Each leaf is associated with one of the sites in the column. The $M - 2$ internal vertices of the tree are associated with random variables $X^{(k)}$, $k = M + 1, \dots, 2M - 2$, that also take values in \mathbb{A} . We now define the substitution model on the phylogenetic tree $t_k = (X^{(k)}, t_i, t_j)$, where $X^{(k)}$ is the random variable associated with vertex k and t_i and t_j the two children. First, we assume that

$$X^{(i)} \perp\!\!\!\perp X^{(j)} \mid X^{(k)}, \Theta_t .$$

where $X^{(i)}$ and $X^{(j)}$ are the random variables associated with vertices i and j . The substitution process is defined as

$$X^{(i)} \mid X^{(k)} = x, \Theta_t = \vartheta \sim M_i \text{Categorical}(\vartheta) + \bar{M}_i \delta_x$$

which was introduced by Felsenstein [1981]. The probability of a mutation from vertex k to vertex i is denoted M_i and depends on the length d of the edge that connects the two vertices, i.e. $M_i = 1 - \exp(-d)$. Furthermore, \bar{M}_i denotes the probability of no mutation, given as $\bar{M}_i = 1 - M_i$. Whenever there is a mutation between two vertices, a new nucleotide is drawn from the categorical distribution. The stationary distribution Θ_t therefore plays a crucial role. If $\{\Theta_t = \vartheta\}$ has low entropy, i.e. only one character of the alphabet \mathbb{A} has high probability, then this character is repeatedly drawn whenever there is a mutation and the column of the alignment will most likely be highly conserved. On the other hand, a high entropic stationary distribution causes a column to be poorly conserved with high probability. This observation motivates the choice of pseudocounts α and β , i.e. for the distribution of patterns we choose $\alpha < 1$ to put mass close to the faces of the probability simplex and for the background we would typically choose $\beta > 1$. See also section 3.1 for a more elaborate discussion of the nucleotide substitution process.

5.2. Process priors

Many applications require mixture models where the number of components is *a priori* not fixed to some finite value. Such priors were first discussed in the context of the *species sampling problem* [cf. e.g. Good,

1953, Bunge and Fitzpatrick, 1993, Zabell, 2005]. From the theoretical perspective, such models can either assume an infinite number of components, or a finite but unknown number.

Among the most commonly used priors for infinite mixtures is the *Dirichlet process*. It was first described with the notion of tailfree measures [cf. Fabius, 1964] by Freedman [1963] and later analyzed in detail by Ferguson [1973], Blackwell and MacQueen [1973]. Antoniak [1974] introduced Dirichlet process mixtures that convolve the random measure with a continuous distribution since random measures drawn from a Dirichlet process are discrete. Good summaries on the topic were written by Ferguson et al. [1992] and Teh et al. [2006]. Closely related to the Dirichlet process is the two-parameter extension discussed by Pitman [1995, 1996], Pitman and Yor [1997]. Mixture models with finite but a priori unknown number of components can be derived from this process [cf. Gnedin, 2010, Poppe, 2015].

Consider the simple model

$$\begin{aligned}\Pi &\sim \text{Dirichlet}(a/K, \dots, a/K) \\ X_n | \Pi = \pi &\sim \text{Discrete}(\pi),\end{aligned}$$

where Π has cardinality K (i.e. Π is the $K - 1$ -dimensional probability simplex) and (X_n) is a sequence of N random variables. Furthermore, let (Y_n) be random variables that relabel the outcomes of (X_n) according to their first occurrences. For instance, consider for $n = 5$ the observation

$$\begin{aligned}(X_1, \dots, X_5) &= (4, 2, 4, 1, 2) \\ (Y_1, \dots, Y_5) &= (1, 2, 1, 3, 2).\end{aligned}$$

Here, $Y_1 = Y_3 = 1$ because 4 is the first label that was observed. For $K \rightarrow \infty$ we call (Y_n) the *Chinese restaurant process*. The name is due to a metaphor by Dubins and Pitman [cf. Aldous, 1985, Pitman, 1995]. In Figure 1a a sample from a Chinese restaurant process is shown. For $a, N \gg 0$ the expected number of distinct labels (clusters) equals approximately

$$a \log \left(1 + \frac{N}{a} \right),$$

which shows that in the limit $n \rightarrow \infty$ we would still observe infinitely many clusters, but the number grows only logarithmically. Hence, some labels appear much more frequent than others, as can be seen in Figure 1a.

In practice it is much more convenient to use the sequential construction of the Chinese restaurant process. Suppose that Y_n is conditionally distributed as

$$Y_n | Y_1 = y_1, \dots, Y_{n-1} = y_{n-1} \sim \sum_{k=1}^{K_n} \frac{c_n(k)}{a+n-1} \delta_k + \frac{a}{a+n-1} \delta_{K_n+1}$$

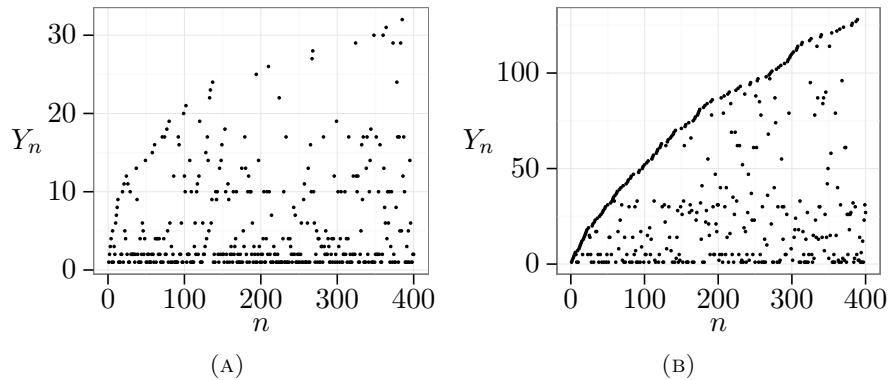


FIGURE 1. (A) Samples from a Chinese restaurant process with $a = 10$. (B) Samples from the Ewens-Pitman process with $a = 10$ and $d = 0.5$.

where $K_n = \max\{y_1, \dots, y_{n-1}\}$ is the number of clusters at sampling step n and $c_n(k) = \sum_{i=1}^{n-1} 1(Y_i = k)$ the number of instances in cluster k . The process is parameterized by $a > -d$ and $d \in [0, 1]$ and usually referred to as the *Ewens-Pitman two-parameter family* [Gnedin and Pitman, 2006]. For $d = 0$ it coincides with the Chinese restaurant process. Interestingly, the process converges to finitely many clusters if $d < 0$. In fact, the number of clusters is determined by $-\alpha/d$. A process with finite but unknown number of clusters can be obtained by mixing over α with $d = -1$ fixed [Gnedin, 2010].

A sample of the process for $d = 1/2$ is shown in Figure 1b. Instead of the logarithmic growth, the Ewens-Pitman process shows a power law behavior. From the conditional distribution of Y_n one can see that the weight of a cluster increases with the number of instances c_n , a property often called *rich get richer*. For practical applications it is very important to be aware of this property, since it introduces a strong prior assumption that has to be justified.

In the context of mixture models, a realization of (Y_n) is used to partition a set of observations $\{Z_n = z_n\}$ into clusters. Each cluster is associated with one component of the mixture model. Usually, a component is a member of a family $\{F(\theta)\}$ of distributions indexed by some parameter θ (for instance, F could be a normal distribution and θ_n its location parameter), so that each event $\{Y_n = y_n\}$ is associated with a member of that family. By assigning a

prior distribution G_0 to the parameters of $\{F(\theta)\}$, we obtain the model

$$\begin{aligned}\Theta_k^* &\sim G_0 \\ Z_n | \{\Theta^* = \theta^*, Y_n = y_n\} &\sim F(\theta_{y_n}^*),\end{aligned}$$

where $\{\Theta^* = \theta^*\} = \{\Theta_1^* = \theta_1^*, \Theta_2^* = \theta_2^*, \dots\}$ is a set of distinct parameters. The combination of this model and the Chinese restaurant process results in the Pólya urn scheme by Blackwell and MacQueen [1973]. In this form, each Z_n is associated with a random variable Θ_n , which is conditionally distributed as

$$\Theta_n | \Theta_1 = \theta_1, \dots, \Theta_{n-1} = \theta_{n-1} \sim \sum_{i=1}^{n-1} \frac{1}{a+n-1} \delta_{\theta_i} + \frac{a}{a+n-1} G_0,$$

so that

$$Z_n | \{\Theta_n = \theta_n\} \sim F(\theta_n).$$

The closely related Dirichlet process can be derived as follows. Let G_0 be again some distribution and consider the model

$$\begin{aligned}\Pi &\sim \text{Dirichlet}(a/K, \dots, a/K) \\ \Theta_k^* &\sim G_0 \\ G | \{\Pi = \pi, \Theta^* = \theta^*\} &= \sum_{k=1}^K \pi_k \delta_{\theta_k^*},\end{aligned}$$

where $\{\Theta^* = \theta^*\} = \{\Theta_1^* = \theta_1^*, \dots, \Theta_K^* = \theta_K^*\}$. For $K \rightarrow \infty$ we call this model the Dirichlet process and use the shorthand notation

$$G \sim \text{DP}(a, G_0).$$

G is a distribution with discrete components and sampling from it is equivalent to the Pólya urn scheme derived above. The corresponding model is

$$\begin{aligned}\Theta_n | G = g &\sim g \\ Z_n | \Theta_n = \theta_n &\sim F(\theta_n).\end{aligned}$$

Although the notation of the Dirichlet process might be convenient for infinite mixture models, the Pólya urn scheme is much easier to interpret and work with.

One reason why the Dirichlet process has received so much attention in the past is that the distribution of $\{\Theta_n\}$ is exchangeable, i.e.

$$\Theta_1, \dots, \Theta_n \stackrel{d}{=} \Theta_{\sigma(1)}, \dots, \Theta_{\sigma(n)}$$

for any permutation σ . Despite the popularity, such *nonparametric models* might suffer from inferential inconsistencies (see section A.8 for a definition). A first example of inconsistency was provided by Diaconis and Freedman [1986], which is however quite artificial. Miller and Harrison [2013] discussed the problem of estimating the number of components K of a finite mixture model. They found that the posterior probability of the true number of components goes to zero as the number of observations increases, which is expected since the Dirichlet process assumes an infinite number of components. On the other hand, posterior estimates of mixture densities are consistent, as shown by Ghosal et al. [1999].

5.3. The label switching problem

Instead of predicting the locations of transcription factor binding sites, one might think of inferring the patterns $\tilde{\Theta}^{(k)}$ directly. This approach turns out to be infeasible due to a technical difficulty. Consider a mixture distribution with K components and density function

$$f(x; \mu) = \sum_{k=1}^K w_k g_{\mu_k}(x),$$

where w_k is the weight of the k th component and $\{g_{\mu_k}\}$ any family of density functions. The parameters $\mu = (\mu_1, \dots, \mu_K)$ are thought to correspond to the patterns $\tilde{\Theta}^{(k)}$. A point estimate of μ might be obtained by computing the mean or median of a set of samples drawn from the posterior distribution. However, the likelihood function f is invariant under permutations of the labels k , i.e.

$$f(x; \mu) = \sum_{k=1}^K w_{\sigma(k)} g_{\mu_{\sigma(k)}}(x)$$

for any permutation σ , which shows that the posterior will usually have $K!$ distinct modes. Hence, when drawing samples from the posterior distribution any two components might switch their parameters. As a result, the estimates of all parameters will approximately be identical (see Figure 2). This problem is well known and often termed the *label switching problem* [see e.g. Celeux et al., 2000, Jasra et al., 2005]. For one-dimensional parameters, i.e. $\mu_k \in \mathbb{R}$, the problem can be solved by *a priori* enforcing a total ordering of the parameters. Obviously, this approach is not feasible in higher dimensions.

Many other solutions were proposed in the past, such as artificial ordering constraints [e.g. Stephens, 2000] and relabeling algorithms [e.g. Celeux, 1998]. Both approaches might bias the resulting estimate in a way which is

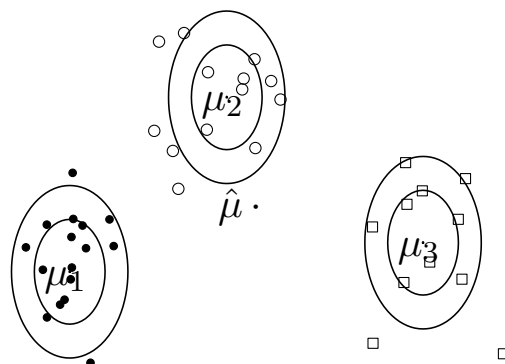


FIGURE 2. Mixture distribution with three normal components. The parameters μ_k specify the locations of the components. $\hat{\mu}$ indicates the position of an estimate for the three location parameters.

hard to quantify. Hence, it seems that the only adequate solution is to first estimate a partition of the data³.

5.4. Predictions

The goal of the statistical analysis is to obtain a prediction of what blocks share a common pattern. From this prediction it is then possible to obtain an estimate of the set of patterns. In the following, we outline the usual Bayesian methodology to obtain a posterior estimate although there exist several technical difficulties that prevent the evaluation of most point estimates on realistic data sets. The first obstacle is the convergence rate of the sampler. Woodard et al. [2013] studied the convergence rate on a model without phylogenetic information and showed that it decreases exponentially with the length of the data set. Although a clear phylogenetic signal might boost the convergence of the sampler, it is likely that the result also holds for the method described here. Another difficulty is the actual computation of the median or mean from posterior samples. A sample partitions the data such that similar conserved sites are clustered together. As explained later, the computation of the mean and median of a set of partitions is NP-hard. Instead, we may pick the sample with minimum (squared)

³As discussed in section 5.2, the inference of the number of components of a *finite* mixture model is inconsistent when a Dirichlet process prior is used. It is easy to see that this result also holds for the estimation of partitions. However, this is not a concern for the inference of transcription factor binding sites, since we assume that in principle we could observe an infinite number of different patterns.

distance to all other posterior samples and thereby obtain a crude approximation of the (mean) median. Still, the computational complexity of this estimate is quadratic in the number of posterior samples. Combining both, the slow convergence of the sampler and the limitation to few samples so that the median or mean can be computed, it is clear that this methodology is restricted to very small data sets. Instead, we may take a similar approach as Siddharthan et al. [2005] and search for a partition with high posterior probability to obtain an approximation to the maximum a posteriori (MAP) estimate. A simple hill climbing method gets easily stuck in one of the modes of the posterior distribution. Hence, the search for good local maxima should at least be partially stochastic. In the following, we derive the MCMC method for drawing samples from the posterior distribution, which accounts for the stochastic part of the search algorithm. The method consists of Gibbs updates of the hidden variables Z_t and Y_n as well as Metropolis-Hastings steps to propose new positions of whole clusters. The MCMC method can then be easily turned into a hill climbing algorithm. For the Gibbs sampler we simply select the update that maximizes the marginal posterior distribution of the respective hidden variable. The resulting algorithm is equivalent to a coordinate ascent method (see also section A.4). Regarding the Metropolis-Hastings algorithm, a proposal is accepted if it maximizes the posterior value. In this way, we may search for the global maximum by alternating between MCMC sampling and hill climbing optimization.

As pointed out earlier, the pattern of the transcription factor binding site is assumed to have many instances in the data set and should therefore be among the largest clusters. We are given a data set $D = \{X_t = x_t \mid t = 1, \dots, T\}$, where an event

$$\{Z_1 = z_1, \dots, Z_T = z_T, Y_1 = y_1, \dots, Y_N = y_N\}$$

defines a partition π of the data, which does not depend on the actual labeling of clusters. Our statistical model enables us to evaluate the posterior probability pr of a partition π given the data D , i.e. $\text{pr}(\pi \mid D)$. In statistical decision theory [cf. Lindley, 1972], a point estimate $\hat{\pi}$ is obtained by minimizing the posterior expected loss

$$\hat{\pi} = \arg \min_{\pi'} \sum_{\pi} \mathcal{L}(\pi', \pi) \text{pr}(\pi \mid D)$$

for an appropriate loss function \mathcal{L} , where the sum is over all possible partitions of the data set. Here, $\mathcal{L}(\pi', \pi)$ quantifies how much we would lose by selecting π' if π would be the better choice. Since it is computationally expensive to evaluate all possible partitions of the data, a Markov chain Monte Carlo (MCMC) method is used to draw samples from the posterior

distribution of partitions. Let $\bar{\pi} = \{\pi_1, \dots, \pi_L\}$ be a set of L samples from the posterior distribution, we use the approximation

$$\hat{\pi} = \arg \min_{\pi'} \sum_{\pi \in \bar{\pi}} \mathcal{L}(\pi', \pi) \approx \arg \min_{\pi'} \sum_{\pi} \mathcal{L}(\pi', \pi) \text{pr}(\pi | D)$$

for a large enough L .

A typical choice for the loss function is

$$\mathcal{L}(\pi_1, \pi_2) = d(\pi_1, \pi_2), \text{ or } \mathcal{L}(\pi_1, \pi_2) = d(\pi_1, \pi_2)^2,$$

where $d(\cdot, \cdot)$ is a metric, so that by minimizing the posterior expected loss we obtain the median respectively mean of the posterior distribution on a given metric space. Among the most common distance functions on the space of partitions is the *symmetric difference distance* (ssd), for which a clear axiomatic characterization exists [Mirkin and Chernyi, 1970]. A partition π induces a unique *equivalence relation*, i.e. a reflexive, symmetric, and transitive relation, denoted $r(\pi)$. The ssd is defined as

$$\begin{aligned} d(\pi_1, \pi_2) &= |r(\pi_1) \Delta r(\pi_2)| = |r(\pi_1) \cup r(\pi_2)| - |r(\pi_1) \cap r(\pi_2)| \\ &= |r(\pi_1) - r(\pi_2)| + |r(\pi_2) - r(\pi_1)|, \end{aligned}$$

where $r(\pi_1) \Delta r(\pi_2)$ is the symmetric difference. Intuitively speaking, the ssd is the number of pairs of elements on which the two partitions disagree. Rand [1971] used a related form to measure the similarity of partitions. The problem of finding a median partition was for instance considered by Régnier [1983], but from the results of Křivánek and Morávek [1986] it can be concluded that the problem is NP-hard [cf. Wakabayashi, 1998]. We therefore compute the point estimate

$$\hat{\pi} = \arg \min_{\pi' \in \bar{\pi}} \sum_{\pi \in \bar{\pi}} \mathcal{L}(\pi', \pi),$$

which is a good approximation unless the posterior distribution has multiple modes that are far apart and there is little mass between them. Local optimizations, such as removing elements from clusters, might then be used to improve $\hat{\pi}$. If the data set is very large, it might be necessary to use a zero-one loss function. The corresponding estimate is the partition with highest posterior probability, which is much easier to compute.

To obtain a set $\bar{\pi}$ of posterior samples, a hybrid sampler is used, which combines a Gibbs and Metropolis-Hastings sampler. For the following discussion, it is more convenient to introduce a different representation of the event

$$\{Z_1 = z_1, \dots, Z_T = z_T, Y_1 = y_1, \dots, Y_N = y_N\}.$$

For each position t of the alignment we introduce a random variable V_t . If position t belongs to the background model, then $V_t = 0$. Otherwise, if t

is the beginning of a block, i.e. $Z_t = 1$, then $V_t \in \{1, \dots, N\}$ is the cluster to which the block is assigned. The Gibbs sampler sequentially draws new values for V_t for all $t = 1, \dots, T$. Suppose a new value for V_t should be drawn and there is no block starting at positions t to $t + L - 1$.⁴ A Gibbs update is a draw from the distribution of $V_t | V_{-t}$. The block is assigned to the background with probability

$$(1 - \lambda)p_{\text{bg}}(x_t, \dots, x_{t+L-1})$$

where $p_{\text{bg}}(x_t, \dots, x_{t+L-1})$ is the joint probability of columns t to $t + L - 1$ under the background model. With probability

$$\lambda \frac{c(k) - d}{a + N} p_k(x_t, \dots, x_{t+L-1} | \mathbf{x}_k)$$

the block is assigned to the k th cluster of the foreground model. A block can be assigned to a cluster either in its original form or as its reverse complement, however, we will refrain from making this formal. With

$$p_k(x_t, \dots, x_{t+L-1} | \mathbf{x}_k)$$

we denote the predictive distribution of the k th cluster given the data

$$\mathbf{x}_k = \{(x_s, \dots, x_{s+L-1}) \mid \text{for all } s \text{ with } V_s = k\}$$

of all blocks that are assigned to this cluster. N denotes the number of blocks that are assigned to the foreground model and $c(k)$ the number of blocks in this particular cluster. The block may also be assigned to a new cluster with probability

$$\lambda \frac{a + dK}{a + N} p_{K+1}(x_t, \dots, x_{t+L-1})$$

where $p_{K+1}(x_t, \dots, x_{t+L-1})$ is the predictive probability of an empty cluster.

A major difficulty is the computation of the predictive distribution

$$p_k(x_t, \dots, x_{t+L-1} | \mathbf{x}_k) = \frac{p_k(x_t, \dots, x_{t+L-1}, \mathbf{x}_k)}{p_k(\mathbf{x}_k)},$$

which involves evaluating the joint marginal likelihood of a set of columns. The likelihood of a single column depends on the phylogenetic tree and the stationary distribution. In our model we assumed that the tree is the same at each position of the alignment, hence we may use the simplified notation

$$\text{pr}_{X|\Theta}(x_s | \vartheta) = \text{pr}_{X_s | \tilde{\Theta}_i^{(k)}, \Pi}(x_s | \vartheta, \pi),$$

which also indicates that the probability of a column is independent of the partition π if the stationary distribution is given. A cluster contains a set of blocks of length L . For simplicity, we discuss computing the predictive

⁴For simplicity, we assume that all clusters have length L .

distribution only for the i th position, $1 \leq i \leq L$, of all blocks in cluster k . Let

$$B_k = \{s + i - 1 \mid \text{for all } s \text{ with } V_s = k\} ,$$

be the set of all such positions. To obtain the predictive distribution, it is necessary to compute

$$\int_{\Delta} \prod_{s \in B_k} \text{pr}_{X|\Theta}(x_s \mid \vartheta) f_{\Theta_i^{(k)}}(\vartheta) d\vartheta ,$$

where evaluating the likelihood $\text{pr}_{X|\Theta}$ requires to sum over all internal vertices of the phylogenetic tree. Solving the integral becomes increasingly difficult the more blocks belong to a cluster and the larger the phylogenetic tree. Instead, we use the approximation

$$q_s(\vartheta) \approx \text{pr}_{X|\Theta}(x_s \mid \vartheta) ,$$

where q_s is a single categorical distribution, as discussed in section 5.5. A similar approximation is also required in the algorithm developed by Sidharthan et al. [2005], however, here it is derived from information theoretic principles. Furthermore, the accuracy of the resulting posterior predictions have been tested in several simulation studies.

A similar computational difficulty arises when the pseudocounts β of the background model are considered as unknown. To avoid integration over β , we draw samples for V_t from the joint posterior distribution V_1, \dots, V_T, β given the data. After each sampling step, the posterior distribution is maximized with respect to β using the RPROP gradient ascent algorithm [Riedmiller and Braun, 1993].

Once a cluster is established it is highly unlikely that the Gibbs sampler will explore partitions where all blocks of the cluster have been moved a few positions left or right [Lawrence et al., 1993, Liu, 1994]. To improve mixing we augment the Gibbs sampler with Metropolis-Hastings (MH) [Metropolis et al., 1953, Hastings, 1970] steps. An MH proposal is obtained by selecting a cluster at random and moving all the elements either left or right. The proposed partition is used as the new sample with the usual MH acceptance probability.

5.5. Phylogenetic likelihood approximation

Let $\{X = x\} = \{X^{(1)} = x^{(1)}, \dots, X^{(M)} = x^{(M)}\}$ denote the observations within a single column of the alignment and $\{\Theta = \vartheta\}$ the stationary distribution specific to this column. Evaluating the likelihood function $\text{pr}_{X|\Theta}(x \mid \vartheta)$ requires to sum over all internal vertices of the phylogenetic tree. Expanding

all terms of this sum results in a polynomial of the form

$$\text{pr}_{X|\Theta}(x|\vartheta) = \sum_i \nu_i \prod_{y \in \mathbb{A}} \vartheta_y^{n_i(y)}$$

with coefficients ν_i and exponents $n_i(y)$, which both depend on the observations $\{X = x\}$ and the phylogenetic tree, as discussed in section 3.3.

To approximate the likelihood function, we interpret it as part of a mixture model with latent variable

$$C \sim \text{Categorical}(\nu)$$

where $\nu = (\nu_i)_i$ are the mixture weights that are assumed w.l.o.g. to sum to one⁵. The variable C indicates from which mixture component an observation is drawn, so that

$$\text{pr}_{X|\Theta,C}(x|\vartheta, i) = \prod_{y \in \mathbb{A}} \vartheta_y^{n_i(y)} .$$

Consider the joint posterior given by

$$f_{\Theta,C|X}(\vartheta, i|x) \propto \text{pr}_{X|\Theta,C}(x|\vartheta, i) f_{\Theta}(\vartheta) \nu_i ,$$

where f_{Θ} is the density of a Dirichlet distribution with pseudocounts $\eta = (\eta_1, \dots, \eta_{|\mathbb{A}|})$. An approximation with a simpler density $g_i(\vartheta) = g(\vartheta) \nu_i$ is obtained by minimizing the Kullback-Leibler divergence

$$D_{\text{KL}}(g_i \| f_{\Theta,C|X}) = \int_{\Delta} \sum_i g_i(\vartheta) \log \frac{g_i(\vartheta)}{f_{\Theta,C|X}(\vartheta, i|x)} d\vartheta ,$$

which is commonly used in variational Bayesian methods to obtain a lower bound on the marginal likelihood. Using variational calculus we find that

$$g(\vartheta) \propto \exp \sum_i \nu_i \log f_{\Theta,C|X}(\vartheta, i|x) ,$$

as discussed in section A.5. By assuming that g is a Dirichlet distribution with pseudocounts $\xi = (\xi_1, \dots, \xi_{|\mathbb{A}|})$ we obtain

$$\xi_y = \eta_y + \sum_i \nu_i n_i(y) .$$

An approximation to the likelihood function $\text{pr}_{X|\Theta}$ is therefore given by

$$q(\vartheta) = \prod_{y \in \mathbb{A}} \vartheta_y^{\sum_i \nu_i n_i(y)} .$$

⁵It is also possible to derive the approximation without the additional latent variable C , however, to arrive at the same result we would need to apply Jensen's inequality.

5.6. Discovery of transcription factor binding sites in *Drosophila*

To demonstrate the proposed method, we consider data from an experiment published by Zinzen et al. [2009]. In this study, the spatio-temporal activity of five transcription factors during early mesoderm development in *Drosophila melanogaster* were analyzed. The Twist (Twi) factor has a global control over the regulatory network. It regulates the expression of Tinman (Tin) and Myocyte enhancing factor 2 (Mef2). While Tin controls the dorsal mesoderm specification, Mef2 is responsible for the muscle differentiation. Tin also initiates the expression of Bagpipe (Bap), which together with Biniou (Bin) regulates the development of the visceral muscle [cf. Zinzen et al., 2009]. Both Bap and Bin have a very similar function and binding preference, which is why we will not consider Bap here.

The first step is to learn the phylogenetic tree from multiple alignments and the question arises as to what the tree should represent. The first option is that it should reflect the average mutation pattern of the genome. In this case, we would learn the gene trees of a sufficiently large set of coding regions and use the average of those trees for the analysis of binding sites. The second option is to infer the tree only from the region that codes for the transcription factor. One might conjecture that the variation between orthologs of the coding region is tightly connected to the variation within ortholog binding sites.

A protein consists of one or more domains that have distinct functions and therefore might be under different selective pressures. If we infer the phylogenetic tree from the protein coding sequence (CDS) of the transcription factor, we may use the entire sequence or select a subsequence that corresponds to one of the protein domains. The CDS associated with the binding domain of the transcription factor is the most natural choice, since we know that it is under heavy selective pressure and can fix our prior assumptions accordingly. In fact, a fundamental assumption of our model is that the binding preference of the transcription factor is identical in all species under consideration. The binding domains of the four transcription factors are shown in Figure 3. All amino acid sequences are highly conserved across the 12 *drosophila* species considered here. In fact, the binding domains of Mef2 show no mutation. This indicates that also the binding preferences of the four transcription factors are conserved.

We learn the phylogenetic tree from the coding regions of the binding domains. Important for the forthcoming analysis of ChIP-seq data is a well justified prior setting for the stationary distributions. We first fix our prior expectations about the levels of conservation within codons and infer the phylogenetic tree. Once the tree is estimated, we may choose prior

parameters for the conservation levels of TFBS *relative* to the prior setting of codons.

A prior setting for the stationary distributions can be obtained by quantifying the codon degeneracy of the genetic code (see also section 4.3). For instance, take the amino acid isoleucine (ile) which is coded by AUA, AUC, and AUU. The first two positions are identical in all three codons, however, the last position can be any of A, C, or U. Quantifying this uncertainty leads to an entropy of zero for the first two positions and $-\log(1/3)$ for the third codon position. For all amino acids we obtain average entropies of 0.095477, 0.031826, and 0.912870 (see Table 1) and we may fix our prior pseudocounts so that the expected entropies match the empirical ones. Accounting for some mutations in the amino acid sequences, we fix our prior pseudocounts at 0.03, 0.01, and 0.6. The estimated median tree is shown in Figure 4a. The posterior uncertainty about the topology is quite large, as can be seen in Figure 4b.

In order to identify TFBS the experimental data from *Drosophila melanogaster* is augmented by aligning ortholog sequences of the 11 related *Drosophila* species. Afterwards, all columns of the alignment that contain gaps in the sequence of *Drosophila melanogaster* are removed. Gaps in the remaining sequences that are longer than six nucleotides are treated as missing data.

For the statistical analysis we have to fix the parameters of the foreground and background model. The transcription factor imposes a selective pressure on the binding sites, which is expected to be slightly less than the selective pressure on the first two codon positions. The background is expected to evolve under less selective pressure than the foreground, but nonetheless it is not expected to evolve fully unconstrained. The prior pseudocounts can be expected to be similar to the ones for the third codon position. Therefore, we fix the foreground prior pseudocounts at $\alpha_x = 0.08$. For the background model we assume that the Dirichlet pseudocounts β are gamma distributed with shape 5.0 and scale 0.2. The remaining model parameters are $\lambda = 0.001$, $a = 1$, and $d = 0$. The length L of the foreground clusters is allowed to range between 6 and 18.

To approximate the maximum a posteriori solution, 10 Markov chains are used, each generating a set of 2000 samples. During the first 500 samples, the chains are heated to increase the chance of reaching a good posterior mode. During the remaining 1500 iterations, the state is optimized at every 20 steps until a local optimum of the posterior distribution is reached. Afterwards, the partition with highest posterior probability is selected as the MAP estimate. The number of patterns and the (unnormalized) posterior values of two chains on the Mef2 data set are shown in Figure 5. The

increased temperature during the first 1000 samples eases the formation of new clusters, as can be observed in the figure.

The results of the analysis are summarized in Figure 6. The motif of Tin almost exactly coincides with the published results by Zinzen et al. [2009]. The analysis of the other data sets revealed several patterns that could be assigned to the transcription factors. For Bin and Mef2 the two predicted motifs are very similar and probably could be joined. The motifs of Twi show much stronger levels of conservation than the published one. From a biological point of view, it seems unlikely that a motif with very low levels of conservation reflects the binding preferences of the transcription factor, since the protein would recognize a large set of different sequences.

AA	codes	first pos.	second pos.	third pos.
Met	AUG	(1, 0, 0, 0)	(0, 0, 0, 1)	(0, 0, 1, 0)
Trp	UGG	(0, 0, 0, 1)	(0, 0, 1, 0)	(0, 0, 1, 0)
Tyr	UAU, UAC	(0, 0, 0, 1)	(1, 0, 0, 0)	(0, 1/2, 0, 1/2)
Phe	UUU, UUC	(0, 0, 0, 1)	(0, 0, 0, 1)	(0, 1/2, 0, 1/2)
Cys	UGU, UGC	(0, 0, 0, 1)	(0, 0, 1, 0)	(0, 1/2, 0, 1/2)
Asn	AAU, AAC	(1, 0, 0, 0)	(1, 0, 0, 0)	(0, 1/2, 0, 1/2)
Asp	GAU, GAC	(0, 0, 1, 0)	(1, 0, 0, 0)	(0, 1/2, 0, 1/2)
Gln	CAA, CAG	(0, 1, 0, 0)	(1, 0, 0, 0)	(1/2, 0, 1/2, 0)
Glu	GAA, GAG	(0, 0, 1, 0)	(1, 0, 0, 0)	(1/2, 0, 1/2, 0)
His	CAU, CAC	(0, 1, 0, 0)	(1, 0, 0, 0)	(0, 1/2, 0, 1/2)
Lys	AAA, AAG	(1, 0, 0, 0)	(1, 0, 0, 0)	(1/2, 0, 1/2, 0)
Ile	AUU, AUC, AUA	(1, 0, 0, 0)	(0, 0, 0, 1)	(1/3, 1/3, 0, 1/3)
Gly	GGU, GGC, GGA, GGG	(0, 0, 1, 0)	(0, 0, 1, 0)	(1/4, 1/4, 1/4, 1/4)
Ala	GCU, GCC, GCA, GCG	(0, 0, 1, 0)	(0, 1, 0, 0)	(1/4, 1/4, 1/4, 1/4)
Val	GUU, GUC, GUA, GUG	(0, 0, 1, 0)	(0, 0, 0, 1)	(1/4, 1/4, 1/4, 1/4)
Thr	ACU, ACC, ACA, ACG	(1, 0, 0, 0)	(0, 1, 0, 0)	(1/4, 1/4, 1/4, 1/4)
Pro	CCU, CCC, CCA, CCG	(0, 1, 0, 0)	(0, 1, 0, 0)	(1/4, 1/4, 1/4, 1/4)
Leu	CUU, CUC, CUA, CUG, UUA, UUG	(0, 2/3, 0, 1/3)	(0, 0, 0, 1)	(1/3, 1/6, 1/3, 1/6)
Ser	UCU, UCC, UCA, UCG, AGU, AGC	(1/3, 0, 0, 2/3)	(0, 2/3, 1/3, 0)	(1/6, 1/3, 1/6, 1/3)
Arg	CGU, CGC, CGA, CGG, AGA, AGG	(1/3, 2/3, 0, 0)	(0, 0, 1, 0)	(1/3, 1/6, 1/3, 1/6)
		0.095477	0.031826	0.912870

TABLE 1. Empirical distributions of nucleotides in the genetic code for the three positions in codons. The last row shows the average entropy of the empirical distributions.

5.6. DISCOVERY OF TRANSCRIPTION FACTOR BINDING SITES IN DROSOPHILA95

```

|175 (Forkhead) |234
DroWil: KPALSYINMIGHAIKESPSGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroGri: KPALSYINMIGHAIKESPSGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroVir: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroMoj: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroPse: KPAMSYINMIGHAIKESPSGKLTSEIYGLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroPer: KPAMSYINMIGHAIKESPSGKLTSEIYGLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroAna: KPALSYINMIGHAIKESPSGKLTSEIYHLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroEre: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroYak: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroMel: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroSim: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
DroSec: KPALSYINMIGHAIKESPTGKLTSEIYAYLQKSYEFFRGPYVGWKNVSRHNLNLSNECFKFLPKGMGVGKPKGKGNWTIDENSAHLFEDEGLRRRPR
* * * * *

```

(A) Bin

```

|3 (MADS box) |58 (MEF2 domain) |86
DroWil: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroVir: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroMoj: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroGri: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroPer: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroPse: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroAna: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroEre: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroYak: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroMel: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroSim: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE
DroSec: RKKIQISRIDERNRQVTFNKRKFGVMKKAYELSVLDCCEIALIIFSSSNKLYQYASTDMRVLKYTEYNEPHESLTNKNIIE

```

(B) Mef2

```

|301 (Homeobox) |360
DroWil: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGS
DroPer: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroPse: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroAna: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroYak: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroEre: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroMel: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroSec: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroSim: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroMoj: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroGri: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
DroVir: KRKPRVLFSAQVLELECFRFRKLYTGAEREIIAQKLNLSATQVKIWFQNRRYKSKRGD
* * *

```

(C) Tin

```

|362 (bHLH) |413
DroWil: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroMoj: NQRVMANVRERQRTQSLNDAFKALQIIPITLPSDKLSKIQLKLATRYIDFL
DroGri: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroVir: NQRVMANVRERQRTQSLNDAFKALQIIPITLPSDKLSKIQLKLATRYIDFL
DroPer: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroPse: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroAna: NQRVMANVRERQRTQSLNDAFKALQIIPITLPSDKLSKIQLKLATRYIDFL
DroYak: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroSim: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroSec: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroEre: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
DroMel: NQRVMANVRERQRTQSLNDAFKSLQIIPITLPSDKLSKIQLKLATRYIDFL
*

```

(D) Twi

FIGURE 3. Multiple sequence alignments of transcription factor binding domains. Stars below the alignment mark columns that contain mutations.

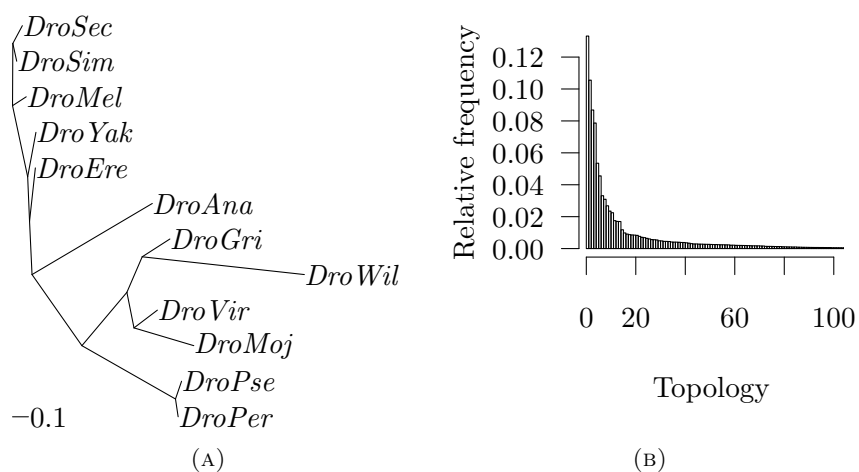


FIGURE 4. (A) Phylogenetic tree estimated from the coding regions of the binding domains of the four transcription factors Bin, Mef2, Tin, and Twi. (B) Relative frequencies of the first 100 most abundant topologies in the posterior samples.

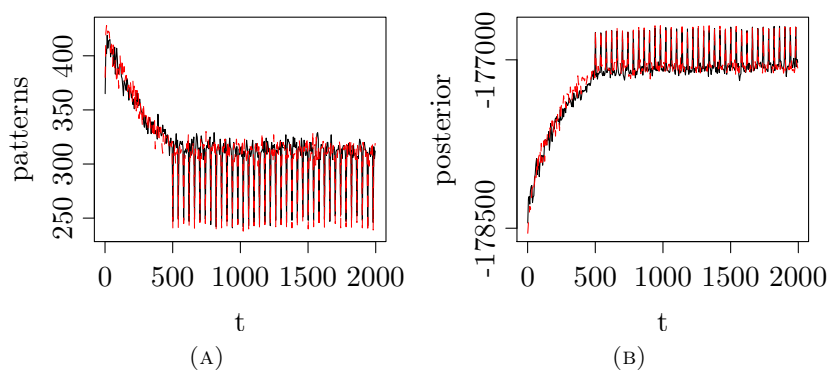


FIGURE 5. Number of patterns (A) and (unnormalized) posterior values from two out of ten MCMC samplers on the Mef2 data set.

5.6. DISCOVERY OF TRANSCRIPTION FACTOR BINDING SITES IN DROSOPHILA97

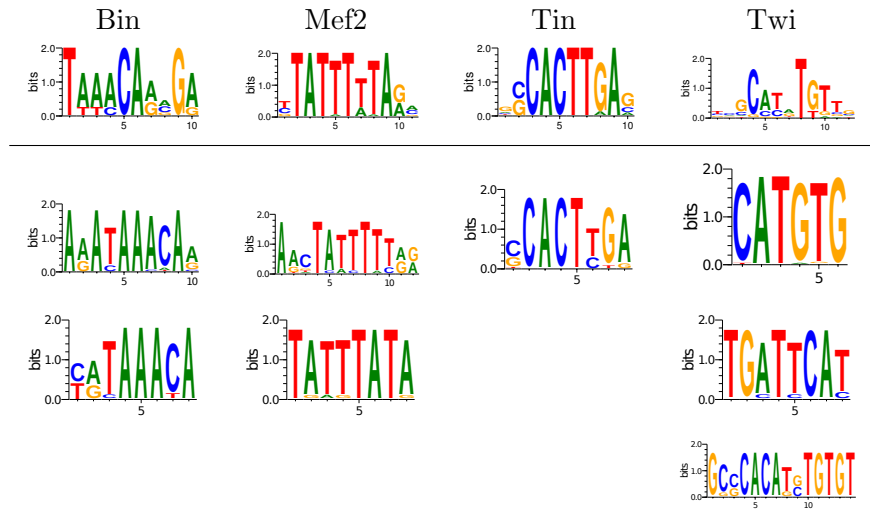


FIGURE 6. Motif predictions for the four transcription factors Bin, Mef2, Tin, and Twi. The first row shows the predictions by Zinzen et al. [2009] who used RSAT, a method developed by Thomas-Chollier et al. [2008]. The next three rows show the predictions with our model. Multiple motifs were found for Bin, Mef2, and Twi.

CHAPTER 6

Conclusion

A method for the identification of transcription factor binding sites (TFBS) from enriched data sets was presented. The inferential step relies on a clear and well justified statistical model, which consists of two main components. First, incorporating data from related species requires a model for the evolution of functional and non-functional sites. Based on a phylogenetic tree, predictions for the selective pressures at each site can be made. Second, the occurrences of repeated patterns (motifs) in the data has to be modeled. By the experimental design, it is known that the data is enriched with binding sites of the transcription factor. However, also other motifs do appear, but are assumed to be less frequent. Because of their *rich-get-richer* property, the Dirichlet and related processes seem to reflect this prior knowledge well. Unfortunately, besides the subjective motivation for these processes, there exist no empirical findings to justify this choice.

In Bayesian inference it is essential to form prior beliefs that are based on a clear line of thought. Predictions might otherwise become arbitrary, especially when the statistical model is complex. The phylogenetic tree is one such parameter, its sound estimation as well as a consistent use of the substitution model is therefore of high importance. Motivated by statistical decision theory, Bayesian estimates are obtained as summaries of the posterior distribution. Common choices include the mean or median, which in Euclidean space are well defined and easy to compute. However, the space of phylogenetic trees is not a simple linear space and its rich geometry complicates the computation of posterior summaries. Nevertheless, the mean and median can be defined as minimizers of appropriate loss functions and recently developed methods allow their approximate computation. In a simulation study it was shown that the so obtained Bayesian estimates provide a more faithful summary than the commonly used majority-rule consensus tree. Unfortunately, little is known about the rate of convergence of the approximation method used for computing summaries in tree space. The results need therefore be treated with care and should be checked thoroughly.

For the identification of TFBS, a substitution model is used which significantly deviates from the ones commonly used to infer phylogenetic trees. In

the model used here, each alignment column equipped with its own stationary distribution. This leads to a different interpretation of the stationary distribution as the selective pressure acting on the respective position in the alignment. Despite the complexity of the model, it was shown here how phylogenetic trees can be efficiently inferred using a combination of analytical and numerical (MCMC) integration. The resulting estimate enables us to quantify subjective expectations of functional and non-functional regions.

The statistical inference of TFBS is based on a model for data sets that are enriched with an a priori unknown number of different motifs. Usually one is interested in the binding sites of a certain protein, and it is assumed that those sites are most abundant in the data. Obtaining posterior predictions for the positions of binding sites is computationally very challenging. To evaluate the (unnormalized) posterior distribution, an approximation of the phylogenetic information was presented that is motivated by variational Bayesian methods. Unfortunately, the dimensionality of the model's parameter space is too high to compute summaries other than the maximum a posteriori (MAP) estimate. Hence, predictions are computed by using MCMC methods augmented with local optimizations to search the posterior distribution for good local maxima.

The method was evaluated on a data set by Zinzen et al. [2009] and it was demonstrated how the setting of prior parameters can be motivated. For some transcription factors, the predictions were in line with previously published results. However, for other transcription factors several highly similar motifs were found, which might indicate that the binding preferences of the protein are best described by a mixture of several motifs. Such mixture models might constitute a simple alternative to Markov models for describing binding preferences [e.g. Benos et al., 2002, Barash et al., 2003, Tomovic and Oakeley, 2007, Zhou and Liu, 2004].

Several extensions of the proposed method can be thought of. First, the attribution of several motifs to a single transcription factor could be made formal. This could be a powerful alternative to model dependencies in the binding preferences of transcription factors. Second, eukaryotic genes are often regulated by several transcription factors that have binding sites organized in cis-regulatory modules (CRMs). Modeling the known structure of regulatory modules can help to identify binding sites [Zhou and Wong, 2004, Siddharthan, 2008, Vandenbon et al., 2008, Lemnian et al., 2013]. So far, this knowledge is not exploited in the present method. Third, a crucial step of the analysis is the generation of multiple sequence alignments. Once a phylogenetic tree is available, it can be used to generate the alignments (for instance with PRANK, Löytynoja and Goldman [2005, 2008]). However, the method relies on a different substitution model and a clear justification for

the parameter settings is not available. Finally, an extensive quantitative comparison to other existing methods should be made, which is inevitable to show the advantages of the method presented here.

Software

The methods presented here were implemented in a C++/Python software package, which consists of almost 30000 lines of code. The software is freely available under

<https://github.com/pbenner/tfbayes>.

The computation of means and medians of phylogenetic trees requires methods from linear programming, for which the *GNU Linear Programming Kit*¹ is used. Parsing phylogenetic trees from file is best achieved with a context free grammar using *GNU Flex*² and *GNU Bison*³. Multiple sequence alignments are parsed from Mef files with *Biopython*⁴. Motifs of transcription factor binding sites are visualized with *Weblogo*⁵. Sampling of phylogenetic trees and the identification of transcription factor binding sites rely on data structures, special functions, and extensive threading methods provided by the *Boost library*⁶, which is also used to build interfaces between C++ and Python.

¹<http://www.gnu.org/software/glpk/>

²<http://flex.sourceforge.net/>

³<http://www.gnu.org/software/bison/>

⁴<http://www.biopython.org>

⁵<http://weblogo.berkeley.edu>

⁶<http://www.boost.org/>

APPENDIX A

Introduction to probability and statistics

“Predictions [...] can only be probable. However solidly founded a prediction may appear to us, we are never absolutely sure that experiment will not prove it to be baseless if we set to work to verify it.” Poincaré [1905].

A.1. Essentials of probability theory

Statistics as the fundamental method for reaching scientific conclusions and making predictions from experimental observations should be based on a sound and well understood theory of probability. Here we introduce basic notions of mathematical probability theory. There exists a vast literature on the subject, see for instance Bauer [2001], Lieb and Loss [2001], and of course Bourbaki and Berberian [2004a,b] for measure and integration theory. Modern probability theory is nicely explained in Bauer [1996], Kallenberg [2002], Klenke [2008], or Loève [1977, 1978].

Since the seminal work of Kolmogoroff [1933], probability theory is inextricably tied to *measure theory*. Assigning a length, area, or volume to a piece of Euclidean space is essentially the same as the assignment of probabilities. Both measures need to satisfy certain requirements or axioms that prevent inconsistencies. The *Banach-Tarski paradox* has shown that σ -additivity is an essential requirement for measures on Euclidean spaces of dimension three or more. The adoption of this axiom however leads to other inconsistencies, as proven by the Italian mathematician Giuseppe Vitali. His argument is based on a family of sets that are nowadays known as *Vitali sets*. It is due to Émile Borel and Henri Lebesgue that the modern definition of a measure is restricted to a class of sets, the *measurable sets*, to avoid any inconsistencies. A more in-depth treatise on the history of measure theory can be found in Elstrodt [2004].

We begin the discussion by recalling some basic notions from topology. Intuitively speaking, a topology can be thought of as a formal description of the collection of neighborhoods of a space. This description is however fully detached from any notions of distance.

DEFINITION A.1.1 (Topology). Let Ω be a set and τ a collection of subsets of Ω . τ is called a *topology* if:

- $\emptyset, \Omega \in \tau$.
- Let $\{A_i \mid A_i \in \tau\}_{i \in I}$ be a collection of sets from τ , then $\bigcup_{i \in I} A_i$ is an element of τ (closed under unions).
- Let $A, B \in \tau$, then $A \cap B$ is an element of τ (closed under countable intersections).

The tuple (Ω, τ) is also called a *topological space*. The sets of a topology τ are called the *open sets* of Ω and a set $A \subset \Omega$ is *closed* if the complement A^c is open. However, a subset of Ω might be neither open nor closed. The motivation behind this definition stems from the properties of open sets in metric spaces. A subset $N \subset \Omega$ is called a *neighborhood* of a point $x \in \Omega$ if there exists an open set A such that $x \in A \subset N$. Simple examples of topologies for a set of points $\Omega = \{1, \dots, n\}$ are

- $\tau = \{\emptyset, \Omega\}$, called the *trivial topology* of Ω , and
- $\tau = 2^\Omega$, which is the *discrete topology*.

There is no concept of distance in topological spaces, since the space is not necessarily equipped with a metric, but other related notions exist. For instance, relative to a given topology concepts such as *continuity* of a function, *convergence* of a sequence of points, or *connectedness* of the space, can be defined. Let (Ω, τ) be a topological space, then Ω is connected if it is not the union of two disjoint non-empty sets $A_1, A_2 \in \tau$. Furthermore, a point $x \in \Omega$ is called *isolated* if $\{x\}$ is an open set. In particular, a topological space is called *discrete* if every point in Ω is isolated. For practical reasons, it is not feasible to specify the topology τ directly. The following definitions provide us with the means to generate a topology from a smaller collection of open sets.

DEFINITION A.1.2 (Topological basis). Let (Ω, τ) be a topological space. A collection $\mathcal{B} \subset \tau$ is called a *basis* of the topology if every open set $A \in \tau$ can be realized as the union of some collection $B \subset \mathcal{B}$.

In particular, such a basis may consist of all open balls defined by a given metric.

DEFINITION A.1.3 (Metric). A metric on a set Ω is a function $d : \Omega \times \Omega \rightarrow [0, \infty]$, which satisfies

- $d(x, y) = 0$ if and only if $x = y$,
- $d(x, y) = d(y, x)$ for any $x, y \in \Omega$ (symmetry),
- $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

As a simple example take the Euclidean space \mathbb{R}^n with metric d and let $B_r(x) = \{y \in \mathbb{R}^n \mid d(x, y) < r\}$ denote the open ball of radius r around x .

The set of open balls

$$\mathcal{B} = \{B_r(x) \mid x \in \mathbb{R}^n, r > 0\} ,$$

defines a basis of the standard topology on Euclidean space. For technical reasons, it is often required to have a *countable basis*. The standard Euclidean topology is also defined through the basis

$$\mathcal{B} = \{B_r(x) \mid x \in \mathbb{Q}^n, 0 < r \in \mathbb{Q}\} .$$

Any topological space with a countable basis is *separable*, i.e. it contains a countable dense subset. The topology generated by a metric is also called the *metric topology*. In general, there are many metrics that can be used to generate a given topology. If we fix the topology, then any metric that induces this topology is said to be *compatible*. In probability theory, it is often not required to fix a certain metric, but rather to ensure that the space allows to define a metric if needed.

DEFINITION A.1.4 (Polish space). A topological space (Ω, τ) is called *Polish space* if there exists

- a countable basis that generates τ , and
- a compatible metric d such that (Ω, d) is a complete metric space (i.e. every Cauchy sequence converges to a point in Ω).

DEFINITION A.1.5 (σ -algebra). Let Ω be a set and Σ a non-empty collection of subsets of Ω . Σ is called a σ -*algebra* of Ω if it satisfies the following conditions:

- If $A \in \Sigma$, then the complement A^c of A is also an element of Σ (closed under complementation).
- If A_1, A_2, \dots is a countable family of sets with $A_i \in \Sigma$, then the union $\bigcup_{i=1}^{\infty} A_i$ is also an element of Σ (closed under countable unions).

The pair (Ω, Σ) is also called a *measurable space*. It follows from the definition that Ω and \emptyset are also elements of every σ -algebra Σ . Furthermore, Σ is also closed under countable intersections. If Σ is only closed under finite unions, then it is called an *algebra of sets*.

DEFINITION A.1.6. Let Ω be a set and S be a family of subsets of Ω . $\sigma(S)$ is called the σ -algebra *generated* by S if it is the intersection of all σ -algebras that contain S .

A σ -algebra generated by a set S is therefore the smallest algebra that contains S . The most prominent example is the Borel σ -algebra \mathcal{B} , which is generated by the topology, i.e. the open subsets of Ω . In particular for the

Euclidean space \mathbb{R}^n the Borel σ -algebra is a strict subset of the powerset, that can be measured without introducing any inconsistencies.

DEFINITION A.1.7 (Measure space). Let (Ω, Σ) be a measurable space. A function $\mu : \Sigma \rightarrow [0, \infty]$ is called a *measure* if $\mu(\emptyset) = 0$ and if it is countable additive (σ -additive), i.e. for any sequence A_1, A_2, \dots of disjoint sets in Σ

$$\mu \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i).$$

The triple (Ω, Σ, μ) is called a *measure space*.

The requirement of countable additivity is a stronger requirement than finite additivity, i.e. a measure which is countable additive is also finite additive (we might choose $A_i = \emptyset$), but not vice versa. A simple example of a measure for a discrete space is given by

$$\mu(A) = \begin{cases} |A| & \text{if } A \text{ is finite,} \\ \infty & \text{otherwise,} \end{cases}$$

which is called the *counting measure*. A measure μ is said to be *finite* if $\mu(\Omega) \in [0, \infty)$. Furthermore, μ is σ -*finite* if there exists a sequence A_1, A_2, \dots of sets in Σ , such that $\mu(A_i)$ is finite and

$$\mu(\Omega) = \sum_{i=1}^{\infty} \mu(A_i).$$

For instance, take the Lebesgue measure $\mu([a, b]) = |b - a|$ on the real line. It is easy to see that this measure is σ -finite but not finite.

DEFINITION A.1.8 (Measurable function). Let (Ω, Σ) and (S, \mathcal{S}) be two measurable spaces. A function $f : \Omega \rightarrow S$ is called *measurable* (or more precisely Σ/\mathcal{S} -measurable) if for every $B \in \mathcal{S}$ the preimage $f^{-1}(B) = \{a \in \Omega \mid f(a) \in B\}$ is an element of Σ . We then also write $f : (\Omega, \Sigma) \rightarrow (S, \mathcal{S})$.

Every measurable function $f : (\Omega, \Sigma) \rightarrow (S, \mathcal{S})$ induces a sub- σ -algebra $\sigma(f) \subseteq \Sigma$ on Ω , i.e.

$$\sigma(f) = \{f^{-1}(B) \mid B \in \mathcal{S}\} \subseteq \Sigma.$$

A popular illustration of a non-measurable function is provided by

EXAMPLE A.1.1. Assume that Ω represents the outcome of rolling a die, i.e. $\Omega = \{\omega_1, \omega_2, \dots, \omega_6\}$ and we are only told whether the outcome is an even or odd number, so that $\Sigma = \{\emptyset, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}, \Omega\}$. A function that maps for instance ω_1, ω_2 , and ω_3 to -1 and all other outcomes to 1 violates the structure of Σ and is not measurable.

Let us now consider a measurable space (Ω, Σ) with two measures μ and ν . The measure ν is said to be dominated by μ , denoted $\nu \ll \mu$, if for every $A \in \Sigma$ and $\mu(A) = 0$ also $\nu(A) = 0$.

THEOREM A.1.9 (Radon-Nikodym). *Let (Ω, Σ) be a measurable space and $\mu, \nu : \Sigma \rightarrow [0, \infty]$ two σ -finite measures with $\nu \ll \mu$, then there exists a function $f : \Omega \rightarrow [0, \infty)$ such that*

$$\nu(A) = \int_A f(x) d\mu(x),$$

for all $A \in \Sigma$.

The function f is also called the *Radon-Nikodym derivative*, denoted by $d\nu/d\mu$. In particular, for Euclidean spaces we may choose μ to be the Lebesgue measure and thereby obtain a simple representation of most measures by selecting a suitable Radon-Nikodym derivative.

We will now turn our attention to the notion of a *probability space*, which is a normed measure space (Ω, Σ, P) such that $P(\Omega) = 1$. In this context, Ω is also called the *sample space* and each $\omega \in \Omega$ is called an *outcome*. Furthermore, the elements of Σ are referred to as *events*. We use P to assign a probability to each event and by requiring that P is a measure we ensure that this assignment is consistent.

REMARK A.1.1. It follows from the definition of P that probabilities are assumed to be σ -additive. In its original definition by Kolmogoroff [1933], it was only required that the function P satisfies $\lim_{i \rightarrow \infty} P(A_i) = 0$ for any countable sequence

$$A_1 \supset A_2 \supset \cdots \supset A_i \supset \cdots$$

with $A_i \in \Sigma$ and

$$\bigcap_{i=1}^{\infty} A_i = \emptyset.$$

However, since P is defined on a σ -algebra, it follows from this axiom that P is also σ -additive.

The definition of probability measures on infinite spaces should always be treated with caution. As Kolmogoroff [1933] pointed out: “Bei einer Beschreibung irgendwelcher wirklich beobachtbarer zufälliger Prozesse kann man nur endliche Wahrscheinlichkeitsfelder erhalten. Unendliche Wahrscheinlichkeitsfelder erscheinen nur als idealisierte Schemata reeller zufälliger Prozesse.” He also emphasizes that the axiom that leads to σ -additivity is hard to justify from the empirical perspective but has been proven to be practicable.

Statistical inference is concerned with learning from experimental observations. In probability theory, an experiment is formalized as a collection of events π , which partitions the sample space Ω . The outcome of an experiment tells us which event of π occurred. This concept is implemented in probability theory using the following

DEFINITION A.1.10 (Random variable). Let (Ω, Σ, P) be a probability space and (X, \mathcal{X}) a measurable space. A function $X : (\Omega, \Sigma) \rightarrow (X, \mathcal{X})$ is called a *random variable*.¹

We will refer to the elements of X as the *experimental outcomes*. A random variable is used to define the partition π of Ω . That is, for each $x \in X$ the preimage $X^{-1}(x)$ is an element of π . In modern terms we say that X induces a sub- σ -algebra $\sigma(X)$ of Σ . The granularity of $\sigma(X)$ relates to how much we can learn from the experiment modeled by X .

EXAMPLE A.1.2. Let $\Omega = \{\omega_1, \dots, \omega_6\}$ represent the possible outcomes of rolling a die. Assume that we do not observe the die directly, but that some other person is rolling the die and afterwards tells us whether the outcome was an even or odd number. This experiment could be modeled by the random variable

$$X(\omega) = \begin{cases} x_1 & \text{if } \omega \in \{\omega_1, \omega_3, \omega_5\}, \\ x_2 & \text{otherwise.} \end{cases}$$

Sometimes the notion of a random variable is restricted to measurable functions with real codomain. Our applications require a more general definition, where the codomain is given by a metric space with the induced Borel σ -algebra. A random variable X induces a probability measure

$$\text{pr}_X(B) = P(X^{-1}(B)) = P(\{a \in \Omega \mid X(a) \in B\})$$

on its codomain (i.e. the *push-forward* of P by X), which is called the *distribution* of X . We will also use the notation $X \sim \text{pr}_X$ to express that X has distribution pr_X . For applications it is common to define random variables indirectly through their distributions on a given measurable space, whereby the definition of a specific sample space becomes dispensable. Therefore, we may impose necessary restrictions directly on the codomain of a random variable, for instance, the requirement that the space is Polish. Continuous distributions are often defined through a *density function*, which is usually given as the Radon-Nikodym derivative of the distribution with respect to the Lebesgue measure.

¹Many authors [e.g. Kallenberg, 2002] use the term random variable only when $X = \mathbb{R}$. Otherwise, X is called a *random vector* if $X = \mathbb{R}^n$ or *random element* for more general spaces. We will not make this distinction here.

Several laws follow directly from the formal definition of probability measures. To summarize some of them, let X be a random variable that takes values on some measurable space $(\mathbf{X}, \mathcal{X})$. Then

- $\text{pr}_X(A) \leq \text{pr}_X(B)$ for all $A \subseteq B \in \mathcal{X}$ (monotonicity),
- $\text{pr}_X(A \cup B) = \text{pr}_X(A) + \text{pr}_X(B) - \text{pr}_X(A \cap B)$ for all $A, B \in \mathcal{X}$ (addition law),
- $\text{pr}_X(\Omega \setminus A) = 1 - \text{pr}_X(A)$ for all $A \in \mathcal{X}$ (inclusion-exclusion principle).

Let (Ω, Σ, P) be a probability space and $\{(\mathbf{X}_i, \mathcal{X}_i)\}_{i=1}^n$ a collection of measurable spaces. We define n random variables X_1, \dots, X_n such that $X_i : (\Omega, \Sigma) \rightarrow (\mathbf{X}_i, \mathcal{X}_i)$ for $i = 1, \dots, n$, and let

$$\bar{X}(\omega) = (X_1(\omega), \dots, X_n(\omega)).$$

The distribution induced by \bar{X} is called the *joint distribution* of X_1, \dots, X_n . It takes a particularly simple form if the n random variables are *independent*, in which case it holds that

$$\text{pr}_{\bar{X}}(A_1, \dots, A_n) = \prod_{i=1}^n \text{pr}_{X_i}(A_i)$$

for any collection of sets $\{A_i \in \mathcal{X}_i\}_{i=1}^n$.

A.2. Conditional probability distributions

For Bayesian statistics it is essential to update probabilities in light of new information. Within the context of the probability framework, we would like to know the probability of an outcome ω being an element of a set A , given the information that ω is an element of some other set B . The concept on which this update of probabilities rests is the *conditional distribution*. Modern probability theory defines conditional distributions based on the notion of conditioning on a random variable. As discussed earlier, random variables are used to partition the sample space and thereby induce a sub- σ -algebra. Conditioning on a random variable X is therefore equivalent to conditioning on $\sigma(X)$. Once the value of a random variable is observed, we know to which element of the partition the outcome belongs, i.e. which event occurred. Defining and showing the existence of a conditional distribution on a discrete sample space is quite simple. However, it becomes much harder for larger spaces when there is a nonnegligible set of events that must have probability zero, as will be discussed in the following. Before introducing the modern definition of conditional probability distributions, it is quite instructive to establish the early definition by Kolmogoroff [1933] first.

A naïve way of defining a conditional measure on a probability space (Ω, Σ, P) is given by

$$Q(A|B) = \frac{P(A \cap B)}{P(B)}$$

for any $A, B \in \Sigma$ such that $P(B) > 0$. Obviously, $Q(A|B) = 0$ if $A \cap B = \emptyset$. It is also easy to see that $A \mapsto Q(A|B)$ is a probability measure, since for any $B \in \Sigma$

- $Q(A|B) \geq 0$ for all $A \in \Sigma$,
- $Q(\Omega|B) = 1$, and
- $Q(\bigcup_{i=1}^{\infty} A_i|B) = P(\bigcup_{i=1}^{\infty} A_i \cap B)/P(B) = \sum_{i=1}^{\infty} Q(A_i|B)$,

where A_1, A_2, \dots is a countable collection of pairwise disjoint events.

For many applications this naïve definition is not eligible, as it is also necessary to condition on sets of probability zero. A first step towards a proper definition is to seek any measure $Q : \Sigma \times \Sigma \rightarrow [0, 1]$ such that

$$P(A \cap B) = Q(A|B)P(B)$$

for all $A, B \in \Sigma$. We observe that $P(A \cap B)$ is zero whenever $P(B)$ equals zero, i.e. $B \mapsto P(A \cap B)$ is dominated by P . Furthermore, assume that P is σ -finite. Hence, by the Radon-Nikodym theorem there exists a function $q : \Sigma \times \Omega \rightarrow [0, \infty)$ such that

$$P(A \cap B) = \int_B q(A|\omega) dP(\omega)$$

for all $B \in \Sigma$. The function q is uniquely defined up to a P -null set and it is easy to see that one possible solution is $q(A|\omega) = 1_A(\omega)$ for all $A \in \Sigma$. The problem of conditioning on a single outcome $\omega \in \Omega$ is therefore easily solved.

Following this line of thought, we extend the above definition to condition on the values of a random variable $Y : (\Omega, \Sigma) \rightarrow (S, \mathcal{S})$. By fixing some $A \in \Sigma$, we know that there exists a function $q : \Sigma \times S \rightarrow [0, \infty)$ such that

$$\begin{aligned} P(A \cap Y^{-1}(B)) &= \int_{Y^{-1}(B)} q(A|Y(\omega)) dP(\omega) \\ (9) \qquad \qquad \qquad &= \int_B q(A|y) d\text{pr}_Y(y) \end{aligned}$$

for all $B \in \mathcal{S}$. The function q might therefore be used to condition on any (measurable) subset of Ω by selecting a suitable Y . However, it should be emphasized that this definition is not constructive.

EXAMPLE A.2.1. Let (Ω, Σ, P) be a probability space with $\Omega = \{a, b, c, d, e, f\}$, $\Sigma = 2^\Omega$, and P the uniform measure on Ω . Furthermore, let

$Y : \Omega \rightarrow \mathbb{N}$ be a random variable defined as

$$Y = \begin{pmatrix} a & b & c & d & e & f \\ 4 & 4 & 2 & 2 & 3 & 7 \end{pmatrix},$$

so that $\sigma(Y) = \sigma\{\{a, b\}, \{c, d\}, \{e\}, \{f\}\}$. The conditional measure q has to satisfy

$$P(A \cap Y^{-1}(B)) = \sum_{y \in B} q(A | y) \text{pr}_Y(y),$$

for all $A \in \Sigma$ and $B \in \mathcal{S}$. For instance, we have

$$P(\{b, c\} \cap Y^{-1}(\{4, 7\})) = 1/2 \cdot 1/3 + 0 \cdot 1/6.$$

While this approach allows an intuitive understanding of how a conditional distribution can be obtained, a disadvantage is that for each $A \in \Sigma$ we might get a different q and therefore may not conclude that $A \mapsto q(A | y)$ is a probability measure for every $y \in S$. Nevertheless, Kolmogoroff [1933] was able to show the following

THEOREM A.2.1. *Any function q as defined in Equation 9 satisfies almost surely*

$$0 \leq q(A | y) \leq 1.$$

Furthermore, let A_1, A_2, \dots be a sequence of pairwise disjoint events such that $A = \bigcup_{i=1}^{\infty} A_i$, then almost surely

$$q(A | y) = \sum_{i=1}^{\infty} q(A_i | y).$$

The term almost surely means that for a given $A \in \Sigma$ there exists a $B \in \mathcal{S}$ such that $\text{pr}_Y(B) = 0$ and the statements of the theorem hold if $y \notin B$. More recent results on conditional distributions rely on a concept called *disintegration* of measures [Chang and Pollard, 1997]. Under stronger assumptions on the probability space it is possible to show the existence of a function q such that $A \mapsto q(A | y)$ is in fact a probability measure for almost all $y \in S$. For our purposes it suffices to say that a conditional distribution always exists and we may simply work with the following

DEFINITION A.2.2 (Conditional probability distribution). Let (Ω, Σ, P) be a probability space, (S_1, \mathcal{S}_1) and (S_2, \mathcal{S}_2) measurable spaces. Furthermore, let $X : (\Omega, \Sigma) \rightarrow (S_1, \mathcal{S}_1)$ and $Y : (\Omega, \Sigma) \rightarrow (S_2, \mathcal{S}_2)$ be random variables. The conditional measure $q : \Sigma \times S_2 \rightarrow [0, 1]$ is any function such that

- $y \mapsto q(A | y)$ is measurable for all $A \in \Sigma$,
- $A \mapsto q(A | y)$ is a probability measure for almost all $y \in S_2$, and
- $P(A \cap X^{-1}(B)) = \int_B q(A | y) d\text{pr}_Y(y)$ for all $A \in \Sigma$ and $B \in \mathcal{S}_2$.

The *conditional distribution* $\text{pr}_{X|Y} : \mathcal{S}_1 \times \mathcal{S}_2 \rightarrow [0, 1]$ is defined as the push-forward of q by X .

A distribution satisfying this definition is often called a *regular conditional distribution* to distinguish it from the early approach by Kolmogoroff [1933].

In statistics, we are often interested in the conditional density $f_{X|Y}$ as a solution to the equation

$$\int_A \int_B f_{X,Y}(a,b) \text{d}a \text{d}b = \int_A \int_B f_{X|Y}(a|b) f_Y(b) \text{d}a \text{d}b$$

for all $A \in \mathcal{S}_1$ and $B \in \mathcal{S}_2$ and one should remember that

$$f_{X|Y}(a|b) = \frac{f_{X,Y}(a,b)}{f_Y(b)}$$

is not the only possible solution (when $f_Y(b) > 0$ for all $b \in \mathcal{Y}$).

A.3. Summarizing a distribution

Probability distributions can be very complex and in high dimensions difficult to inspect and visualize. Summarizing a distribution therefore is important, especially in statistics when a decision has to be reached. In Euclidean space, many characteristics of a distribution can be defined, such as the *moments* and *cumulants*. Only some of those characteristics generalize to more abstract metric spaces and we will restrict the discussion to those.

DEFINITION A.3.1 (Expectation and variance, cf. Sturm [2003]). Let X be a random variable in a metric space (X, d) and let $f : X \rightarrow X$ be a measurable function. The expectation of X with respect to f is defined as

$$\mathbb{E} f(X) = \arg \inf_{y \in X} \int_X d(y, f(x))^2 \text{dpr}_X(x).$$

Furthermore, the variance of X with respect to f is defined as

$$\text{var} f(X) = \inf_{y \in X} \int_X d(y, f(x))^2 \text{dpr}_X(x).$$

The expectation $\mathbb{E} X$ is also called the Fréchet mean or the *barycenter* of the distribution pr_X . In particular in Hadamard spaces, which are metric spaces of nonpositive curvature, d^2 is strongly convex and therefore the expectation exists and is unique. When considering the Euclidean space \mathbb{R}^n , the expectation and variance take a much simpler form. For instance, the expectation simplifies to

$$\mathbb{E} f(X) = \int_X f(x) \text{dpr}_X(x),$$

whereas the variance can be written as

$$\text{var}(X) = \mathbb{E}(X - \mathbb{E} X)^2 = \int_{\mathbb{X}} (x - \mathbb{E} f(X))^2 \text{dpr}_X(x).$$

In Euclidean space, probability distributions can be characterized by their *moments*. The p -th moment μ_p is defined as

$$\mu_p(X) = \mathbb{E} X^p = \int_{\mathbb{X}} x^p \text{dpr}_X(x)$$

Given a sequence of reals c_1, c_2, \dots , the *moment problem* asks for the existence and uniqueness of a random variable X such that $\mu_1(X) = c_1, \mu_2(X) = c_2, \dots$, which is a well-studied problem and necessary and sufficient conditions have been found [cf. e.g. Shohat and Tamarkin, 1943].

In this form, the definition of moments does not permit a generalization to metric spaces. Only the p -th *absolute central moment* $|\mu|_p(X) = \mathbb{E} |X - \mathbb{E} X|^p$ may be written as

$$|\mu|_p(X) = \int_{\mathbb{X}} d(\mathbb{E} X, x)^p \text{dpr}_X(x).$$

However, by taking the absolute value, much information about the distribution is lost. Instead, it seems more natural to generalize the expectation and variance directly, which leads to

DEFINITION A.3.2 (p -expectation and p -variance). For $p \in [1, \infty)$ and a random variable X on a metric space the p -*expectation* $\mathbb{E}_p f(X)$ (or p -*mean*) with respect to a function $f : \mathbb{X} \rightarrow \mathbb{X}$ is defined as

$$\mathbb{E}_p f(X) = \arg \inf_{y \in \mathbb{X}} \int_{\mathbb{X}} d(y, f(x))^p \text{dpr}_X(x).$$

Furthermore, the p -*variance* of X with respect to f is given by

$$\text{var}_p f(X) = \inf_{y \in \mathbb{X}} \int_{\mathbb{X}} d(y, f(x))^p \text{dpr}_X(x).$$

This definition is particularly important in Bayesian statistics and statistical decision theory. While the p -variances can be seen to characterize the widths of the distribution, the p -expectations measure the asymmetries. We also call $\mathbb{E}_1 X$ the *geometric median* and $\mathbb{E}_2 X$ the *Fréchet mean*. As will be discussed later, the median and mean are well defined in CAT(0) spaces since the distance function in spaces of nonpositive curvature is convex. Whether the sequence $(\mathbb{E}_1 X, \text{var}_1 X), (\mathbb{E}_2 X, \text{var}_2 X), \dots$ uniquely characterizes the distribution of X seems to be an open problem.

For instance, assume that X is a real random variable with normal distribution. Due to the symmetry of the distribution, all p -expectations are located at the mean of the distribution. If instead we consider a skewed gamma distribution, see Figure 1, we observe that the p -expectations move

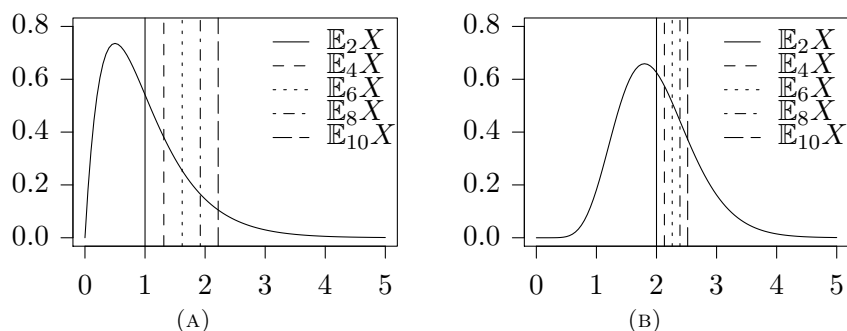


FIGURE 1. Density function of two gamma distributions with different shape k and scale θ parameters. The p -expectations for even p are shown as vertical lines. (A) Shape $k = 2$ and scale $\theta = 2$. (B) Shape $k = 10$ and scale $\theta = 5$.

with increasing p in the direction of the tail, while the distance between successive expectations decreases very slowly.

A.4. Markov chain Monte Carlo methods

The actual computation of a summary of some distribution μ can in practice be very difficult. For instance, assume that

$$\mu(\cdot) = \frac{1}{Z} \pi(\cdot),$$

where $Z = \int d\pi$ is the normalization constant of μ . To evaluate μ , we have to compute the normalization constant Z , which in most but very simple cases is infeasible. A summary of μ can therefore only be obtained by circumventing direct evaluations of μ .

Markov chain Monte Carlo (MCMC) methods can be used to construct an *ergodic* Markov chain that has μ as its invariant distribution. Simulating the Markov chain *ad infinitum* is therefore equivalent to sampling from μ . Let $X^{(0)}, \dots, X^{(t)}, \dots$ denote the random variables of the Markov chain, where each $X^{(t)}$ takes values in $\mathsf{X} = \text{dom}(\mu)$. For *any* initial value $X^{(0)}$ and all integrable functions $g : \mathsf{X} \rightarrow \mathbb{R}$ we have by the *ergodic theorem* that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T g(X^{(t)}) \rightarrow \int_{\mathsf{X}} g(x) d\mu(x),$$

converges almost surely (for suitable X).² Intuitively speaking, μ is approximated by an infinite mixture of Dirac distributions located at $\{X^{(t)} =$

²The chain converges almost surely for *every* starting value if it is Harris recurrent [cf. e.g. Robert and Casella, 1999]. However, one should not forget that any MCMC

$x^{(t)}\}_{t=1}^{\infty}$. This result is used by MCMC methods to approximate summaries of μ . The following discussion provides a highly simplified review of MCMC methods. Further details can be found in Robert and Casella [1999].

The distribution of a (time-homogeneous) Markov chain $M = (X^{(t)})$ is defined through its initial distribution $\mu_0 = \text{pr}_{X^{(0)}}$ and a transition kernel

$$\kappa(x, Y) = \text{pr}_{X^{(t+1)} | X^{(t)}}(Y | x),$$

for all $t \in \mathbb{N}$, $x \in \mathsf{X}$, and $Y \in \mathcal{B}(\mathsf{X})$. A σ -finite measure μ is called an *invariant* or *stationary* distribution of M if it satisfies

$$\mu(Y) = \int_{\mathsf{X}} \kappa(x, Y) d\mu(x),$$

for all $Y \in \mathcal{B}(\mathsf{X})$.

EXAMPLE A.4.1. Let M be a time-homogeneous Markov chain on a discrete space X and assume that M has a single stationary distribution μ . Furthermore, let $K = (\kappa(x, y))$ denote the transition kernel matrix, so that $\mu_t = \mu_0 K^t$ is the probability of each state after t iterations of the chain. The stationary distribution can be written as the limit

$$\mu = \lim_{t \rightarrow \infty} \mu_t$$

and can be computed as the normalized eigenvector of M that belongs to the eigenvalue one.

To implement MCMC methods in order to draw samples from a given distribution μ it is necessary to show that μ is the unique stationary distribution of M . In practice it is often more convenient to show that *detailed balance* (also called *local balance*), i.e.

$$\int_{\mathsf{X}} \int_{\mathsf{X}} f(x_1, x_2) d\mu(x_1) d\kappa(x_1, x_2) = \int_{\mathsf{X}} \int_{\mathsf{X}} f(x_2, x_1) d\mu(x_2) d\kappa(x_2, x_1)$$

holds for all bounded and measurable functions f [cf. e.g. Cappé et al., 2005]. If M is in detailed balance, it is *reversible* and has μ as its invariant distribution (the reverse does in general not hold).

There exist a variety of different MCMC methods and we will have to restrict our attention to the most basic ones. We begin the discussion with the Gibbs sampling method, which was developed by Geman and Geman [1984], Gelfand and Smith [1990]. In the following, we assume that X is a product space of N components, i.e. $\mathsf{X} = \mathsf{X}_1 \times \cdots \times \mathsf{X}_N$, so that $X^{(t)} = (X_1^{(t)}, \dots, X_N^{(t)})$. The distribution of

$$X_n^{(t+1)} | X_1^{(t+1)} = y_1, \dots, X_{n-1}^{(t+1)} = y_{n-1}, X_{n+1}^{(t)} = x_{n+1}, \dots, X_N^{(t)} = x_N$$

implementation on a computer is discrete. The effects of this discretization were studied by Roberts et al. [1998].

is defined as

$$(10) \quad d\nu_n(y_n | y_1, \dots, y_{n-1}, x_{n+1}, \dots, x_N) = \frac{d\mu(y_1, \dots, y_n, x_{n+1}, \dots, x_N)}{\int_{\mathcal{X}_n} d\mu(y_1, \dots, y_n, x_{n+1}, \dots, x_N)}$$

for $n = 1, \dots, N$, and $t \in \mathbb{N}$. The Gibbs sampler can be understood as a stochastic version of a coordinate ascent algorithm. In each iteration, a coordinate n is fixed and a new value for X_n is drawn, based on the current state $(X_1, \dots, X_{n-1}, X_{n+1}, \dots, X_N)$. The distribution ν_n is much easier to compute than μ , since the normalization constant of μ drops due to the ratio in Equation 10.

THEOREM A.4.1 (Stationary distribution of the Gibbs sampler). *Let M be the Markov chain of the Gibbs sampler with transition kernel*

$$(11) \quad d\kappa(x, y) = \prod_{n=1}^N d\nu_n(y_n | y_1, \dots, y_{n-1}, y_{n+1}, \dots, x_N)$$

where $x = (x_1, \dots, x_N)$, and $y = (y_1, \dots, y_N)$. The Markov chain M is reversible and μ is the unique stationary distribution of M .

The proof of the theorem can for instance be found in Robert and Casella [1999] and we not be repeated here.

Although the Gibbs sampler seems to solve the problem of approximating summaries of a distribution, there still remains the technical difficulty of computing the integral in Equation 10, which is why this approach is mostly restricted to discrete spaces \mathcal{X} . The Metropolis-Hastings algorithm [Metropolis et al., 1953, Hastings, 1970] circumvents this problem, but instead requires a family of distributions $\{q_x\}_{x \in \mathcal{X}}$ from which proposals can be drawn. Let $M = (X^{(t)})$ again denote a Markov chain on \mathcal{X} and $(P^{(t)})$ a sequence of random variables with

$$P^{(t)} | X^{(t)} = x \sim q_x.$$

$P^{(t)}$ is used as a proposal for the state of the Markov chain at time $t + 1$. If the proposal $\{P^{(t)} = p\}$ is accepted, the Markov chain switches its state to p . Otherwise, it stays in its old state, i.e. $X^{(t+1)} = X^{(t)}$. The state at time $t + 1$ is distributed according to

$$X^{(t+1)} | P^{(t)} = p, X^{(t)} = x \sim \rho(x, p)\delta_x + (1 - \rho(x, p))\delta_p,$$

where

$$\rho(x, p) = \min \left\{ \frac{d\mu(p)dq_p(x)}{d\mu(x)dq_x(p)}, 1 \right\}.$$

THEOREM A.4.2 (Stationary distribution of the Metropolis-Hastings algorithm). *The transition measure of the Metropolis-Hastings algorithm is given by*

$$\kappa(x, Y) = \rho(x, y)q_x(Y) + (1 - r(x))\delta_x(Y),$$

where $r(x) = \int_{\mathcal{X}} \rho(x, y)dq_x(y)$. For suitable proposal distributions $\{q_x\}$ the Markov chain satisfies the detailed balance condition and therefore has μ as its invariant distribution [Robert and Casella, 1999].

The theorem holds for almost any family of proposal distributions, which is why the method has become one of the most commonly used MCMC algorithms.

A.5. Variational Bayes

Variational Bayes is another method to approximate a complicated (posterior) distribution μ . Instead of drawing samples from μ and utilizing the law of large numbers or the ergodic theorem, a class of simpler distributions Ξ is chosen from which we select the closest distribution to μ [cf. e.g. Ghahramani and Beal, 2001, Beal, 2003, Wainwright and Jordan, 2008]. For this approach to be useful, we need a class Ξ such that the most important features of μ can be captured, but the distributions must also be simple enough to allow the computation of important statistics.

First, we need to review some basic facts of variational calculus. Functionals are the central object of variational calculus, which are mappings from a set of functions to the real numbers. The objective of variational calculus is to find functions that maximize or minimize a given functional. As an example, take the set of all curves on some manifold that connect two points a and b . The shortest path between a and b is the curve with minimal length. It is important to notice that the space of all curves is an infinite dimensional function space. To illustrate this, we may approximate any curve by a polygonal chain of n points (see Figure 2). However, for an exact representation of the curve we would in general need $n = \infty$ points. Variational calculus was developed to deal with such infinite dimensional function spaces. Excellent treatments of the subject can be found in Kielhöfer [2010] and Jost and Li-Jost [1998].

In the following, $C[a, b]$ denotes the set of continuous functions from the compact interval $[a, b]$ to the real numbers and $C^p[a, b]$ the set of p times continuously differentiable functions, i.e. every $y \in C^p[a, b]$ has a p th derivative $y^{(p)} \in C[a, b]$. To study extremal points of a functional $J : C^1[a, b] \rightarrow \mathbb{R}$ we use the *Gâteaux derivative*

$$dJ(y, h) = \lim_{t \rightarrow 0} \frac{J(y + th) - J(y)}{t}$$

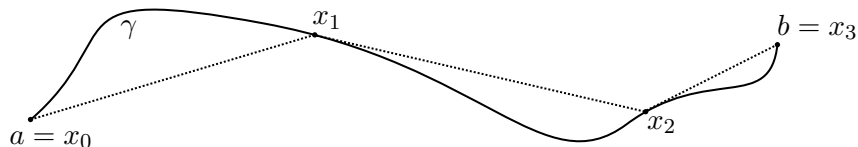


FIGURE 2. Approximation of a curve γ by a polygonal chain (x_0, x_1, x_2, x_3) .

at a point y in the direction of h . If the Gâteaux derivative at a point y is linear in h , we write

$$dJ(y, h) = \delta J(y)h = \left. \frac{d}{dt} J(y + th) \right|_{t=0},$$

which is called the *first variation* of J .

EXAMPLE A.5.1. Let $f(x) = |x|$ which is non-differentiable at $x = 0$ and therefore the Gâteaux derivative $df(0, h) = |h|$ is nonlinear.

Variational calculus is mostly concerned with functionals $J : C^2[a, b] \rightarrow \mathbb{R}$ of the form

$$J(y) = \int_a^b F(x, y(x), y'(x)) dx = \int_a^b F(x, y, y') dx$$

where $F : [a, b] \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ is assumed to be twice continuously differentiable and called the *Lagrange function*. It should be emphasized that F is not a functional and to avoid any confusion we use the notation F_y and $F_{y'}$ for the partial derivatives with respect to the second and third variable. The Gâteaux derivative $dJ(y, h)$ is linear in h and can be written as

$$\delta J(y)h = \int_a^b [hF_y(x, y, y') + h'F_{y'}(x, y, y')] dx.$$

A necessary condition for a point y to be extremal is that the Gâteaux derivative is zero. Using partial integration and the *fundamental lemma of variational calculus*, a critical point y of J satisfies the *Euler-Lagrange equation*

$$\frac{d}{dx} F_{y'}(x, y, y') = F_y(x, y, y').$$

Optimization problems in statistics often involve additional constraints which restrict the set of feasible functions. For instance, a minimizer of a functional might be required to be a probability distribution. Recall that extrema of real valued functions subject to constraints are found by using the method of Lagrange multipliers. Assume that we are seeking a solution

to the optimization problem

$$\begin{aligned} & \text{minimize } f(x) \\ & \text{subject to } g(x) = c. \end{aligned}$$

At a solution x_0 the gradients of f and g satisfy $\nabla f(x_0) = \lambda \nabla g(x_0)$ for some $\lambda \in \mathbb{R}$ (see Figure 3). Therefore, solutions of the constrained optimization problem can be found by first identifying the critical points of the Lagrangian

$$\Lambda(x, \lambda) = f(x) + \lambda(g(x) - c).$$

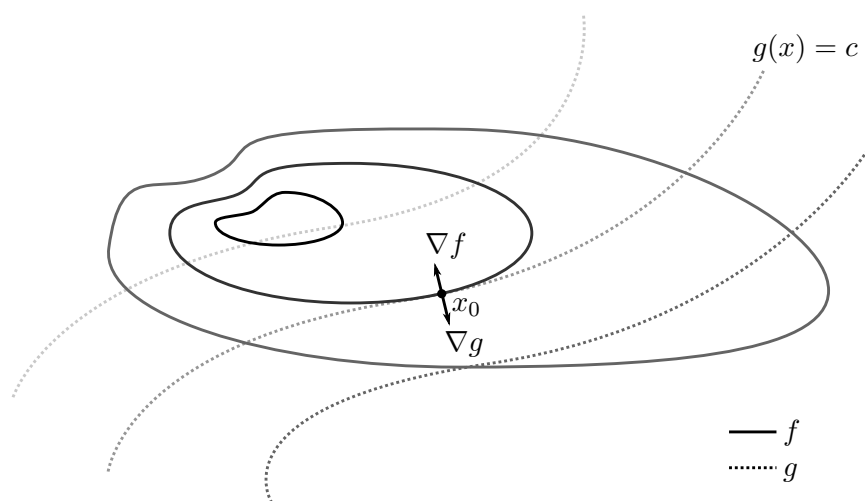


FIGURE 3. Two functions $f, g : \mathbb{R}^2 \rightarrow \mathbb{R}$ where $g(x) = c$ is used as a constraint for finding critical points of f . At a critical point x_0 the gradients satisfy $\nabla f(x_0) = \lambda \nabla g(x_0)$ for some $\lambda \in \mathbb{R}$.

In a similar way the Euler-Lagrange equation can be extended to account for the set of feasible functions. Assume we want to find a solution to the optimization problem

$$\begin{aligned} & \text{minimize } J(y) = \int_a^b F(x, y, y') dx \\ & \text{subject to } K(y) = \int_a^b G(x, y, y') dx = c, \end{aligned}$$

where K is called an *isoperimetric* constraint. By including the constraint in the Euler-Lagrange equation we find that for a solution y there exists a $\lambda \in \mathbb{R}$ such that

$$\frac{d}{dx} [F_{y'}(x, y, y') + \lambda G_{y'}(x, y, y')] = F_y(x, y, y') + \lambda G_y(x, y, y').$$

We may now return to the problem at hand and illustrate the most common use case of the variational Bayes approach. Assume that μ is a distribution of a n -dimensional random variable $\Theta \in \mathbb{R}^n$. For a so-called *mean field* approximation we select a class Ξ of distributions such that every $\xi \in \Xi$ has a density function

$$d\xi(\theta)/d\theta = \prod_{i=1}^n \xi_i(\theta_i).$$

The distance between distributions is typically measured with the Kullback-Leibler divergence, although it is not symmetric. Hence, we seek a solution to the optimization problem

$$\underset{\xi \in \Xi}{\text{minimize}} \left[D_{\text{KL}}(\xi||\mu) = \int_{\Theta} \log \left(\frac{d\xi(\theta)}{d\mu(\theta)} \right) d\xi(\theta) \right].$$

It is easy to verify that by minimizing the Kullback-Leibler divergence a lower bound on the marginal likelihood is obtained. Let

$$\mu = \text{pr}_{\Theta|X}(\cdot|x)$$

for some (discrete) observation $\{X = x\}$. The marginal likelihood pr_X can be written as

$$\log \text{pr}_X(x) = D_{\text{KL}}(\xi||\mu) - D_{\text{KL}}(\xi||\text{pr}_{X,\Theta}(x, \cdot)).$$

Hence, by minimizing $D_{\text{KL}}(\xi||\mu)$ we obtain the lower bound

$$\log \text{pr}_X(x) \geq -D_{\text{KL}}(\xi||\text{pr}_{X,\Theta}(x, \cdot)).$$

It is common to not require that the elements of Ξ are normalized, so that we may instead consider the equivalent constrained optimization problem

$$\begin{aligned} & \underset{\xi \in \Xi}{\text{minimize}} \quad D_{\text{KL}}(\xi||\mu) \\ & \text{subject to} \quad \int_{\Theta} d\xi(\theta) = 1. \end{aligned}$$

To stay within the class of feasible functions Ξ when solving for the optimum we need to employ a small trick. Knowing that $\Xi = \Xi_1 \times \Xi_2 \times \dots \times \Xi_n$ factorizes into n components, a solution can be obtained with an iterative *coordinate descent* algorithm by optimizing one component at a time [cf. e.g. Bertsekas, 1999]. At every iteration t we have a solution ξ^t . We pick one of the components ξ_i^t , $i = (t \bmod n) + 1$, and find a solution ξ_i^{t+1} to

the optimization problem

$$\begin{aligned} & \underset{\xi_i^{t+1} \in \Xi_i}{\text{minimize}} \int_{\Theta} \log \left(\frac{\xi_i^{t+1}(\theta_i) \xi_{-i}^t(\theta_{-i})}{f_{\mu}(\theta)} \right) d\xi^t(\theta) \\ & \text{subject to} \int_{\Theta_i} \xi_i(\theta_i) d\theta_i = 1, \end{aligned}$$

to yield $d\xi^{t+1}/d\theta = \xi_i^{t+1} \xi_{-i}^t$, where $\xi_{-i}^t = \prod_{j \neq i} \xi_j^t$ is given and $f_{\mu} = d\mu/d\theta$ is the density function of μ . The algorithm is iterated until convergence. Since the Kullback-Leibler divergence is convex, the solution ξ_i^{t+1} is fully determined by the Euler-Lagrange equation

$$\frac{\partial}{\partial \xi_i^{t+1}} \xi_i^{t+1} \int_{\Theta_{-i}} \xi_{-i}(\theta_{-i}) \log \left(\frac{\xi_i^{t+1} \xi_{-i}^t(\theta_{-i})}{f_{\mu}(\theta)} \right) d\theta_{-i} + \lambda = 0.$$

By solving for ξ_i^{t+1} we obtain the general solution

$$\xi_i^{t+1}(\theta_i) = \frac{1}{Z_i^{t+1}} \exp \left\{ \int_{\Theta_{-i}} \xi_{-i}^t(\theta_{-i}) \log (f_{\mu}(\theta)) d\theta_{-i} \right\},$$

where Z_i^{t+1} is the normalization constant of the density. The solution already shows that this approach is especially fruitful if μ belongs to the exponential family. The computation can also be highly structured by taking advantage of conditional independencies if the distribution μ is represented as a graphical model, such as a *Bayesian network* or a *factor graph* [cf. e.g. Winn and Bishop, 2005, Dauwels, 2007].

A.6. Probability in statistics

While mathematical probability theory is mostly concerned with a consistent axiomatization of probability and the development of a sound calculus, in statistics we demand much more than that. Assigning probabilities to events requires a clear interpretation of what probabilities mean as well as having a persuasive justification for the axioms that constrain our assignments [cf. e.g. Kadane, 2011, Lindley, 1972].

When we speak about an *aleatory*³ probability, we refer to some physical process, which might be truly or just appear to be random. Closely tied to this notion is the *frequentist* school of probability, where probabilities are defined as the limiting frequencies of the events if the process is observed indefinitely. Clearly, in many situations it is not possible to actually observe an event several times. Therefore, we instead adopt the *epistemic*⁴ interpretation of probability, as pioneered by Bayes [1763] and Laplace [1820],

³*Alea* is from the Latin and means die or game of die.

⁴*Episteme* is derived from the Greek and means knowledge.

where probabilities describe the certainty of propositions. Many authors have contributed to this interpretation of probability and developed it into several distinct directions. For instance, statistics can be seen as an extension of inductive logic, as suggested by Keynes [1921], Cox [1961], and Carnap [1962]. However, we will mostly use the *subjective* interpretation, which describes a personal degree of belief. Important contributions to this interpretation were made by Ramsey [1931], Koopman [1940], de Finetti [1964], and Savage [1972].

Naturally the question arises as to why our personal beliefs should follow the axioms of probability. Let Ω be a sample space with σ -algebra Σ . A possible approach to motivate subjective probabilities is to define a relation \succ on Σ such that the statement $A \succ B$ reflects our belief that A is more probable than B . Under suitable and compelling restrictions on \succ there exists a probability measure $p : \Sigma \rightarrow [0, 1]$ with

$$A \succ B \quad \text{iff} \quad p(A) > p(B)$$

for all $A, B \in \Sigma$ [cf. e.g. de Finetti, 1964, Fishburn, 1986, Savage, 1972]. A more intuitive approach was discussed by de Finetti [1964] [cf. also Lindley, 1972]. Suppose you are betting on a set of pairwise disjoint events $A_1, \dots, A_n \in \Sigma$ with $\cup_{i=1}^n A_i = \Omega$ offered by a bookie. In this gambling game, the bookie fixes a set of odds for each A_i which represents his personal belief $p(A_i)$ that A_i will occur. Afterwards, you decide on a stake $s(A_i)$, which you pay to the bookie for each $i = 1, \dots, n$. The peculiarity of this game is that a stake $s(A_i)$ may also be negative, in which case the bookie pays you an amount of $-s(A_i)$. If the event A_i occurs, you receive a prize of $s(A_i)/p(A_i)$ from the bookie so that your net gain is

$$g_i = \frac{s(A_i)}{p(A_i)} - \sum_{j=1}^n s(A_j).$$

We say that the personal beliefs of the bookie are *coherent* if you are not able to fix stakes such that you will always win. Such a set of stakes in combination with the proposed odds is also called a *Dutch book*. If we consider the $p(A_i)$ as unknowns, we obtain a system of linear equations with determinant

$$\det \begin{pmatrix} 1/p(A_1) - 1 & -1 & \dots & -1 \\ -1 & 1/p(A_2) - 1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & 1/p(A_n) - 1 \end{pmatrix} = \frac{1 - \sum_{i=1}^n p(A_i)}{\prod_{i=1}^n p(A_i)}.$$

If the determinant is non-zero, i.e. the probabilities do not sum up to one, you are able to fix stakes such that your gain will always be positive, no

matter what event occurs. For finite probability spaces, it can be shown that the personal beliefs have to satisfy the axioms of probability. For an infinite probability space, it turns out that the σ -additivity axiom is hard to justify. A comprehensive discussion on this matter was provided by Kadane et al. [1999]. Nevertheless, Dutch book arguments were found, see for instance Williamson [1999], Freedman [2003]. A notable extension of the results of de Finetti [1964] was provided by Freedman and Purves [1969]. They showed that in light of new information, a bookie should update his beliefs according to Bayes' rule.

A.7. Statistical inference

Predictive inference is the process of forming beliefs about future events from our past observations. The problem of how this process can be justified is nowadays mostly attributed to the philosopher David Hume and his highly influencing book “A Treatise of Human Nature” [Zabell, 2005], where he raised the question: *Why should the future resemble the past?* We will not discuss the problem of induction in detail, but give a simple answer that allows us to use our methodology for reaching scientific conclusions. To put it simply, we will not only have to assume that future events will resemble the past, but also how they do. In statistics our assumptions are most elegantly expressed by symmetry assumptions on the probability function of future and past events [Zabell, 2005].

Exchangeability is one of the most basic and notable symmetry assumptions on a random sequence. It was introduced by Haag [1924] and further developed by de Finetti [1931, 1964]. Assume that X_1, \dots, X_n, \dots are binary random variables for which

$$\bar{X}_n = X_1, \dots, X_n \stackrel{d}{=} X_{\pi(1)}, \dots, X_{\pi(n)}$$

for all $n \in \mathbb{N}$ and all permutations π . The famous *representation theorem* by de Finetti [1931] shows that for an exchangeable sequence there exists a distribution μ such that the joint probability function takes the form

$$\text{pr}_{\bar{X}_n}(\bar{x}) = \int_{[0,1]} \theta^{k(\bar{x})} (1 - \theta)^{n - k(\bar{x})} d\mu(\theta),$$

where $\bar{x} = (x_1, \dots, x_n) \in \{0, 1\}^n$ and $k(\bar{x}) = \sum_{i=1}^n x_i$. We may introduce a random variable Θ so that

$$\text{pr}_{\bar{X}_n | \Theta}(\bar{x} | \theta) = \theta^{k(\bar{x})} (1 - \theta)^{n - k(\bar{x})}$$

can be interpreted as the likelihood function, which describes the *parametric model*, and $\text{pr}_{\Theta} = \mu$ as the prior distribution. Hence, the representation theorem, which is a result from pure probability theory, justifies the Bayesian

approach to inductive inference [cf. e.g. Bernardo, 1996]. Using basic probability theory, we obtain the prediction

$$\text{pr}_{X_{n+1} | \bar{X}_n}(x | \bar{x}) = \frac{\text{pr}_{X_{n+1}, \bar{X}_n}(x, \bar{x})}{\text{pr}_{\bar{X}_n}(\bar{x})}$$

for X_{n+1} based on the observed events $\{\bar{X}_n = \bar{x}\}$. Similarly, our prior belief about Θ can be updated using *Bayes' rule*

$$\text{dpr}_{\Theta | \bar{X}_n}(\theta | \bar{x}) = \frac{\text{pr}_{\bar{X} | \Theta}(\bar{x} | \theta) \text{dpr}_{\Theta}(\theta)}{\text{pr}_{\bar{X}_n}(\bar{x})}.$$

The event $\{\Theta = \theta\}$ can be interpreted as a hypothesis for the observed event $\{\bar{X}_n = \bar{x}\}$.

There exist several generalizations of the representation theorem by de Finetti. For instance, random sequences of finite cardinality were discussed by Hewitt and Savage [1955]. Diaconis and Freedman [1980a,b] analyzed finite exchangeable sequences and showed a representation theorem for Markov chains. However, in practice, the approach of obtaining the model from symmetry assumptions is often too difficult. Hence, it is common to define the model directly, which then induces certain symmetries of the joint distribution.

A.8. Statistical decision theory

What is nowadays understood as statistical decision theory has been developed over the past century from many different perspectives. The first derivation of epistemic probability and utility from a system of preference axioms can be traced back to Ramsey [1931]⁵. However, it was probably Savage [1972] who first developed a complete and sound axiomatization of decision making under uncertainty. His treatise relied partially on the work of von Neumann and Morgenstern [1953] in the context of economic game theory. Notable contributions were also made by Wald [1950], although from a more frequentist perspective. He developed the idea of admissible decision rules and thereby established a strong argument for the Bayesian paradigm. Other important contributions were made by DeGroot [1962, 2004], who first applied this theory to sequential experimentation. In the following, we will give a brief introduction to decision theory, but without discussing personal preferences to derive subjective probabilities [similar to the discussion of Lindley, 1972, Young and Smith, 2005]. Good reviews on the approach by Savage [1972] can be found in Fishburn [1982], and Kadane [2011]. An

⁵For an excellent discussion of the essay by Ramsey [1931] see Zabell [2005, Chapter 6].

excellent summary of the most important milestones in the development of decision theory was written by Fishburn [1989].

To motivate the following discussion, consider first a simple example where you need to decide between two construction procedures for a bridge [cf. Fishburn, 1970]. Procedure A costs 150 million, whereas procedure B costs only 100 million. Engineers provide you with probabilities $p_A(t)$ and $p_B(t)$ for each procedure for completing the bridge within t years, which are based on their previous experience and observations. Assume that the probabilities are given as

$$p_A(t) = \begin{cases} 0 & \text{if } t \leq 2 \\ \frac{t-2}{3} & \text{if } 2 < t \leq 5 \end{cases}, \quad p_B(t) = \begin{cases} 0 & \text{if } t \leq 3 \\ \frac{t-3}{4} & \text{if } 3 < t \leq 7 \end{cases}.$$

In order to come to a decision, you have to specify your personal utility for both procedures and the time of completion t . Assume your utilities are

$$u(A, t) = -(t - 2)^2 - 5, \quad u(B, t) = -(t - 3)^2,$$

so that we obtain the *expected utilities*

$$\bar{u}(A) = \int_2^5 u(A, t)p_A(t)dt = -8, \quad \bar{u}(B) = \int_3^7 u(B, t)p_B(t)dt = -10.33.$$

Procedure A has a higher expected utility and therefore you should decide for this option.

We may formalize the process of decision making by introducing an action space \mathcal{A} and a random variable A that takes values in \mathcal{A} . Furthermore, let Y be an observable random variable in \mathcal{Y} and assume that we have observations $\{X = x\}$ in \mathcal{X} from which we would like to predict Y . A bounded function $u : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ quantifies our subjective utility $u(a, y)$ of an action $a \in \mathcal{A}$ in case the event $\{Y = y\}$ occurs. The *expected utility* is defined as

$$\bar{u}(a) = \int_{\mathcal{Y}} u(a, y) d\text{pr}_{Y|A, X}(y | a, x),$$

and best action to choose is the one with *maximal expected utility*, i.e.

$$\hat{a} = \arg \sup_{a \in \mathcal{A}} \bar{u}(a).$$

REMARK A.8.1. In this context, the theory by Savage [1972] can be framed as follows. Let (Ω, Σ) and (S, \mathcal{S}) be two measurable spaces and \mathbb{F} a space of random variables such that each $F \in \mathbb{F}$ is a mapping $F : (\Omega, \Sigma) \rightarrow (S, \mathcal{S})$. A total ordering \succ on \mathbb{F} is used to implement personal (subjective) preferences. If we adopt the original interpretation of random variables by Kolmogoroff, the statement $F \succ G$ means that we prefer the experiment F over G since we might believe that F is more informative than G . Savage [1972] developed axioms that guarantee the existence of a unique probability

measure P and a utility function u , such that (Ω, Σ, P) is a probability space and

$$F \succ G \quad \text{iff} \quad \mathbb{E} u(F) > \mathbb{E} u(G),$$

for all $F, G \in \mathbb{F}$.⁶ In the theory by Savage, Ω contains the possible *states of nature* and S is called the space of *consequences*. We may assume that $S = \mathcal{A} \times \mathbf{Y}$ and thereby obtain an axiomatic justification of the decision theory outlined earlier.

EXAMPLE A.8.1 (Poppe, Benner, and Elze [2012]). Consider an experiment where a psychometric function is measured with an adaptive method. For each measurement one of L possible stimuli is presented and the subject can choose between K responses. Let $A_n \in \mathbf{A}$ denote the stimulus and $X_n \in \mathbf{X}$ the response of the n th measurement. Based on the past observations

$$\{\bar{A}_n = \bar{a}_n, \bar{X}_n = \bar{x}_n\} = \{A_1 = a_1, X_1 = x_1, \dots, A_n = a_n, X_n = x_n\}$$

a new stimulus $\{A_{n+1} = a_{n+1}\}$ is selected. The decision is based on the expected utility

$$\bar{u}(a) = \sum_{x \in \mathbf{X}} u(a, x) \text{pr}_{X_{n+1} | \bar{A}_n, \bar{X}_n}(x | \bar{a}_n, \bar{x}_n),$$

where $u(a, x)$ quantifies our subjective utility of observing x in response to a stimulus a . A good choice for the utility function is the certainty by which the psychometric function can be predicted, which implicitly depends on the past observations $\{\bar{A}_n = \bar{a}_n, \bar{X}_n = \bar{x}_n\}$.

The objective of many applications is not to predict observations directly, but to estimate an unobserved parameter Θ of the statistical model that takes values in some parameter space Θ . The parameter Θ captures all essential information in X about Y , so that

$$Y \perp\!\!\!\perp X | A, \Theta.$$

The expected utility then becomes⁷

$$\begin{aligned} \bar{u}(a) &= \int_{\mathbf{Y}} u(a, y) \text{dpr}_{Y | A, X}(y | a, x) \\ &= \int_{\Theta} \int_{\mathbf{Y}} u(a, y) \text{dpr}_{Y | A, \Theta}(y | a, \theta) \text{dpr}_{\Theta | X}(\theta | x), \end{aligned}$$

⁶In fact, Savage [1972, p. 43] agrees with de Finetti [1964] that probabilities should not be σ -additive. Hence, an additional axiom is required to obtain a true probability measure P [Villegas, 1964, Fishburn, 1986].

⁷In this context x is assumed to be fixed so that we may write $\bar{u}(a)$ instead of $\bar{u}(a, x)$.

and by defining

$$u^*(a, \theta) = \int_{\mathcal{Y}} u(a, y) \mathrm{dpr}_{Y|A, \Theta}(y | a, \theta),$$

we may write the expected utility as

$$\bar{u}(a) = \int_{\Theta} u^*(a, \theta) \mathrm{dpr}_{\Theta|X}(\theta | x).$$

DEFINITION A.8.1 (Point estimate). Assume that $\mathcal{A} = \Theta$ and let $u^* : \mathcal{A} \times \Theta \rightarrow \mathbb{R}$ be a bounded utility function. An estimate

$$\hat{\theta} = \arg \sup_{\theta' \in \Theta} \bar{u}(\theta') = \arg \sup_{\theta' \in \Theta} \int_{\Theta} u^*(\theta', \theta) \mathrm{dpr}_{\Theta|X}(\theta | x)$$

is called a *point estimate* of Θ .

A different approach to decision theory was developed by Wald [1950]. We augment the setup so far by introducing a decision space \mathcal{D} which consists of functions $\delta : \mathcal{X} \rightarrow \mathcal{A}$, i.e. δ assigns an action to each observation $x \in \mathcal{X}$. Furthermore, the utility function is replaced by a loss function $\mathcal{L} : \mathcal{A} \times \Theta \rightarrow [0, \infty)$, where $\mathcal{L}(a, \theta)$ quantifies the subjective loss of selecting action a for an observation $\{X = x\}$ when θ is the true parameter. Central to the decision theory of Wald [1950] is the definition of the risk function

$$R(\theta, \delta) = \int_{\mathcal{X}} \mathcal{L}(\delta(x), \theta) \mathrm{dpr}_{X|\Theta}(x | \theta).$$

EXAMPLE A.8.2. Assume that $X = (X_1, X_2, \dots, X_n)$ are i.i.d. normal random variables with mean parameter μ and variance σ^2 . Assume that $\mathcal{A} = \Theta$ and let $\mathcal{L}(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$. We may consider the decision functions

$$\delta_1(x) = \frac{1}{n} \sum_{i=1}^n (x_i), \quad \delta_2(x) = x_1, \quad \delta_3(x) = \frac{1}{2},$$

for observations $x = (x_1, \dots, x_n)$, so that we obtain the risks

$$R(\mu, \delta_1) = \frac{\sigma^2}{n}, \quad R(\mu, \delta_2) = \sigma^2, \quad R(\mu, \delta_3) = \frac{1}{4} + (\mu - 1)\mu.$$

The risk of δ_1 goes to zero for $n \rightarrow \infty$, whereas for δ_2 it is constant. The decision rule δ_3 does not depend on the data x and has zero risk only if $\mu = 1/2$.

DEFINITION A.8.2 (Dominated decision rule, Wald [1950]). Let Θ be a parameter space. A decision rule δ is said to be *dominated* by a decision rule δ^* if

$$R(\theta, \delta) \geq R(\theta, \delta^*)$$

for all $\theta \in \Theta$ and

$$R(\theta, \delta) > R(\theta, \delta^*)$$

for at least one $\theta \in \Theta$.

A decision function δ is hence dominated by a function δ^* if the risk of δ^* is lower for at least one parameter and for no parameter it is higher.

DEFINITION A.8.3 (Admissible decision rule, Wald [1950]). Let \mathcal{D} be a space of decision functions. A decision function $\delta \in \mathcal{D}$ is *admissible* if there exists no $\delta^* \in \mathcal{D}$ which dominates δ .

If we consider again Example A.8.2 and assume that $\mathcal{D} = \{\delta_1, \delta_2, \delta_3\}$, we see that δ_2 is dominated by δ_1 for $n > 1$. The set of admissible decision rules is therefore given by $\{\delta_1, \delta_3\}$. This example already shows that in general there are many admissible decision rules and we may therefore ask for a criterion to select a single rule which is regarded as optimal. A possible choice for such a criterion is the *minimax* decision rule

$$\hat{\delta} = \arg \inf_{\delta \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, \delta).$$

The solution $\hat{\delta}$ may not be unique and therefore also include decision rules that are inadmissible. For instance, consider two decision rules δ_1 and δ_2 such that δ_1 dominates δ_2 and

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) = \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If δ_1 is a minimax decision rule, then so is δ_2 , however, δ_2 is inadmissible because it is dominated by δ_1 [cf. e.g. Young and Smith, 2005].

In game theory, minimax rules are frequently used in zero-sum games, cf. von Neumann and Morgenstern [1953]. In this context, $L(a, \theta)$ represents your loss when selecting action a and your opponent chooses θ . Since the objective of your opponent is to win the game, it can be assumed that he will select θ such that your loss is maximized. However, in statistics it seems overly conservative to select decision rules according to their maximum risk. In a Bayesian setting we would instead explicate our prior knowledge about the parameter Θ by selecting an appropriate distribution and choose the decision rule $\hat{\delta}$ that minimizes the *average risk*, i.e.

$$\hat{\delta} = \arg \inf_{\delta \in \mathcal{D}} \int_{\Theta} R(\theta, \delta) d\text{pr}_{\Theta}(\theta).$$

By rewriting the minimum average risk, we obtain

$$\inf_{\delta \in \mathcal{D}} \int_{\Theta} R(\theta, \delta) d\text{pr}_{\Theta}(\theta) = \int_{\mathcal{X}} d\text{pr}_X(x) \inf_{\delta \in \mathcal{D}} \int_{\Theta} \mathcal{L}(\delta(x), \theta) d\text{pr}_{\Theta|X}(\theta|x),$$

where

$$\text{pr}_X(x) = \int_{\Theta} \text{pr}_{X|\Theta}(x|\theta) d\text{pr}_{\Theta}(\theta).$$

With $\mathcal{L}(a, \theta) = c - u(a, \theta)$ for an appropriate constant c we see that the minimum average risk is indeed very similar to the principle of maximum expected utility, where we focus only on a particular observation $\{X = x\}$ ⁸.

DEFINITION A.8.4 (Bayes estimator). The decision rule

$$\hat{\delta}(x) = \arg \inf_{a \in \mathcal{A}} \int_{\Theta} \mathcal{L}(a, \theta) d\text{pr}_{\Theta|X}(\theta|x)$$

minimizes the average risk and is called a *Bayes estimator* with respect to the prior distribution pr_{Θ} .

Under mild assumptions on the prior distribution, it can be shown that the Bayes estimator is admissible, cf. Wald [1950].

An important concept is also the *consistency* of a Bayesian estimator, which refers indirectly to the convergence of the posterior distribution. We say that an estimator is consistent if it converges to the “true” value of the parameter. To make this notion more precise, consider the following

DEFINITION A.8.5 (Consistency). Let $\{X_1 = x_1, \dots, X_n = x_n\}$ be a sequence of observations and let $\{p_{\theta}\}_{\theta \in \Theta}$ be a family of distributions. Each random variable X_i is distributed as p_{θ^*} , where $\theta^* \in \Theta$. We call a Bayesian estimator $\hat{\delta}$ *consistent* iff

$$\hat{\delta}(x_1, \dots, x_n) \rightarrow \theta^*$$

almost surely for all $\theta^* \in \Theta$.

It should be noted that this definition is much stronger than what is required for the highly celebrated theorem by *Doob* [Theorem 10.10 Van der Vaart, 2000, Doob, 1949]. Indeed, not every Bayesian estimator is consistent in this sense and especially nonparametric models can suffer from inconsistencies [Kleijn and van der Vaart, 2006, Ghosal et al., 2007, van der Vaart and Van Zanten, 2008].

⁸The fact that we may express the loss function in this way depends on the assumption that the utility function is bounded.

Bibliography

- R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows: theory, algorithms, and applications. 1993.
- D. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198, 1985.
- C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.
- E. Archer, I. M. Park, and J. Pillow. Bayesian entropy estimation for countable discrete distributions. *arXiv preprint arXiv:1302.0328*, 2013.
- J. C. Avise, J. Shapira, S. W. Daniel, C. F. Aquadro, and R. A. Lansman. Mitochondrial dna differentiation during the speciation process in peromyscus. *Molecular Biology and Evolution*, 1(1):38–56, 1983.
- M. Bačák. The proximal point algorithm in metric spaces. *Israel Journal of Mathematics*, 194(2):689–701, 2013.
- M. Bačák. Computing Medians and Means in Hadamard Spaces. *SIAM J. Optim.*, 24(3):1542–1566, 2014a. ISSN 1052-6234.
- M. Bačák. *Convex analysis and optimization in Hadamard spaces*, volume 22 of *De Gruyter Series in Nonlinear Analysis and Applications*. Walter de Gruyter & Co., Berlin, 2014b.
- H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in applied mathematics*, 7(3):309–343, 1986.
- H.-J. Bandelt and A. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in mathematics*, 92(1):47–105, 1992.
- Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-dna binding sites. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 28–37. ACM, 2003.
- J.-P. Barthelemy. From copair hypergraphs to median graphs with latent vertices. *Discrete Mathematics*, 76(1):9–28, 1989.
- H. Bauer. *Probability Theory*, volume 23 of *Studies in Mathematics*. Walter de Gruyter, 1996.
- H. Bauer. *Measure and integration theory*, volume 26 of *Studies in Mathematics*. Walter de Gruyter, 2001.

- T. Bayes. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions (1683-1775)*, 53:370–418, 1763.
- M. J. Beal. *Variational algorithms for approximate Bayesian inference*. PhD thesis, University of London, 2003.
- P. Benner and M. Bačák. Computing the posterior expectation of phylogenetic trees. *arXiv preprint arXiv:1305.3692*, 2013.
- P. Benner, M. Bačák, and P.-Y. Bourguignon. Point estimates in phylogenetic reconstructions. *Bioinformatics*, 30(17):i534–i540, 2014.
- P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein–dna interactions: how good an approximation is it? *Nucleic acids research*, 30(20):4442–4451, 2002.
- J. M. Bernardo. The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences*, 4:111–122, 1996.
- D. P. Bertsekas. *Nonlinear programming*. 1999.
- D. P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129(2, Ser. B):163–195, 2011. ISSN 0025-5610. doi: 10.1007/s10107-011-0472-0. URL <http://dx.doi.org/10.1007/s10107-011-0472-0>.
- D. P. Bertsekas and J. N. Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.
- L. Billera, S. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Adv. in Appl. Math.*, 27(4):733–767, 2001. ISSN 0196-8858. doi: 10.1006/aama.2001.0759. URL <http://dx.doi.org/10.1006/aama.2001.0759>.
- D. Blackwell and J. B. MacQueen. Ferguson distribution via pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.
- L. Bottou. Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998.
- N. Bourbaki and S. K. Berberian. *Integration I*. Springer, 2004a.
- N. Bourbaki and S. K. Berberian. *Integration II*. Springer, 2004b.
- P.-Y. Bourguignon. personal communication, 2012.
- S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- M. R. Bridson and A. Haefliger. *Metric spaces of non-positive curvature*, volume 319. Springer, 1999.

- D. Bryant. A classification of consensus methods for phylogenetics. *DIMACS series in discrete mathematics and theoretical computer science*, 61:163–184, 2003.
- D. Bryant, N. Galtier, and M.-A. Poursat. Likelihood calculation in molecular phylogenetics. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 33–62. Oxford University Press, 2005.
- O. P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the archaeological and historical sciences*, 1971.
- J. Bunge and M. Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.
- O. Cappé, E. Moulines, and T. Rydén. *Inference in hidden Markov models*, volume 6. Springer, 2005.
- R. Carnap. Logical foundations of probability. 1962.
- G. Celeux. Bayesian inference for mixture: The label switching problem. In *Compstat*, pages 227–232. Springer, 1998.
- G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970, 2000.
- J. T. Chang and D. Pollard. Conditioning as disintegration. *Statistica Neerlandica*, 51(3):287–317, 1997.
- R. T. Cox. *The algebra of probable inference*. Johns Hopkins Press Baltimore, 1961.
- J. Dauwels. On variational message passing on factor graphs. In *IEEE International Symposium on Information Theory*, pages 2546–2550. IEEE, 2007.
- M. W. Davis. Nonpositive curvature and reflection groups, 1994.
- B. de Finetti. Funzione Caratteristica Di un Fenomeno Aleatorio. 6. Memorie, pages 251–299. Accademia Nazionale del Linceo, 1931.
- B. de Finetti. Foresight: Its Logical Laws in Subjective Sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 93–158. Wiley, 1964.
- M. H. DeGroot. Uncertainty, information, and sequential experiments. *The Annals of Mathematical Statistics*, 33(2):404–419, 1962.
- M. H. DeGroot. *Optimal statistical decisions*. Wiley-Interscience, Hoboken and N.J., 2004.
- P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980a.
- P. Diaconis and D. Freedman. de finetti’s theorem for markov chains. *The Annals of Probability*, pages 115–130, 1980b.
- P. Diaconis and D. Freedman. On the consistency of bayes estimates. *The Annals of Statistics*, pages 1–26, 1986.

- J. L. Doob. Application of the theory of martingales. *Le calcul des probabilités et ses applications*, pages 23–27, 1949.
- A. Dress, K. Huber, J. Koolen, V. Moulton, and A. Spillner. *Basic phylogenetic combinatorics*. Cambridge University Press, Cambridge, 2012. ISBN 978-0-521-76832-0.
- A. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 1(214), 2007. doi: 10.1186/1471-2148-7-214. URL <http://dx.doi.org/10.1186/1471-2148-7-214>.
- S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.
- J. Elstrodt. *Maß- und Integrationstheorie*. Springer, 2004.
- J. Fabius. Asymptotic behavior of bayes' estimates. *The Annals of Mathematical Statistics*, 35(2):846–856, 1964.
- J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.
- T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.
- T. S. Ferguson, E. G. Phadia, and R. C. Tiwari. Bayesian nonparametric inference. In *Lecture Notes-Monograph Series; Current Issues in Statistical Inference; Essays in Honor of D. Basu*, volume 17, pages 127–150. 1992.
- P. C. Fishburn. *Utility theory for decision making*. Publications in operations research. Wiley, 1970.
- P. C. Fishburn. *The foundations of expected utility.*, volume 31 of *Theory and Decision Library*. Springer Netherlands, 1982.
- P. C. Fishburn. The axioms of subjective probability. *Statistical Science*, pages 335–345, 1986.
- P. C. Fishburn. Foundations of decision analysis: along the way. *Management Science*, 35(4):387–405, 1989.
- D. A. Freedman. On the asymptotic behaviour of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403, 1963.
- D. A. Freedman. Notes on the dutch book argument. *Lecture Notes, Department of Statistics, University of Berkley at Berkley*, <http://www.stat.berkeley.edu/~census/dutchdef.pdf>, 2003.
- D. A. Freedman and R. A. Purves. Bayes' method for bookies. *The Annals of Mathematical Statistics*, pages 1177–1186, 1969.
- O. Gascuel. *Mathematics of evolution and phylogeny*. Oxford University Press, 2005.

- A. Gavruskin and A. J. Drummond. The space of ultrametric phylogenetic trees. *arXiv preprint arXiv:1410.3544*, 2014.
- A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.
- S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
- C. J. Geyer. Markov chain monte carlo maximum likelihood. 1991.
- C. J. Geyer and E. A. Thompson. Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920, 1995.
- Z. Ghahramani and M. J. Beal. Propagation algorithms for variational bayesian learning. *Advances in neural information processing systems*, pages 507–513, 2001.
- S. Ghosal, J. K. Ghosh, and R. Ramamoorthi. Posterior consistency of dirichlet mixtures in density estimation. *Annals of Statistics*, 27(1):143–158, 1999.
- S. Ghosal, A. Van Der Vaart, et al. Posterior convergence rates of dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.
- J. H. Gillespie. *The Causes of Molecular Evolution*. Oxford Series in Ecology and Evolution, 1991.
- A. Gnedin. A species sampling model with finitely many types. *Elect. Comm. Probab*, 15:79–88, 2010.
- A. Gnedin and J. Pitman. Exchangeable gibbs partitions and stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.
- A. V. Goldberg and R. E. Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988.
- I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.
- M. Gromov. *Hyperbolic groups*. Springer, 1987.
- J. Haag. Sur un probleme général de probabilités et ses diverses applications. In *Proceedings of the International Congress of Mathematicians, Toronto*, volume 1, pages 659–674, 1924.
- W. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57:97–109, 1970.
- J. Hess and N. Goldman. Addressing inter-gene heterogeneity in maximum likelihood phylogenomic analysis: yeasts revisited. *PloS one*, 6(8):e22783, 2011.

- E. Hewitt and L. J. Savage. Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, pages 470–501, 1955.
- M. Holder, J. Sukumaran, and P. Lewis. A Justification for Reporting the Majority-Rule Consensus Tree in Bayesian Phylogenetics. *Systematic Biology*, 57:814–821, 2003.
- S. Holmes. Statistical approach to tests involving phylogenies. In O. Gascuel, editor, *Mathematics of evolution and phylogeny*, pages 91–120. Oxford University Press, Oxford, UK, 2005.
- J. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.
- P. M. Huggins, W. Li, D. Haws, T. Friedrich, J. Liu, and R. Yoshida. Bayes Estimators for Phylogenetic Reconstruction. *Systematic Biology*, 60(4):528–540, 2011. doi: 10.1093/sysbio/syr021. URL <http://sysbio.oxfordjournals.org/content/60/4/528.abstract>.
- A. Isaev. *Introduction to Mathematical Methods in Bioinformatics*. Springer, 2006. ISBN 3540219730.
- A. Jasra, C. Holmes, and D. Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, pages 50–67, 2005.
- J. Jost. Convex functionals and generalized harmonic maps into spaces of non positive curvature. *Commentarii Mathematici Helvetici*, 70(1):659–673, 1995.
- J. Jost. *Nonpositive curvature: geometric and analytic aspects*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 1997. ISBN 3-7643-5736-3.
- J. Jost. Nonlinear dirichlet forms. In J. Jost, W. Kendall, U. Mosco, M. Röckner, and K.-T. Sturm, editors, *New directions in Dirichlet forms*, pages 1–47. AMS and International Press, 1998.
- J. Jost and X. Li-Jost. *Calculus of variations*, volume 64. Cambridge University Press, 1998.
- T. H. Jukes and C. R. Cantor. Evolution of protein molecules. 1969.
- J. B. Kadane. *Principles of uncertainty*. CRC Press, 2011.
- J. B. Kadane, M. J. Schervish, and T. Seidenfeld. Statistical implications of finitely additive probability. *Rethinking the Foundations of Statistics*, page 211, 1999.
- O. Kallenberg. *Foundations of modern probability*. Probability and its applications. Springer, 2 edition, 2002.
- S. Karlin and H. Taylor. *A First Course in Stochastic Processes*. Elsevier Science, 1975. ISBN 9780080570419.
- K. G. Karol, R. M. McCourt, M. T. Cimino, and C. F. Delwiche. The closest living relatives of land plants. *Science*, 294(5550):2351–2353, 2001.

- J. M. Keynes. *A treatise on probability*. MacMillan, 1921.
- H. Kielhöfer. Variationsrechnung: Eine einföhrung in die theorie einer unabhangigen variablen mit beispielen und aufgaben. 2010.
- M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, 1985.
- M. Kimura et al. Evolutionary rate at the molecular level. *Nature*, 217: 624–626, 1968.
- B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional bayesian statistics. *The Annals of Statistics*, pages 837–877, 2006.
- A. Klenke. *Probability theory: A comprehensive course*. Springer, 2008.
- A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Verlag von Julius Springer, 1933.
- B. O. Koopman. The axioms and algebra of intuitive probability. *Annals of Mathematics*, pages 269–292, 1940.
- M. Křivanek and J. Moravek. NP-hard problems in hierarchical-tree clustering. *Acta Informatica*, 23(3):311–323, 1986.
- S. Kumar, A. J. Filipski, F. U. Battistuzzi, S. L. K. Pond, and K. Tamura. Statistics and truth in phylogenomics. *Molecular biology and evolution*, page msr202, 2011.
- P. S. Laplace. *Théorie analytique des probabilités*. Courcier, 1820.
- N. Lartillot and H. Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109, 2004. doi: 10.1093/molbev/msh112. URL <http://dx.doi.org/10.1093/molbev/msh112>.
- N. Lartillot, T. Lepage, and S. Blanquart. PhyloBayes 3: A Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17):2286–2288, 2009.
- D. S. Latchman. Transcription factors: an overview. *The international journal of biochemistry & cell biology*, 29(12):1305–1312, 1997.
- E. L. Lawler. *Combinatorial optimization: networks and matroids*. Courier Dover Publications, 1976.
- C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, 262(5131):208–214, 1993.
- J. M. Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer, 1997.
- I. M. Lemnian, R. Eggeling, and I. Grosse. Extended sunflower hidden markov models for the recognition of homotypic cis-regulatory modules. In *OASIS-OpenAccess Series in Informatics*, volume 34. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2013.

- P. O. Lewis, M. T. Holder, and K. E. Holsinger. Polytomies and bayesian phylogenetic inference. *Systematic Biology*, 54(2):241–253, 2005.
- S. Li, D. K. Pearl, and H. Doss. Phylogenetic tree construction using markov chain monte carlo. *Journal of the American Statistical Association*, 95(450):493–508, 2000.
- W. Li, C. Luo, and C. Wu. Evolution of dna sequences. *Molecular evolutionary genetics. Plenum, New York*, 1:1–94, 1985.
- E. H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate studies in mathematics*. American Mathematical Society, Providence, RI, 2 edition, 2001.
- T. M. Liggett. *Continuous time Markov processes: an introduction*, volume 113. American Mathematical Soc., 2010.
- D. Lindley. *Bayesian statistics: A review*. SIAM, 1972.
- J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.
- M. Loève. *Probability Theory I*, volume 45 of *Graduate texts in mathematics*. Springer-Verlang New York Inc., 4th edition, 1977.
- M. Loève. *Probability Theory II*, volume 46 of *Graduate texts in mathematics*. Springer-Verlang New York Inc., 4th edition, 1978.
- A. Löytynoja and N. Goldman. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10557–10562, 2005.
- A. Löytynoja and N. Goldman. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–1635, 2008.
- W. P. Maddison and L. L. Knowles. Inferring phylogeny despite incomplete lineage sorting. *Systematic biology*, 55(1):21–30, 2006.
- T. Margush and F. R. McMorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43(2):239–244, 1981.
- B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Rev. Française Informat. Recherche Opérationnelle*, 4 (Ser. R-3):154–158, 1970.
- L. A. McCue, W. Thompson, C. S. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic acids research*, 29(3):774–782, 2001.
- C. A. Meacham. A manual method for character compatibility analysis. *Taxon*, pages 591–600, 1981.
- C. A. Meacham. Theoretical and computational considerations of the compatibility of qualitative taxonomic characters. In *Numerical taxonomy*, pages 304–314. Springer, 1983.

- N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.
- E. Miller, M. Owen, and J. S. Provan. Averaging metric phylogenetic trees. *arXiv preprint arXiv:1211.7046*, 2012.
- J. W. Miller and M. T. Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. In *Advances in Neural Information Processing Systems*, pages 199–206, 2013.
- B. G. Mirkin and L. Chernyi. Measurement of the distance between distinct partitions of a finite set of objects. *Automation and Remote Control*, 31(5):786, 1970.
- E. Mossel and E. Vigoda. Phylogenetic mcmc algorithms are misleading on mixtures of trees. *Science*, 309(5744):2207–2209, 2005.
- R. Neal. Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC, 2011.
- R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, 1995.
- R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and computing*, 6(4):353–366, 1996.
- I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. *arXiv preprint physics/0108025*, 2001.
- M. Owen. Computing geodesic distances in tree space. *SIAM Journal on Discrete Mathematics*, 25(4):1506–1529, 2011.
- M. Owen and S. Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 8:2–13, 2011.
- N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):123–231, 2013.
- P. J. Park. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, 2009.
- J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI*, pages 133–136, 1982.
- J. Pitman. Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields*, 102(2):145–158, 1995.
- J. Pitman. Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series*, pages 245–267, 1996.
- J. Pitman and M. Yor. The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2): 855–900, 1997.
- H. Poincaré. *Science and hypothesis*. Science Press, 1905.

- S. Poppe. *Inductive Characterization and Statistical Application of Sequential Prediction Rules*. PhD thesis, Mathematisches Institut der Universität Leipzig, 2015.
- S. Poppe, P. Benner, and T. Elze. A predictive approach to nonparametric inference for adaptive sequential sampling of psychophysical experiments. *Journal of mathematical psychology*, 56(3):179–195, 2012.
- M. Ptashne and A. Gann. Transcriptional activation by recruitment. *Nature*, 386(6625):569–577, 1997.
- F. P. Ramsey. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198, 1931.
- W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.
- S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *Mathématiques et Sciences humaines*, 82:13–29, 1983.
- M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Neural Networks, 1993., IEEE International Conference on*, pages 586–591. IEEE, 1993.
- C. Robert. *The Bayesian choice*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2001. ISBN 0-387-95231-4. From decision-theoretic foundations to computational implementation, Translated and revised from the French original by the author.
- C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 1 edition, 1999. ISBN 038798707X.
- G. O. Roberts, J. S. Rosenthal, and P. O. Schwartz. Convergence properties of perturbed markov chains. *Journal of applied probability*, pages 1–11, 1998.
- R. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM J. on Control and Optimization*, 14(5), 1976.
- M. Salemi and A.-M. Vandamme. *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press, 2003.
- L. Salichos and A. Rokas. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature*, 497(7449):327–331, 2013.
- M. J. Sanderson and H. B. Shaffer. Troubleshooting molecular phylogenetic analyses. *Annual Review of Ecology and Systematics*, pages 49–72, 2002.
- L. J. Savage. *The foundations of statistics*. Dover Publications, Inc., 1972.
- C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003. ISBN 0-19-850942-1.

- J. A. Shohat and J. D. Tamarkin. *The Problem of Moments*, volume 1. American Mathematical Soc., 1943.
- R. Siddharthan. Phylogibbs-mp: module prediction and discriminative motif-finding by gibbs sampling. *PLoS computational biology*, 4(8): e1000156, 2008.
- R. Siddharthan, E. D. Siggia, and E. van Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS computational biology*, 1(7):e67, 2005.
- A. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. *Journal of Computational Biology*, 11(2-3):413–428, 2004.
- A. Siepel and D. Haussler. Phylogenetic hidden markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer, 2005.
- J. B. Slowinski and R. D. Page. How should species phylogenies be inferred from sequence data? *Systematic Biology*, pages 814–825, 1999.
- S. Song, L. Liu, S. V. Edwards, and S. Wu. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, 109(37):14942–14947, 2012.
- M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 795–809, 2000.
- K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, volume 338 of *Contemp. Math.*, pages 357–390. Amer. Math. Soc., Providence, RI, 2003.
- J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis. Lie markov models. *Journal of theoretical biology*, 298:16–31, 2012a.
- J. G. Sumner, P. D. Jarvis, J. Fernández-Sánchez, B. T. Kaine, M. D. Woodhams, and B. R. Holland. Is the general time-reversible model bad for molecular phylogenetics? *Systematic biology*, 61(6):1069–1074, 2012b.
- Y. Suzuki, G. V. Glazko, and M. Nei. Overcredibility of molecular phylogenies obtained by bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, 99(25):16138–16143, 2002.
- S. Tavaré. Some probabilistic and statistical problems in the analysis of dna sequences. *Lectures on mathematics in the life sciences*, 17:57–86, 1986.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476), 2006.
- M. Thomas-Chollier, O. Sand, J.-V. Turatsinze, M. Defrance, E. Vervisch, S. Brohée, J. van Helden, et al. Rsat: regulatory sequence analysis tools. *Nucleic acids research*, 36(suppl 2):W119–W127, 2008.

- A. Tomovic and E. J. Oakeley. Position dependencies in transcription factor binding sites. *Bioinformatics*, 23(8):933–941, 2007.
- M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144, 2005.
- A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- A. W. van der Vaart and J. H. Van Zanten. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, pages 1435–1463, 2008.
- A. Vandenberg, Y. Miyamoto, N. Takimoto, T. Kusakabe, and K. Nakai. Markov chain-based promoter structure modeling for tissue-specific expression pattern prediction. *DNA research*, 15(1):3–11, 2008.
- C. Villegas. On qualitative probability/sigma-algebras. *The Annals of Mathematical Statistics*, pages 1787–1796, 1964.
- J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton UP, 1953.
- M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.
- Y. Wakabayashi. The complexity of computing medians of relations. *Resenhas IME-USP*, 3(3):323–349, 1998.
- A. Wald. Statistical decision functions. 1950.
- W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Review Genetics*, 5(4):276–287, 2004.
- J. Williamson. Countable additivity and subjective probability. *The British Journal for the Philosophy of Science*, 50(3):401–416, 1999.
- J. M. Winn and C. M. Bishop. Variational message passing. In *Journal of Machine Learning Research*, pages 661–694, 2005.
- D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841, 1995.
- D. B. Woodard, J. S. Rosenthal, et al. Convergence rate of markov chain methods for genomic motif discovery. *The Annals of Statistics*, 41(1): 91–124, 2013.
- Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, 10(6):1396–1401, 1993.
- Z. Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution*, 11(9):367–372, 1996.

- Z. Yang. Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. *Molecular biology and evolution*, 24(8):1639–1655, 2007.
- Z. Yang and B. Rannala. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, 14(7):717–724, 1997.
- Z. Yang and B. Rannala. Branch-length prior influences bayesian posterior probability of phylogeny. *Systematic biology*, 54(3):455–470, 2005.
- G. A. Young and R. L. Smith. *Essentials of statistical inference*, volume 16. Cambridge University Press, 2005.
- S. L. Zabell. *Symmetry and its discontents: essays on the history of inductive probability*. Cambridge University Press, 2005.
- B. Zhong, L. Liu, Z. Yan, and D. Penny. Origin of land plants using the multispecies coalescent model. *Trends in plant science*, 18(9):492–495, 2013.
- Q. Zhou and J. S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–916, 2004.
- Q. Zhou and W. H. Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proceedings of the national academy of sciences of the United States of America*, 101(33):12114–12119, 2004.
- R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, 2009.
- E. Zuckerkandl and L. Pauling. Molecular disease, evolution and genetic heterogeneity. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, 1962.
- E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, 1965.

Daten zum Autor:

Name: Philipp F. Benner

geboren am: 09.04.1984 in Göttingen, Deutschland

10/2005-07/2008: Bachelor-Studium der Kognitionswissenschaft an der Universität Osnabrück, Wahlpflichtmodule: Informatik, Mathematik, Neuroinformatik, Künstliche Intelligenz und Neurobiologie (B.Sc. mit Auszeichnung)

07/2007-12/2007: Auslandssemester an der Technisch-Naturwissenschaftlichen Universität Norwegens (NTNU-Trondheim), Fachbereich Informatik

09/2008-04/2010: Master-Studium am Institut für Neuroinformatik, Eidgenössische Technische Hochschule (ETH) Zürich (M.Sc. ETH mit Auszeichnung)

06/2010-: Doktorand am Max-Planck-Institut für Mathematik in den Naturwissenschaften in Leipzig

Bibliographische Daten

P. Benner, M. Bacák, and P.-Y. Bourguignon. *Point estimates in phylogenetic reconstructions*. *Bioinformatics*, 30(17):i534-i540, 2014.

P. Benner and S. Poppe. *Stochastische Prozesse und Bayessches Schätzen*. In Nicole J. Saam and Norman Braun, editors, *Handbuch Modellbildung und Simulation in den Sozialwissenschaften*, Springer, 2014.

S. Poppe, P. Benner, and T. Elze. *A predictive approach to nonparametric inference for adaptive sequential sampling of psychophysical experiments*. *Journal of Mathematical Psychology*, 56(3):179–195, 2012.

R. Stoop, S. Martignoli, P. Benner, R. Stoop, and Y. Uwate. *Shrimps: Occurrence, scaling and relevance*. *International Journal of Bifurcation and Chaos*, 22(10), 2012.

R. Stoop, P. Benner, and Y. Uwate. *Real-world existence and origins of global shrimp organization on spirals*. *Physical Review Letters*, 105(7) : 074102, 2010.

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 11. Mai 2015

.....

(Philipp F. Benner)